iRepertoire, Inc.

iRweb: Data Analysis Guide

V20220509

iRepertoire, Inc.

Version Updates

Update Classification	Update Description	Version
Minor	Removed prior YPL version descriptions and added YPL6 description.	V202205
Minor	Removed references to obsolete features.	V202204
Minor	Removed extraneous note page 10 regarding TCR filtering.	V202108
Minor	Added note specifying version differences.	V202107
Minor	Updated styling. Added YPL5 description.	V202106
Minor	Added information related to UMI pipeline versions, pupated table of contents, altered formatting in some pages.	V202104
Moderate	Updated graphics, included newer Summary Table information, updated calculation descriptions for D50, Shannon Entropy, and Diversity Index (Di.), updated information for dam-PCR libraries and UMI, added Terms and Definitions, added QuickStart & Applications	V201901
Moderate	Updated information on Diversity Index (Di), Added information about Entropy (Shannon Entropy)	V201810

iRepertoire, Inc.

Table of Content

Introduction	5
Quick Start	6
About the Pipeline	7
Data Structure & Design	10
Logging In & Accessing Data	12
Demo Data Sets	15
Sample Analysis Menu	17
Show 2D Map	19
Show 3D Map	21
List CDR3 new	22
List CDR3 old	23
List CDRs	23
CDR3 algebra	24
D50	27
Diversity Index (Di)	29
Entropy (Shannon)	30
Tree Map	30
Distribution Analyses (V-gene, J-gene, Nucleotide trimming, etc.) & Normalized	31
Raw Data	33
Raw Data: ######_pep.csv	34
Frequently Asked Questions	35
Terms & Definitions	36
Contact Information	36

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. If the method utilized incorporates unique molecular identifiers, a separate algorithm is applied in order to utilize this information for frequency and error detection. 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 utilizes IMGT nomenclature.
- 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, and Raw Data contains raw formats of all of the data on the site.
- 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.

iRepertoire, Inc.

Quick Start & Applications

The information on this page is geared towards more experienced iRweb users. If you are new to iRweb, or new to immune repertoire sequencing overall, we recommend that you read through the whole guide before using the information here.

How do I look for shared CDR3s?

Select a sample you wish to compare to others. In the left-hand menu, click CDR3 algebra and select all other samples you wish to use in the comparison. Click “Share”. Please see pages 24-26 for more.

How do I compare diversity across samples?

Please download the Summary Table. Comparisons can be made either by D50, Diversity Index (Di), or by Shannon entropy (Entropy). Please be sure to check the notes on pages 27-28, 29, and 30 respectively when your sample has more than ten-thousand unique CDR3s.

How do I compare CDR3s across different time points?

Just as with the description for shared CDR3s mentioned previously on this page, please use CDR3 Algebra to make your comparisons. If you wish to include the top 100 unique CDR3s for each sample, regardless of sharing, click “Extra”. Be sure to select the sample you wish to use as your base line or first time point and calculate CDR3 Algebra from the Sample Analysis menu specific for that sample.

Where is the full-length read?

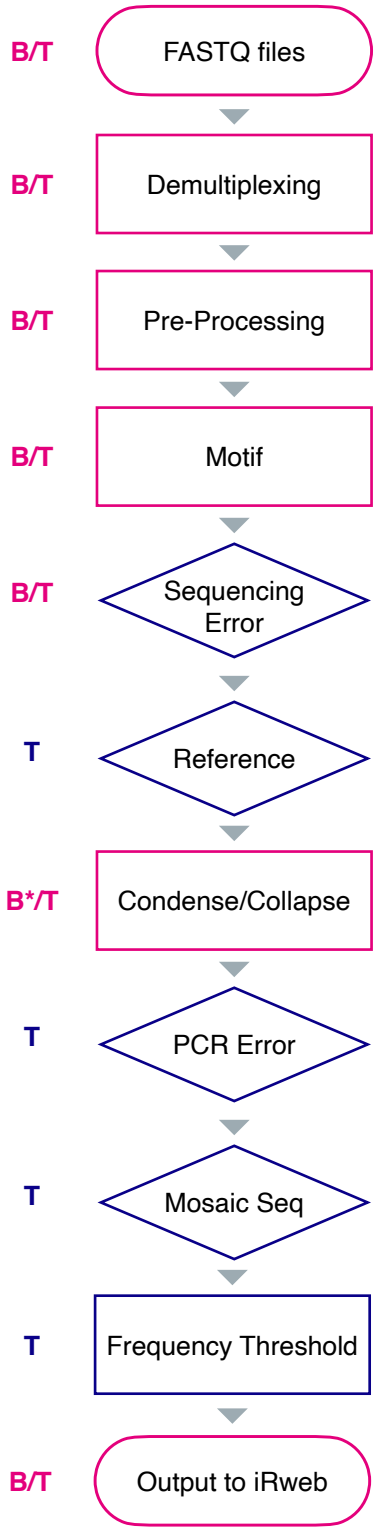
If you wish to access full-length reads, please download the Raw Data for either the whole Project or for a given Sample. Within the directory you download, the ######_pep.csv file will contain a "joinedSeq" column that has the full-length read.

Since a small portion of FR1 is missing, can we obtain the full sequence of the V-gene from the beginning of FR1?

Yes, although this is not provided in iRweb. We have a program that accepts both .ipair files and pep.csv files and is available at inferseq.irepertoire.com.

iRepertoire, Inc.

About the non-UMI RepSeq Pipeline: BCR & TCR



The flow chart to the left is a general overview of the proprietary non-umi bioinformatic pipeline. It is important to note that all filtering with the iRweb non-umi pipeline does not include error correction. Errors are identified and removed.

BCR & TCR data is analyzed differently within the pipeline.

In the flowchart to the left, when both TCR and BCR data are analyzed, the process or stage will be noted in purple or with either a "T" or B" to emphasize which stages apply to which type of data.

BCR data is subject to somatic hypermutation and in an effort to preserve these data, we have omitted the use of some of the filters and processes that are used for TCR data. One of the side effects of such processing involves the creation of large numbers of singleton uCDR3s which *may* be disregarded in downstream analyses, but allows us to preserve a vast majority of signal that might otherwise be removed by the very stringent analyses applied to TCR data [Yang, et al. eLife (2015)]. It is highly recommended to generate BCR libraries using RepSeq+ as the incorporation of umi assist with differentiating PCR error from somatic hypermutation.

**There are four separate stages to the Condense/Collapse stage of the pipeline. With BCR data, only one level occurs (sequences are identical at a nucleotide level).*

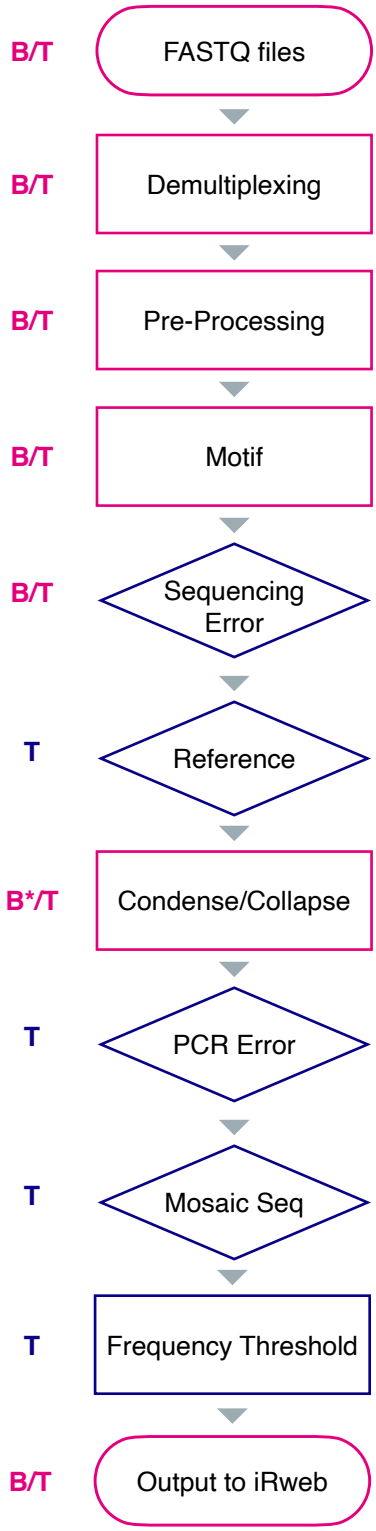
Demultiplexing & Pre-Processing

Once raw fastq files are available, demultiplexing occurs. If dual indices are included in the primer system, this occurs first, but all of our primer systems include a proprietary molecular barcode that is also demultiplexed. Additional pre-processing includes trimming the ends of reads for quality, overlapping and stitching, mapping to the IMGT™ reference sequences, and locating the CDRs based upon positional information.

The **Motif** filter is the first decision point in the data: *Does the CDR3 region contain the canonical amino acid motifs flanking it upstream and downstream?* Typically, the uCDR3 sequence begins with a conserved cysteine at the beginning and an F/WGXG motif at the end.

iRepertoire, Inc.

About the non-UMI RepSeq Pipeline: BCR & TCR



Our **Sequencing Error** filter will then look within the stitched overlap from pre-processing. Any reads that are not 100% identical within the overlap are thrown out.

All of those sequences that pass the Sequencing Error filter will then move on to the **Reference** filter. If a sequence varies from the reference, it is removed from the data.

Reads passing the Reference filter are then **Condensed** and/or **Collapsed** by nucleotide sequence identity. This step is what creates the final frequency or copy number that is output to iRweb.

The **PCR error** filter will remove indel and substitution errors by PCR.

The **Mosaic** filter will detect and remove chimeric sequences created during PCR.

The **Frequency Threshold** filter will remove all those reads that have a frequency of 1 or less. Provided that a read has not collapsed/condensed with the previous stages and only has a frequency of 1.

Additional processes include parsing completely analyzed data into the outputs seen on iRweb: V- & J-gene usage, D50, Di, Entropy, N addition, etc.

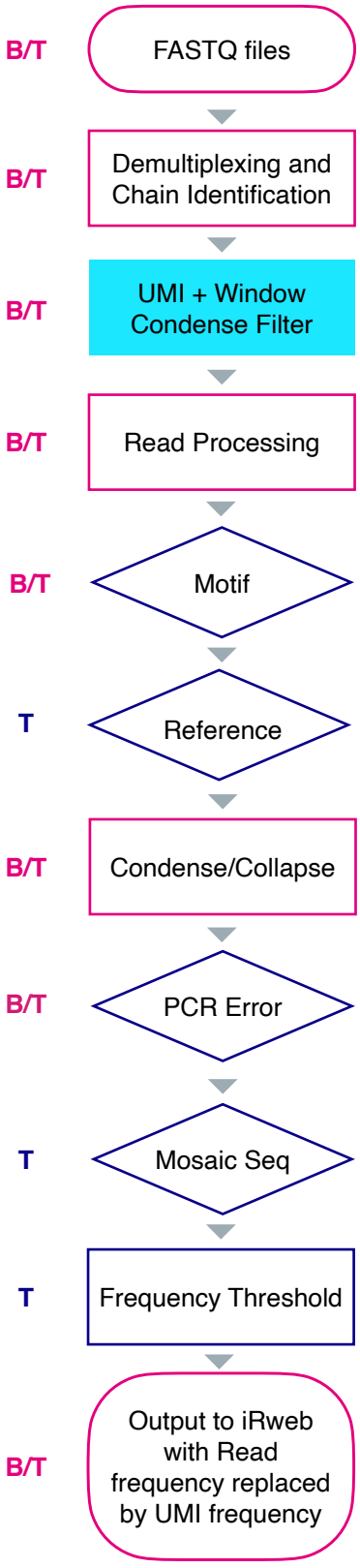
Note: If demultiplexed and stitched reads are requested, please note they come from the Demultiplexing and Pre-Processing steps seen in this flowchart. No additional processing beyond trimming and stitching for quality have occurred. In the case of demultiplexed reads only, trimming has not even occurred.

Important: As mentioned on the first page, BCR analysis is slightly different than TCR analysis. BCR libraries, in an effort to preserve somatic hypermutation, do not go through all of these filtering processes. BCR data goes through all pre-processing stages, the Motif filter, and the Sequencing filter. It is highly recommended to generate BCR libraries using RepSeq+ as the incorporation of umi assist with differentiating PCR error from somatic hypermutation.

Wang, C. *et al.* Data Analysis Pipeline for Immune Repertoire Sequencing Data. (2009)

iRepertoire, Inc.

About the UMI RepSeq+ Pipeline YPL6: BCR & TCR



The flow chart to the left is a general overview of the proprietary YPL6-UMI bioinformatic pipeline. BCR data sets have a reduced set of filters applied to them in order to maintain data relevant to somatic hypermutation. The steps labeled "B/T" apply to both BCR and TCR datasets, while steps labeled "T" only apply to TCR datasets.

Demultiplexing & Chain Identification: The reads are demultiplexed first. Then the chain information is identified for further separation before the next step.

The **UMI+Window Condense Filter** step incorporates an algorithm whereby each umi is identified along with a nucleotide window in the middle of the sequence. If a single umi has multiple associated window sequences, only the sequence with the highest number of copies is kept. The other reads are discarded.

Once the reads have been collapsed by umi, they undergo **Read Processing**. Single end reads are reverse complemented, and paired end reads are trimmed for quality and then stitched/paired together. Afterward the genes are mapped to IMGT reference sequences and the CDRs are identified by positional information. At the end, the read counts are reattached to each read independent of and in addition to the umi counts.

The **Motif** filter is the first decision point in the data: *Does the CDR3 region contain the canonical amino acid motifs flanking it upstream and downstream?* Typically, the uCDR3 sequence begins with a conserved cysteine at the beginning and an F/WGXG motif at the end.

Those sequences that have the canonical motifs move to to the **Reference** filter. Reads are compared to IMGT references and discarded if they do not match.

Reads passing the Reference filter are then **Condensed** and/or **Collapsed** by nucleotide sequence identity into representative clonotypes with a read and umi count. In this step, the selection criteria for B cell differs notable by using a larger target sequence region, more than just the CDR3, to account for somatic hypermutation in the V region.

The **PCR Error** filter will remove indel and substitution errors caused by PCR.

The **Mosaic Seq** filter will detect and remove chimeric sequences created during PCR and sequencing.

The **Frequency Threshold** filter removes any clonotypes with a read count of 1 unless a qualifying exception such as another read under the same UMI is detected in order to maintain clones that may still be real.

iRepertoire, Inc.

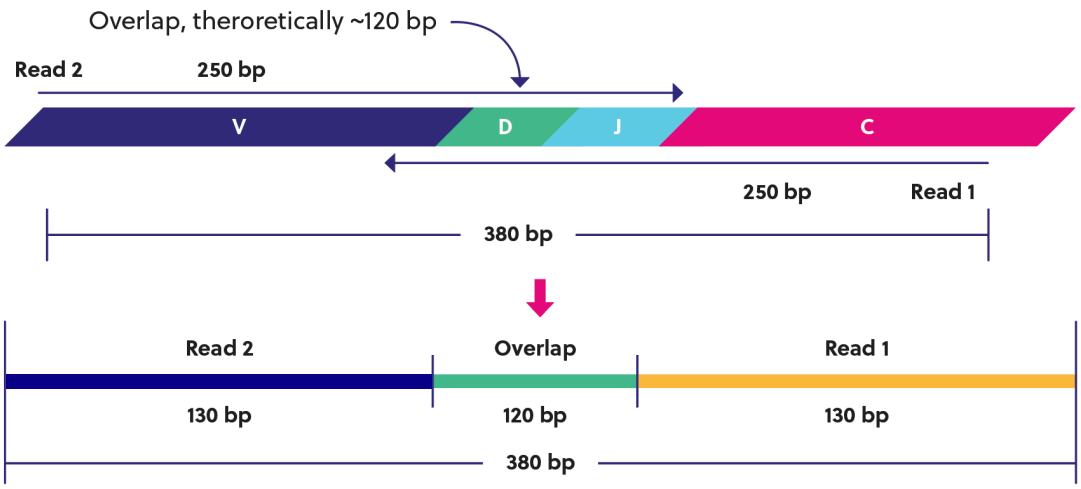
Data Structure & Design

iRepertoire's offers two different types of library preparation via amplicon rescue multiplex PCR (arm-PCR) and dimer avoided multiplex PCR (dam-PCR). dam-PCR has the option of including universal molecular identifiers (UMI). Primer designs for both systems are built around the paired-end or single-end reads available with the Illumina© MiSeq©, NextSeq©, NovaSeq© 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. Data can also be prepared as single-end read. In this case, the data for a strand of DNA is read from Read 1 uni-directionally from the C-region into the V-region. Data for FR1 and CDR1 are lost but coverage of the CDR3 and into CDR2 is provided.

For arm-PCR libraries: RepSeq

For all of our arm-PCR 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.

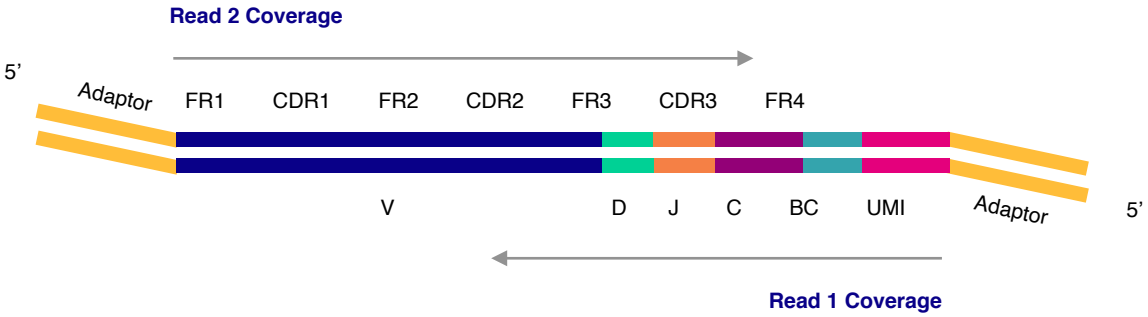
For human, the average MiSeq amplicon length is **500 bp** *including adaptors*. Subtract approximately 120 bp for the adaptors, and **the insert is on average 380 bp**. Remember that this value varies depending upon receptor editing.



iRepertoire, Inc.

For dam-PCR libraries: RepSeq+ PER

Read 1 (R1) starts within the constant region and will capture sequencing information in the reverse direction along the V(D)J-C recombination (see Figure 2). Read 2 (R2) starts within FR1 for human samples and will capture information on the sense strand heading towards the V(D)J recombination site.



For dam-PCR libraries: RepSeq+ SER

In order to take advantage of smaller cycle kits (such as NextSeq 300 cycle kits), it is possible to read the RepSeq+ libraries as single end read (SER) from Read 1 (R1). R1 starts within the constant region and will capture sequencing information in the reverse direction along the V(D)J-C recombination (see Figure 2). In general, sequencing data will be available through FR2 providing coverage of the CDR3 region and a large portion of the V-gene.

Note: When reads are stitched within the analysis pipeline, stitched information is converted to uppercase and single-read information is in lower case. If, in the case of single-end data, in which only R1 is used, all nucleotide sequence information will be capitalized.

An example pulled from the joinedSeq column of the downloaded Raw Data directory for sample data follows:

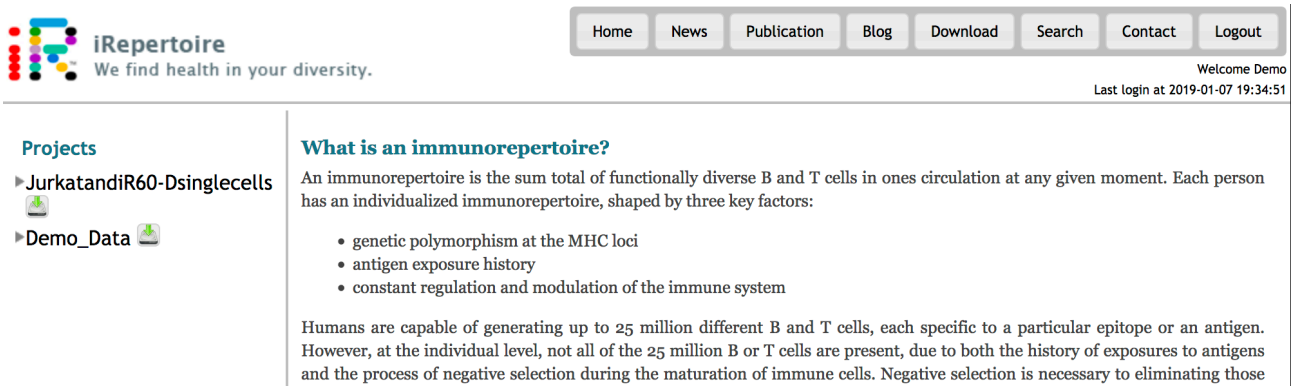
agactctcctgtacagcgtctggattcacctttagcagctatgccatgagctgggtccgccaggctccagggaaggggctggagtgggtc
tcagctattagtggtagtggtggtagcacatactacgcagactccgtgaAGGGCCGGTTCACCATCTCCAGAGACAATTCCAAGAACACG
Ctgtatctgcaaatgaacagcctgagagccgaggacacggccgtatattactgtgcgagaagtcctgtggtgactgcccccgaagactac
tggggccagggaaccctggtcaccgtctcctcagggagtgcatccgccccaaccctttttccccctcgtct

iRepertoire, Inc.

Logging In and Accessing Data

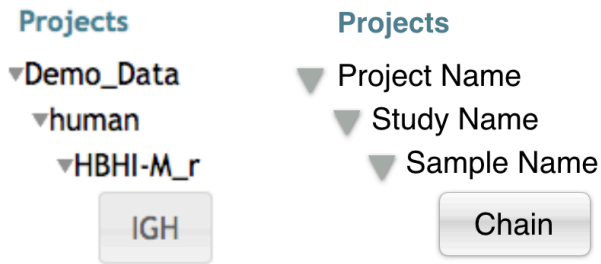
To access data on iRweb, please go to https://irweb.irepertoire.com/nir/ and log in with the credentials provided to you by Data Support. If you would like to look at a demo data set on iRweb, please see page 15 for more information.

If login was successful, you will be able to see a header like in the image below.



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 Table of their data. The Summary Table (next page) 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.

If users have multiple projects and studies available to them, the arrows will not be available and only a Summary Table will be made available via clicking the Project Name. This was done in an effort to reduce loading times for data.



Project	Study	Sample	Sample id	Species	Chain	Reads	CDR3	Unique CDR3	D50	Diversity Index	Entropy
Demo_Data	human	HBHI-M	21747	h	IGH	1420700	1403796	106292	28.2	36.4	13.0
Demo_Data	human	HBHI-M_r	21748	h	IGH	1502092	1485800	104070	28.4	36.1	13.0
Demo_Data	human	HBKLI-M	21744	h	IGL	130838	123272	3366	0.9	5.9	8.0
Demo_Data	human	HBKLI-M	21746	h	IGK	1268618	1204844	11493	0.5	3.3	8.5
Demo_Data	human	HBKLI-M_r	21743	h	IGL	54293	50444	3464	1.1	9.2	8.5
Demo_Data	human	HBKLI-M_r	21745	h	IGK	1142987	1083301	14869	0.5	4.9	8.9
Demo_Data	human	HTAIvc_01	21583	h	TRA	2249690	2048876	124357	17.7	29.6	12.5
Demo_Data	human	HTBI-M	21014	h	TRB	832285	827305	72352	0.3	15.6	9.2
Demo_Data	human	HTDI-M	21552	h	TRD	197342	184142	4631	3.0	11.7	9.6
Demo_Data	human	HTGI-M	21551	h	TRG	93685	57117	2221	1.1	11.2	8.0
Demo_Data	mouse	MBHI-M_SortedCell	21555	m	IGH	340136	330852	12195	4.4	8.4	10.7
Demo_Data	mouse	MBKLI-M-05	21462	m	IGL	624207	483743	971	0.1	1.8	1.2
Demo_Data	mouse	MBKLI-M-05	21464	m	IGK	19113	18215	1250	2.1	8.1	7.4
Demo_Data	mouse	MTBI-M	21013	m	TRB	1007946	964388	96716	32.3	39.2	13.1
Demo_Data	mouse	MTDI-M	25671	m	TRD	46231	35176	2399	11.1	19.8	10.2
Demo_Data	mouse	MTGI-M	25670	m	TRG	251950	158619	1755	0.6	6.5	6.1
Project	Study	Sample	Sample id	Species	Chain	Reads	CDR3	Unique CDR3	D50	Diversity Index	Entropy

Clicking either the **Sample** or **Sample id** will bring you to a separate Sample Analysis window.

Sample id Generated and assigned automatically by the pipeline as samples complete analysis. This ID can be used as a reference for the sample in addition to the Sample Name. When communicating with data support, please also include the sample id as this is an unambiguous sample identifier.

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 (see page 27 for more information)

Diversity Index Another proprietary index (see page 29 for more information)

Entropy This value is the calculated Shannon entropy for the sample (more information on page 30)

Note: Clicking the download icon on the Summary Table will download a .csv version of the Summary Table.

To access the demo data:

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 Data Support at datasupport@irepertoire.com.

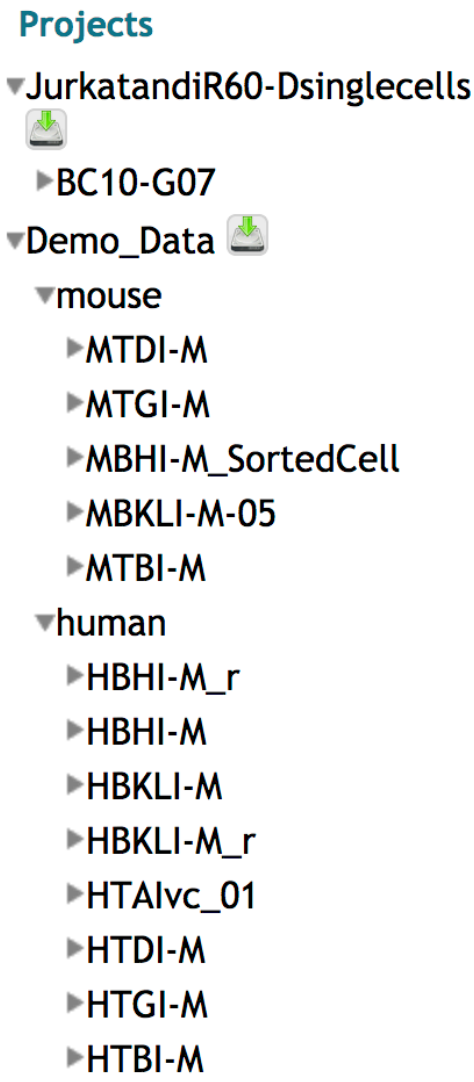
Username: demo
Password: 12345

Note: 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.

iRepertoire, Inc.

Demo Data Sets

We have made a number of types of data available in the demo of iRweb. Once you have logged in, the left-hand menu will contain the following:



The BC10-G07 sample is from our iPair services. The data here is from a single cell, sorted into a well on a 96-well plate with co-amplification of TCRa and TCRb. Note, iPair data should be viewed in the iPair Analyzer, not on iRweb.

All of the data included in this section of the demo pertains to our long-read primer systems for mouse.

All of the data included in this section of the demo is for our human primer systems. The majority of the data is representative of our long-read systems (-M), but we do have data for one of our short-read systems for TCR alpha (HTAIvc_01). Anything in this sample set with an underscore r (_r) is a technical repeat of the same sample. In the case of HBHI-M and HBHI-M_r, these are the same sample, amplified with two different barcodes, pooled, and then sequenced together.

iRepertoire, Inc.

JurkatandiR60-Dsinglecells

As mentioned previously, this is an example of the data available from our single cell sequencing service, iPair. iPair data is not typically viewed in iRweb. In general, we recommend the iPair Analyzer, a separate application for reading .ipair files. The data contained in the demo for this is representative of a single cell in a single well of a 96-well plate. Human TCR alpha and TCR beta chains were co-amplified together in the same reaction. The single cell data serves to demonstrate the types of noises in Next Generation Sequencing data, in particular RepSeq data. The two images below are from the List CDR3 old for both the TCR alpha and TCR beta chains for the sample. Because of the scale of single cell data, noise due to PCR error and sequencing error tends to be exaggerated, which we refer to as a Cosmic Effect. The Cosmic Effect entails that for each similar or identical amino acid CDR3s, clonotypes with a frequency at or below a ratio of 1:500 with respect to the highest frequency clone are disregarded as sequencing or PCR error. Single cell data is an upper bound limit to the Cosmic Effect.

CDR3 list

Show 100 entries Search:

Frequency	Peptide	V	J
144	AETISGNTPLV	hTRAV13-2	hTRAJ29
2	AETVSGNTPLV	hTRAV13-2	hTRAJ29

Showing 1 to 2 of 2 entries

CDR3 list

Show 100 entries Search:

Frequency	Peptide	V	J
347	ASSSRTGGNTGELF	hTRBV7-2	hTRBJ2-2
5	ASSSRTGGDTGELF	hTRBV7-2	hTRBJ2-2
4	ASSSRTGGSTGELF	hTRBV7-2	hTRBJ2-2
3	ASSSRAGGNTGELF	hTRBV7-2	hTRBJ2-2
3	ASSSGTGGNTGELF	hTRBV7-2	hTRBJ2-2
2	ASSFRTGGNTGELF	hTRBV7-2	hTRBJ2-2

Showing 1 to 6 of 6 entries

Mouse

With the exception of the sample labeled “SortedCell” all of the data here are representative of use of our long-read mouse primer systems on total RNA of PBMCs from whole blood. The “SortedCell” sample was from total RNA of B cells sorted from whole blood.

Human

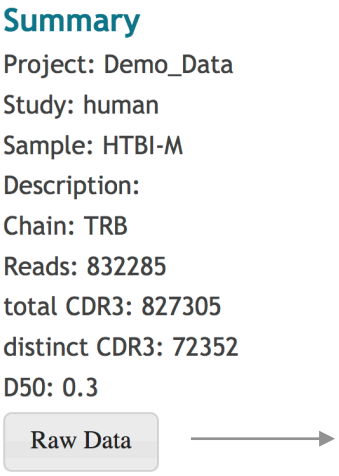
HTAIvc_01 is a representation of our short-read primer systems for human. The HBHI-M and HBKLI-M samples, with their respective technical repeats (ending in _r) are repeat amplifications from the same RNA extraction.

iRepertoire, Inc.

Sample Analysis Menu

Once a sample has been selected either via clicking the Sample ID or Sample Name, 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 (see pages 31-33 for more information).

This Sample Analysis Menu consists of four sections: Summary, Analysis, Distribution, and Normalized.



The information in the Summary consists of the same information that was previously available in the Summary Table.

Raw Data will generate a .zip of tab-separated (.csv) versions of almost all outputs for data on iRweb. More information is available on page ___. This download *does not* contain the raw sequencing data for the sample.

2D heat map of V- & J-gene combinations in relation to frequency (page 19)	
3D chart of V- & J-gene combinations in relation to frequency (page 21)	
List of CDR3s w/ hierarchical tree maps for V(D)J-C combinations (page 22)	
Another iteration of V-J gene combinations by peptide (page 23)	
A list of all called CDRs for a given peptide sequence (page 23)	
Shared CDR3 calculations across samples (page 24)	
D50 Calculation Graph	
Diversity Index Graph	
A graphical representation of the V-J combinations (page 30)	

iRepertoire, Inc.

Sample Analysis Menu

Once a sample has been selected either via clicking the Sample ID or Sample Name, 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 (see pages 31-33 for more information).

Distribution

V usage	A percentage use of V-genes within the sample (page 31)
J usage	A percentage use of J-genes within the sample (page 31)
V trimming	V-gene nucleotides trimmed through recombination events (page 32)
J trimming	J-gene nucleotides trimmed through recombination events (page 32)
CDR3 length	CDR3 lengths calculated for the sample (page 32)
N addition	Nucleotide additions within the sample based on references (page 32)

Normalized

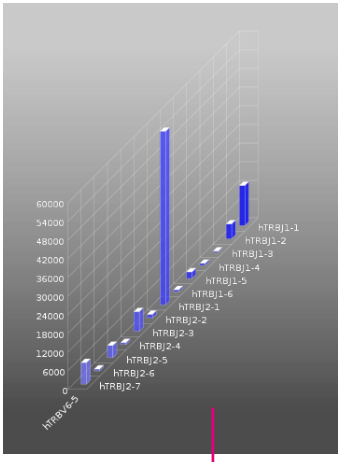
- V usage
- J usage
- V trimming
- J trimming
- CDR3 length
- N addition

In brief during Normalization, 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 removes the effect that a single or handful of dominant clones can have on the rest of the perspective of the repertoire. If the diversity of the repertoire is high, then the normalized result will look very similar to the distribution with all reads incorporated. However, if the diversity is low, this indicates there is clonal dominance and the effect of these clones on these metrics will be reduced by this type of normalization.

iRepertoire, Inc.

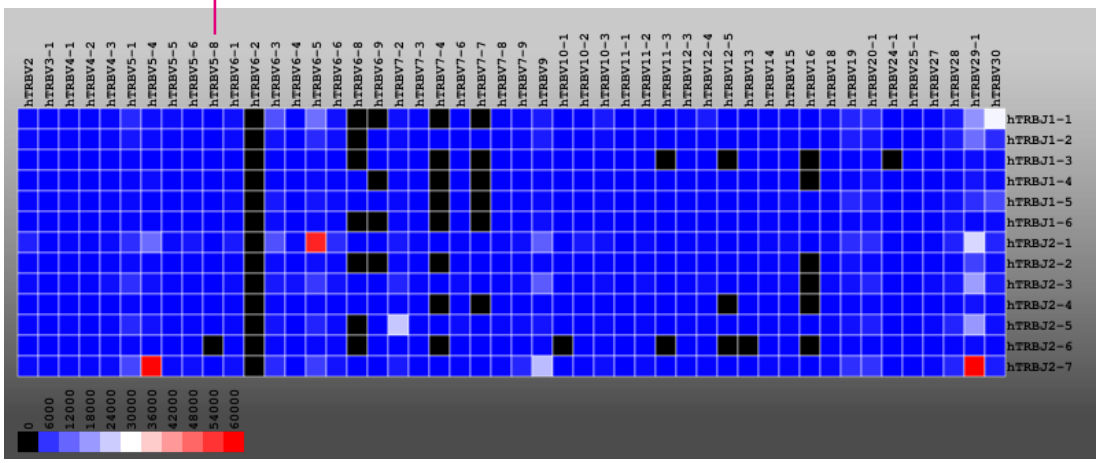
Analysis: Show 2D Map

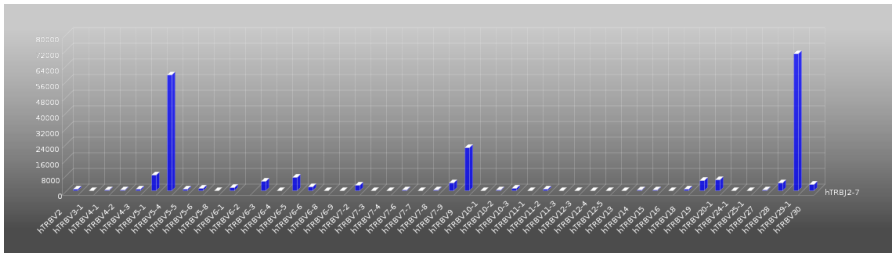
The 2D Map is a heat map of the relative frequencies of given V-gene and J-gene combinations. Along the top x-axis are all of the functional V-genes within IMGT that were used as references. Along the right-hand y-axis are all of the functional J-genes that were used as references. As you can see by the scale used below the 2D Map that a transition from blue to white to red ultimately shows relative frequencies of given gene combinations, with red being the highest frequencies.



Clicking a given V-gene or J-gene name will pop out a 3D bar chart of the frequency data for that specific gene combination. All images can be saved as .PNG files via Right-Click and "Save Image As…" or the as a PDF by clicking the download icon ( )

Clicking a given V-J combination will provide the top 10 most frequent alignments for that gene combination (more on the next page).





iRepertoire, Inc.

Clicking a specific V-J gene combination in the 2D Map will generate alignments for that given selection. The alignments are sorted by frequency, but include:

Line 1: A fasta-like header, gene calls (V, D, J, and C), as well as frequency/copy number and the amino acid unique CDR3 associated with the alignment
Line 2: A counter for 10, 50, and 100 nucleotides
Line 3: The sample sequence; In the case of stitched reads the regions of the read specific to either R1 or R2 will be lowercase and at the 3' or 5' regions of the sequence, respectively. Overlapping regions with 100% identity will be in all uppercase characters
Line 4: The germline reference sequence

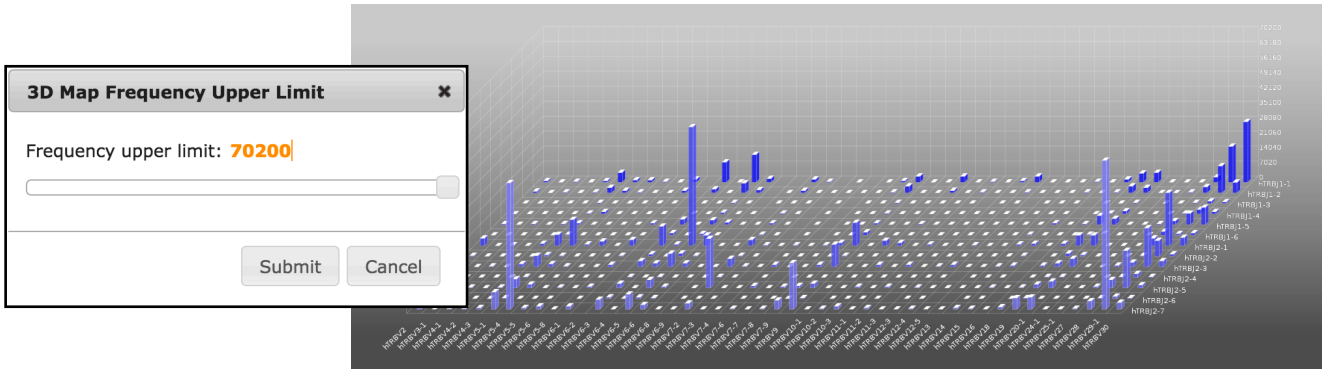
```
>M01518:63:000000000-AL2BP:1:1109:21713:25713 COPY:55879 hTRBV5-4*01[0] hTRBD2*02 hTRBJ2-7*01[3] hTRBC2*02 ASSLGLAGYYEQY
 N  T  R  G  Q  Q  V  T  L  R  C  S  S  Q  S  G  H  N  T  V  S  W  Y  Q  Q  A  L  G  Q  G  P  Q  F  I  F  Q  Y  Y  R  E
          .        .        .        .        †        .        .        .        .        ‡        .        .
aacacgagaggacagcaagtgactctgagatgctcttctcagtctgggcacaacactgtgtcctggtaccaacaggccctgggtcaggggccccagtttatctttcagtattataggGAG
   ACGAGAGGACAGCAAGTGACTCTGAGATGCTCTTCTCAGTCTGGGCACAACACTGTGTCCTGGTACCAACAGGCCCTGGGTCAGGGGCCCCAGTTTATCTTTCAGTATTATAGGGAG

 E  E  N  G  R  G  N  F  P  P  R  F  S  G  L  Q  F  P  N  Y  S  S  E  L  N  V  N  A  L  E  L  D  D  S  A  L  Y  L  C  A
          .        .        †        .        .        .        .        .        ‡        .        .        .        .
GAAGAGAATGGCAGAGGAAACTTCCCTCCTAGATTCTCAGGTCTCCAGTTCCCTAATTATAGCTCTGAGCTGAATGTGAACGCCTTGGagctggacgactcggccctgtatctctgtgcc
GAAGAGAATGGCAGAGGAAACTTCCCTCCTAGATTCTCAGGTCTCCAGTTCCCTAATTATAGCTCTGAGCTGAATGTGAACGCCTTGGAGCTGGACGACTCGGCCCTGTATCTCTGTGCC

 S  S  L  G  L  A  G  Y  Y  E  Q  Y  F  G  P  G  T  R  L  T  V  T  E  D  L  K  N  V  F  P  P  E  V  A  V  F  E  P  S
          †        .        .        .        .        ‡        .        .        .        .        .        †
agcagcttgggactagcgggatactacgagcagtacttcgggccgggcaccaggctcacggtcacagaggacctgaaaaacgtgttcccacccgaggtcgctgtgtttgagccatcaga
AGCAGCTTGGGACTAGCGGGA  CTACGAGCAGTACTTCGGGCCGGGCACCAGGCTCACGGTCACAGAGGACCTGAAAAACGTGTTCCCACCCGAGGTCGCTGTGTTTGAGCCATCAGA
```

Note: In an effort to preserve somatic hypermutation in BCR sequences, fewer of our proprietary filters are used. Mutations from germline are highlighted in red within the sample sequence. Though exceptionally rare, TCR sequences that differ from the germline references will be indicated with red as well.

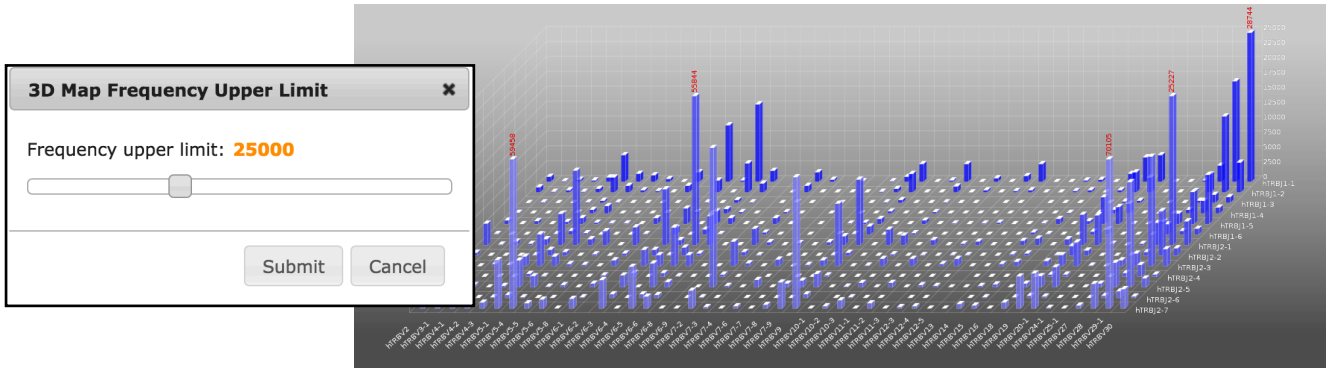
iRepertoire, Inc.

Analysis: Show 3D Map

Show 3D Map provides a 3D bar chart of the same information that is available in the 2D Map; It is another way to visualized the relative frequencies of given V- and J-gene combinations within a sample. Once this Analysis is clicked, a menu will pop up, asking about the upper limit you wish to gate with for your plot. By default, there is a rounded-up value for the highest frequency V-J gene combination.



The Frequency upper limit can be manually typed in or the bar can be dragged to set the upper limit. All values above this upper limit will be indicated in red numbers above those capped columns. It is often best to visualize this graphic by Right-Click and “View Image”.

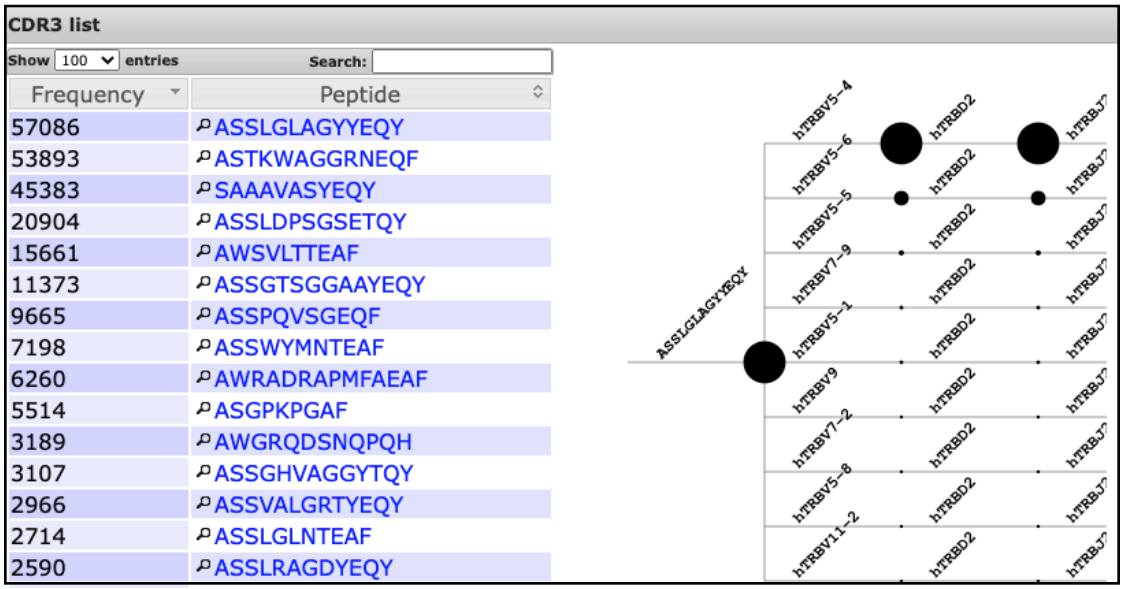


iRepertoire, Inc.

Analysis: List CDR3 new

List CDR3 new provides a simple table of frequencies for amino acid CDR3s, as well as a broken down hierarchical tree for the V, D, J, and C-gene combinations that went in to making a given amino acid CDR3. The Search function within the window can be used to look for a specific unique CDR3 within the sample.

Note: The magnifying glasses have the functionality of being a CDR3 search. Clicking the magnifying glass should open a window that will allow you to search within your own available data for the same unique CDR3.



Though the window containing hierarchical tree can be scrolled through, it is often easiest to view by Right-Click and "View Image".

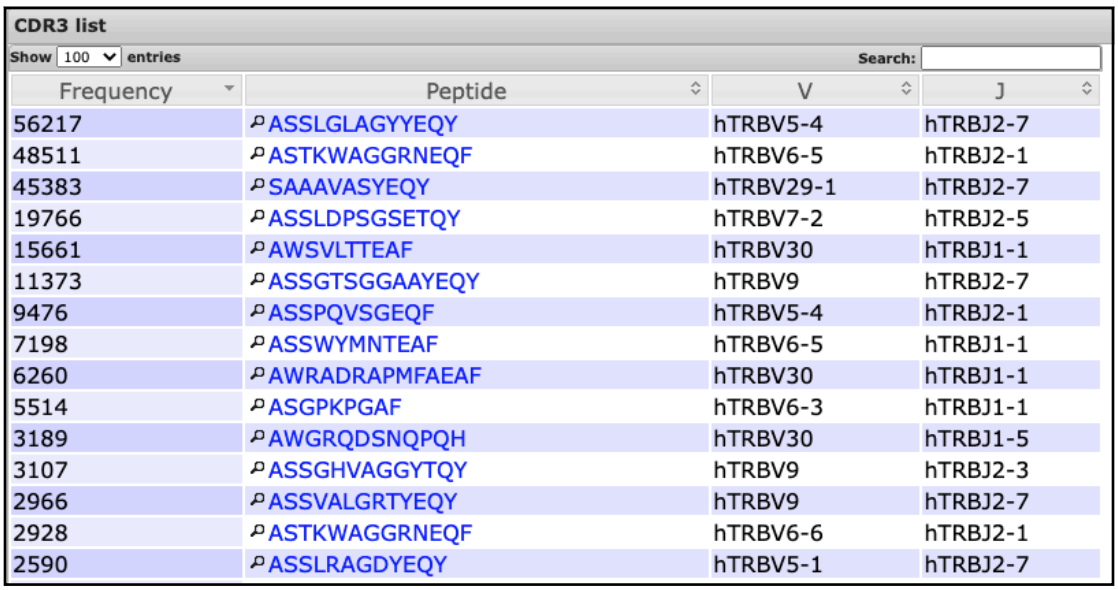
Clicking any of the red dots on the far right, while inside the window, will give you the top ten most frequent alignments. This is a sorted set of alignments from the highest frequency to the lowest. It is important to note that these alignments are not comprehensive.

iRepertoire, Inc.

Analysis: List CDR3 old

List CDR3 old provides a simple table of frequencies for V-J-CDR3aa combinations. It is important to note that the frequencies here will be slightly different than those of the List CDR3 new, simply because the List CDR3 old breaks down the listing of amino acid CDR3s into subcategories or combinations of V- and J-gene usage as well. As with the List CDR3 new, clicking the magnifying glass icon next to any of the CDR3 sequences will open up a CDR3 Search.

The Search function within the window can be used to look for a specific unique CDR3 within the sample.



CDR3 list

Show 100 entries Search:

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	AWSVLTTEAF	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	AWGRQDSNQPQH	hTRBV30	hTRBJ1-5
3107	ASSGHVAGGYTQY	hTRBV9	hTRBJ2-3
2966	ASSVALGRTYEQY	hTRBV9	hTRBJ2-7
2928	ASTKWAGGRNEQF	hTRBV6-6	hTRBJ2-1
2590	ASSLRAGDYEQY	hTRBV5-1	hTRBJ2-7

Clicking any of the blue CDR3aa sequences will provide a list of top ten most frequent alignments, as described previously.

Analysis: List CDRs

Where possible, List CDRs can be used to look at the amino acid sequences for CDR1s, CDR2s, and CDR3s within the sample. It is important to note that our short-read primer systems will only capture sequence information for the CDR3, so in instances of use of short-read primer systems, this data will be limited. Our mouse long-read systems will capture only CDR2 and CDR3.

iRepertoire, Inc.

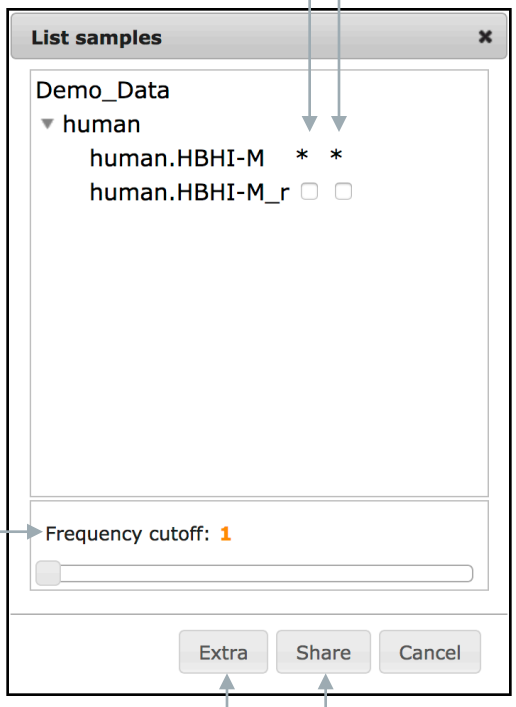
Analysis: CDR3 Algebra

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 below. 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.

"Inclusion" selection; Select this option to discover CDR3s that are shared.

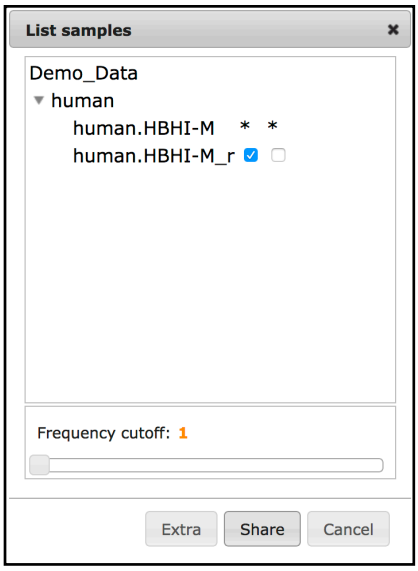
"Exclusion" selection; More on the next page. Will remove, or exclude, CDR3s that are shared.



The Frequency cutoff will allow for removal of unique CDR3s under a determined frequency threshold. The bar can be dragged or values can be typed in orange.

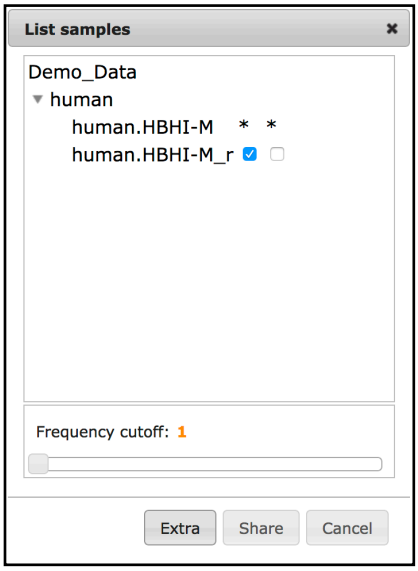
Clicking "Extra" instead of "Share" will provide the top 100 unique CDR3s for all samples in the comparison, regardless of sharing.

Click "Share" to begin the sharing calculation or "Extra" for other functionality as described to the left.



Clicking “Share” once your samples have been chosen will give you a list of only those unique CDR3s that are shared.

	A	B	C
1	CDR3	human.HBHI-M_r	human.HBHI-M
2	AKTTYYYDSSGYQTPYYFDY	60499	53690
3	ARHGQQLALA	13608	12266
4	ARSAAVVATAFTWRSYKGMDV	12754	12850
5	ARILKDSSGWYHFDY	11932	9239
6	ARGAYSSNYARIDD	9422	8904
7	ASLVGTGKDY	7544	3725
8	ARDGTYFGS	7410	4302
9	VKSAQYCDSNSCWRGYYSYYLDV	6939	5178
10	ARGRGYRESYYAFDI	6393	6489
11	ATDQPGFGFEV	6272	8569
12	ARGPWESWNYPEADYSFDY	6245	4031
13	ARRVPEPTGHKYYFDY	6212	5371



Clicking “Extra” once your samples have been chosen will give you a list of all those unique CDR3s that are shared, as well as the top 100 (in terms of frequency) regardless of sharing.

	A	B	C
1	CDR3	human.HBHI-M_r	human.HBHI-M
2	AKGGVRGWNWVDP	3560	
3	ARVCAAGSCFRAFDI	2914	
4	VRARTVVPNLLDY	2645	
5	ARGMYQLVSSAFDP		2835
6	ARGPGLAAGKRYLDY		2507
7	YTTAVVARDY	6	7
8	YATAVVARDY	1070	1154
9	WSGWFGGY	6	7
10	WRDPTDSSGWYPPNDVFDI	121	78
11	WRASAPEPYDSPTYSDT	20	14
12	WGGWSGGY	6	14
13	WGGWFGSY	6	7
14	WGGWFGGY	2900	2272

With CDR3 algebra, there is also an “exclusion” 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 Share.

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.

Important: All CDR3 Algebra calculations are performed under normalized conditions. These normalized conditions scale all frequencies as if there were 10 million reads available for each sample. In effect, this serves as looking at each shared clonotype’s frequency as a percentage of its original repertoire prior to comparison.

As an example:
Let’s say that a given unique CDR3, ASSLGLAGYYEQY, has a frequency of 57,086 in a sample that has a total of 832,285 total reads. Provided that this particular CDR3 is shared with another sample in the CDR3 algebra calculation, it will appear as a frequency ~686,000 in the output .csv as it the original frequency had it as ~6.86% of the original sample (6.86% of 10M reads).

iRepertoire, Inc.

Analysis: D50

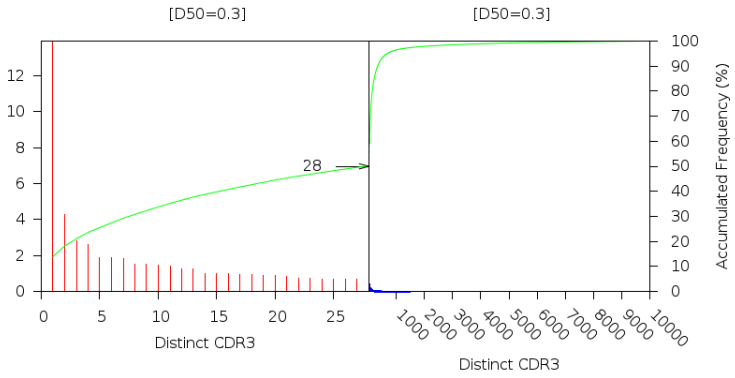
In order to describe and compare the relative diversity of libraries, we have developed a proprietary analysis, termed D50, which assigns a single value that defines the diversity of a library. The D50 is a quantitative measure of the degree of diversity of T cells or B cells within a sample. The D50 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:

Samples with greater than 10,000 unique CDR3s

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 (or 0.3)

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



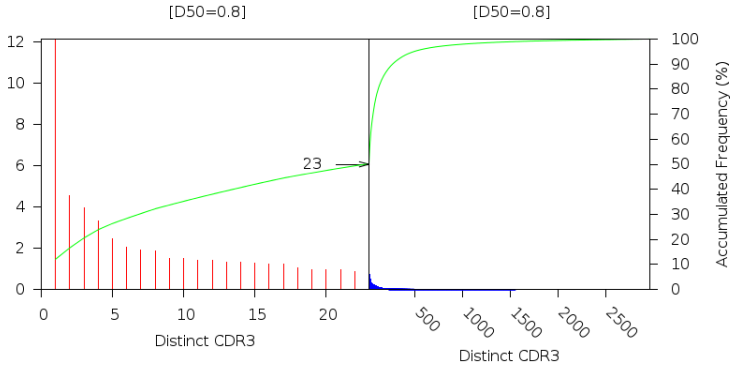
Reads: 2943775
total CDR3: 2920471
distinct CDR3: 14971
D50: 0.3

Samples with fewer than 10,000 unique CDR3s

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 (or 0.8)

(No. of uCDR3s that make up 50% of the total reads *100) / No. of uCDR3s = D50



Reads: 356546
total CDR3: 352860
distinct CDR3: 2963
D50: 0.8

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

Left-hand panel

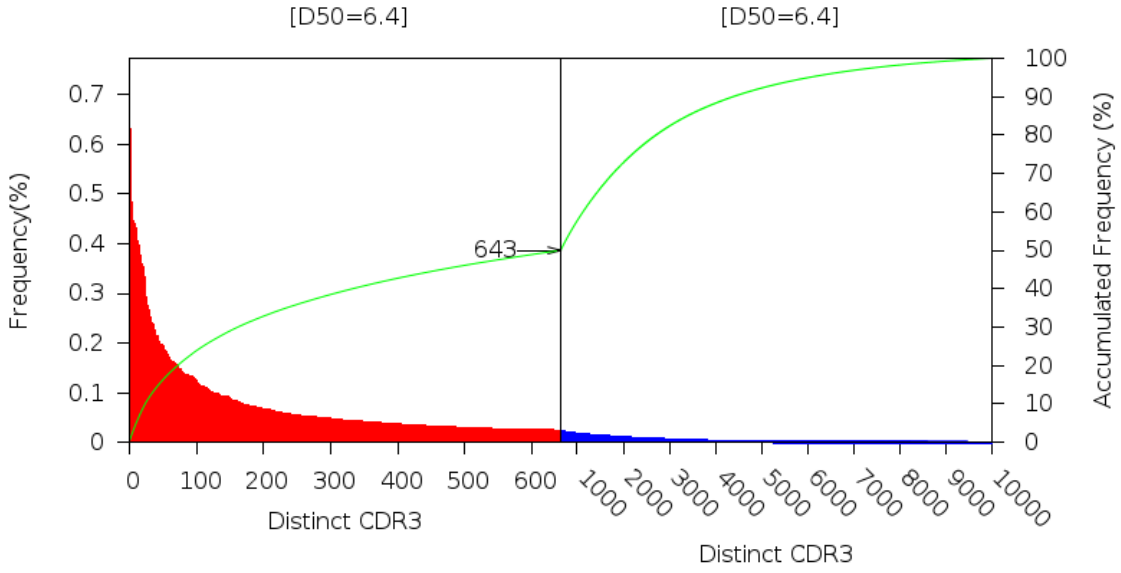
The red bars represent the % of the number of reads for the top uCDR3s for 50% of the top 10k uCDR3s that make up all of the top 10k uCDR3s. In the case of Figure 16, the peptide ARDPSSGWYGDDY (not pictured) is the most dominant clone with a frequency of 6,727. When this frequency is counted in relation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



iRepertoire, Inc.

Analysis: Diversity Index

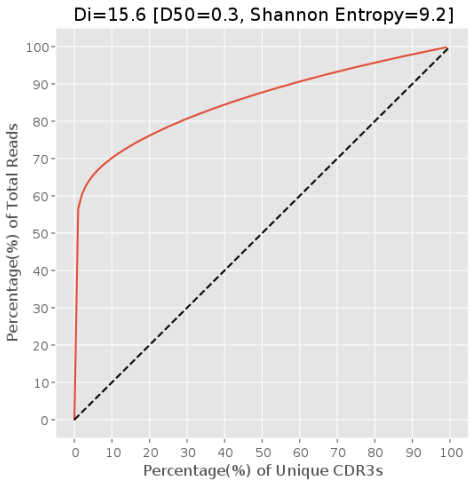
iRepertoire, Inc. has two proprietary diversity indices available via iRweb: D50 and the Diversity Index (Di). The Diversity Index, or Di, is officially defined as:

100 minus the area under the curve between the percentage of total reads and the percentage of unique CDR3s, when unique CDR3s are sorted by frequency from largest to smallest.

Assume that $\underbrace{r_1 \geq r_2, \cdots \geq r_i \geq r_{i+1} \cdots \geq r_n}_{n}$ where r_i is the frequency of the *i-th* CDR3 and n is the total number of unique CDR3s

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

The line created by these calculations assembles a red curve that describes the overall diversity of the sample. The dotted black line with a slope of 1 is representative of an imaginary sample in which all unique CDR3s exist at the same frequency. The area under this curve is used for the calculation of the Diversity Index, i.e., 100 - area under the curve.



Important: The Diversity Index (Di) is calculated one of two ways: If a sample has fewer than ten-thousand unique CDR3s, all unique CDR3s are used. If the sample has more than ten-thousand unique CDR3s, only the top 10k unique CDR3s (and their respective read counts) are used.

iRepertoire, Inc.

Analysis: Shannon Entropy

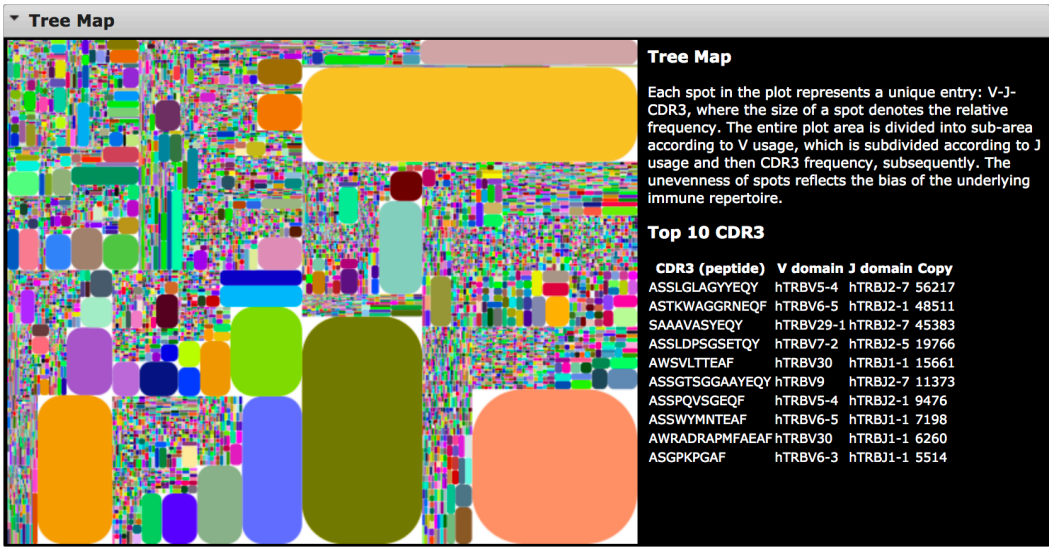
Shannon Entropy (often represented as “Entropy” on iRweb) is calculated in a similar fashion to D50 and Diversity Index. In instances of samples that have greater than ten-thousand unique CDR3s, only the top ten-thousand are unique CDR3s are used in the calculation. With samples that have fewer than ten-thousand unique CDR3s, all of them are used in the calculation. The formula are included below for either calculation, but pi corresponds to the probability of a given unique CDR3 (i.e., frequency of uCDR3/total number of reads).

$$-\sum_{i=1}^{n} p_i \ log_2 p_i$$

$$-\sum_{i=1}^{10000} p_i \ log_2 p_i$$

Analysis: Tree Map

The 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. 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.



iRepertoire, Inc.

Distribution Analyses

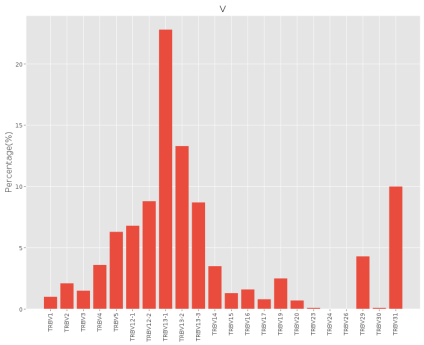
The software also provides several types of distribution analysis including V-usage, J-usage, V-trimming, J-trimming, CDR3 length, and N-addition. The same analyses are also provided as normalized distributions. The difference between the regular distribution and normalized distribution is how the data are counted. The regular distribution is based on the number directly observed from the read count data.

Distribution Analyses: Normalized

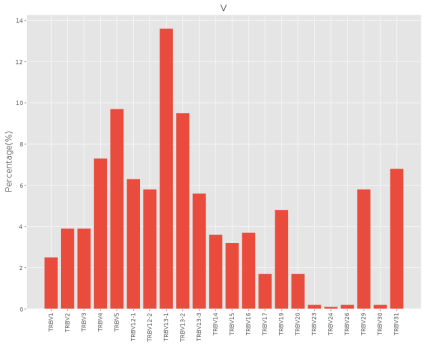
The normalized distribution counts the value (for V, J, N-addition, CDR3 length, etc.) of each distinct CDR3 as one, no matter how many of the particular CDR3s are observed. In short, each uCDR3-VDJ combination is treated as a quantity of 1 regardless of read count, and then analyzed for V usage, J usage, 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

The bar chart provided by V usage under Distribution includes all of the relative frequencies of a given V-gene. Below are the Distribution and Normalized versions of the V usage charts for the MTBI-M sample within the demo data set.



Distribution
The y-axis has mTRBV13-1 as upwards of 24% of the total reads within the sample.

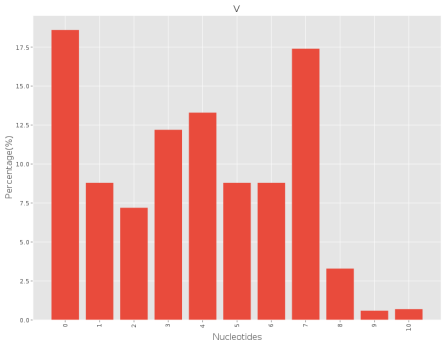


Normalized
The y-axis has been revamped, with mTRBV13-1 now representing ~13.5% of the sample. This is as if the usage of each V-gene counted only as 1 and is not based on frequency.

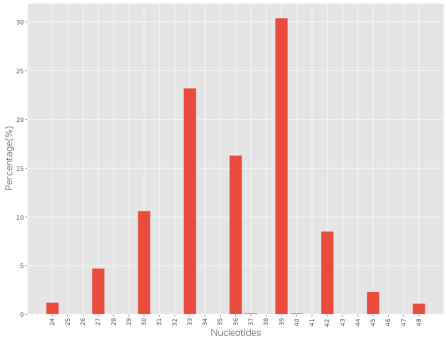
iRepertoire, Inc.

V and J trimming

The recombination events that generate the final immune receptor will, in some cases, “nibble” or “trim” from the genes used to produce the receptor. Using our germline reference sequences (via IMGT), we can show summaries about trimming from the 3’ end of the V-gene used and the 5’ end of the J-gene used, depending on the coverage within the reference sequence.

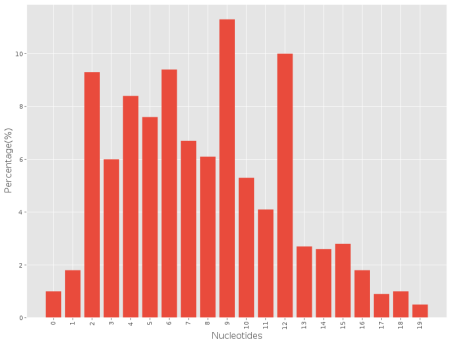


The percentages used here are reflective of the total number of reads within the sample. In this example, just over 17.5% of the sample has zero nucleotides trimmed from the end of the V-gene.



CDR3 length

CDR3 length provides the percent frequencies of the lengths of all unique CDR3s at a nucleotide level. In most instances, a distinct triplet-spacing can be observed, and this is tied to the triplet nucleotide codons that produce each amino acid within the final CDR3 sequence.



N addition

Nucleotide addition during V(D)J Recombination is calculated, similar to V and J trimming, by comparisons to the germline reference seqeunces.

Important: With chains that include D-genes (TCRb, TCRd, IgH), it is known that possible tandem D-gene usage is possible - i.e., multiple D-genes back-to-back (citation). Our pipeline will perform the a best match alignment for only one D-gene and then will calculate all other nucleotides in the gaps between V- or J-genes as N addition.

iRepertoire, Inc.

Raw Data

Raw Data contains all of the documents necessary to replicate most of 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 via clicking the Raw Data button, or by clicking the green download icon next to the Project Name. With either method of access, iRweb will output a blue message that "Preparing raw data for project[PROJECTNAME], will refresh in a minute". The window may refresh multiple times while this is occurring, so please do not click Back in your browser. Once ready, a .zip containing 17 files per Sample ID should be available for download.

Please see more details about the contents of the Raw Data download below.

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).

diversity	Contains all of the diversity measurements for the sample (D50), Di, and Shannon Entropy)
######_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.

iRepertoire, Inc.

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.

iRepertoire, Inc.

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 (Psuedogene) are not covered.

Is subisotype information available?

Isotype information is available for all Ig samples. Subisotype 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 Data Support (datasupport@irepertoire.com) 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 under our Document Center. If this form is unavailable in your country or otherwise inaccessible, please contact Customer Service about getting an Excel verison 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 files names and Study Names will delay return of your analyzed data. Data can be uploaded via SFTP (please contact Data Support for a login), shared with info@irepertoire.com 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.

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.

iRepertoire, Inc.

Terms & Definitions

Frequency
In practice, this number is generated by the collapse/condense stage of our pipeline. Nucleotide-level sequencing reads are collapsed based on identity and with each new line added (effectively, a read) the frequency increases by 1. A frequency of 450 means that there were 450 lines of sequencing data that shared the same nucleotide sequence and unique CDR3. This frequency would correlate to the performance of the amplification and to the levels of original RNA template in the original sample.

Copy number
Please use the description above for Frequency. The two values are interchangeable on iRweb.

Cosmic Effect
A rule of thumb applied in-house at iRepertoire. Provided two identical or exceptionally similar unique CDR3s, if the ratio of the lower frequency uCDR3 is less than or equal to 1:500 with respect to the higher frequency CDR3, the lower is disregarded as noise from either sequencing error or PCR error.

Clone(s)
Oftentimes, the use of “clone”, “unique CDR3”, or “uCDR3” stand in for the same concept: a given amino acid CDR3 with a frequency of X. To say a sample is “very diverse” is to describe it as having high numbers of “unique CDR3s” or “clones”.

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