

VDJbase: Documentation

Introduction

We present a new database application that stores genotype and haplotype data inferred from adaptive immune receptor repertoire sequencing (AIRR-seq) datasets. The input is a pre-processed dataset. We align each of the pre-processed datasets, detect novel allele and infer genotypes and haplotypes. Our application allows users to query and filter data according to selectable criteria. It is also possible to plot the data according to interest terms such as allele, gene, heterozygosity, inferred genotype, and haplotype. Users can freely access the database using any web browser and see the results of their queries immediately. The results are displayed as a table along with samples and their related metadata which provides files and figures that can also be downloaded to the user's own computer. We used the database application to perform a responsive analysis. The analysis allows for easy comparison between samples to explore the genetic predispositions across the population. See figure 1 for a schematic diagram of VDJbase.

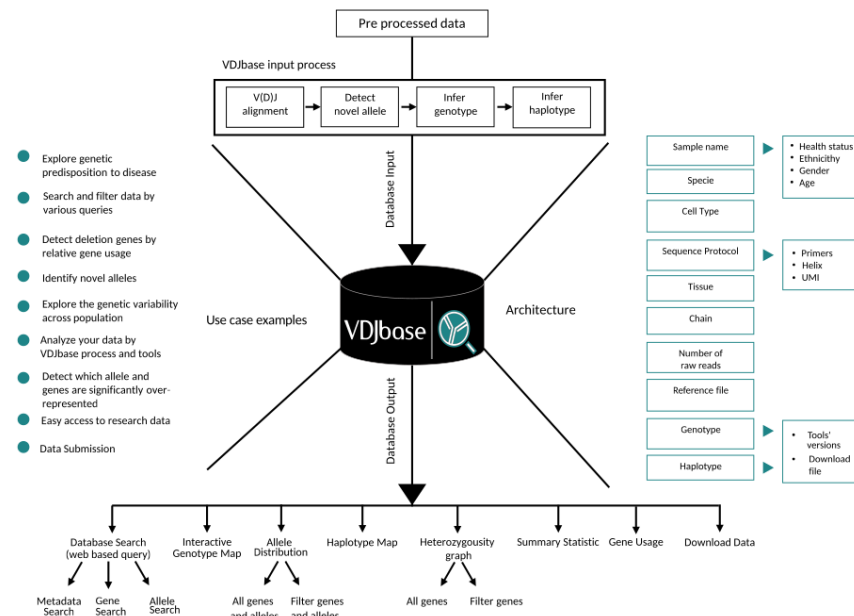


Figure 1 : Schematic diagram of VDJbase

Implementation

Filtering options in VDJbase

Data

Users can select samples by checking the square next to each sample.

The search section is intended to filter the samples according to selectable criteria (using ctrl for multiple selection), and the graphs are the output of the filtered samples. In case the user does not select specific samples, graphs will be generated for all samples.

Graphs

Users can generate a plot for single individual by the right genotype and haplotype columns. There are three format options: downloadable PDF, interactive graph by HTML, and downloadable summary statistics table file. Graphs for comparing multiple samples are provided by the left black bar. The 'advanced filter' allows filtering out the figures by gene type, pseudogenes, ORF's genes and the certainty level of inference (Kdiff). The defaults are zero Kdiff and pseudogenes are filtered out.

VDJbase

Graphs

Advanced Filters

Gene types

Genes

Kdiff

With pseudo genes

Genotype

Haplotypes

Name:

Tissue:

Cell type:

Helix:

Sequencing length:

Sex:

Subject:

Health status:

Umi:

Minimal no. of reads:

Sequencing platform:

Primer 3 prime:

Primer 5 prime:

Haplotype by:

Rows/Page:

Studies:









Alleles:

Search









Advanced Search

Clear search

Clear Selection

	Name	Species	Chain	Reads	Cell Type	UMI	Genotype	Haplotype
<input type="checkbox"/>	P1_I100_S1	human	Heavy	12231	B-CELL - Naive	✓	   	IGHJ6 IGHJ2-8
<input type="checkbox"/>	P1_I10_S1	human	Heavy	33297	B-CELL - Naive	✓	   	IGHJ6

Click on the sample row to view added information about this sample.

<input type="checkbox"/>	Name	Species	Chain	Reads	Cell Type	UMI	Genotype	Haplotype
<input type="checkbox"/>	P1_I100_S1	human	Heavy	12231	B-CELL - Naive	✓	   	IGHJ6 IGHJ2-8
Study								
Subject								
Sequence Protocols								
Tissue Processing								
Genotype								
<input type="checkbox"/>	P1_I10_S1	human	Heavy	33297	B-CELL - Naive	✓	   	IGHJ6

Allele search (path: "Database Search/Allele")

This page provides a summary collection of the alleles that appear in the database. It provides information about their sequence and indicates the number of samples that contain the allele in their genotype. Clicking on this number opens a new page with these samples.

VDJbase

Home

Database Search

Explore Data

User Guide

About

Name:

Gene:

Identical sequence:

Search

Clear

Name	Sequence	Sequence length	Identical sequence	Appeared in population
IGHV1-18*01	cagggttcagctgggtgcagctctggagct...gagggtgaagaagcctggggc ctcagtgaaagctctcctgcgaagcctctgtgtacacottt..... ..accagctatggtatcagctgggtgcacaggccctggacaaggcctt gagtggtatgggatggatcagcgtttac.....aatggttaacacaaacta tgacacagaagctccag...ggcagagtcacccatgaccacagacacatcca ogagcacagcctacatggagctgaggagcctgagatctgacgacacaggcc gtgtattactgtgogagaga	320		306
IGHV1-18*01_a190g		320		2
IGHV1-18*01_a196g		320		1
IGHV1-18*01_a318g		320		1
IGHV1-18*01_c291g		320		2

Data visualizations in VDJbase

Keywords

Allele and gene annotations in figures

Unk: In genotype graph, if a gene does not appear in a specific genotype sample, they are marked as unknown(Unk).

In haplotype inference, If the evidence for a gene is not strong enough (K is lower than a certain threshold (set to 1000 by default)), this allele is set to “Unk”.

Del:Single or double chromosome deletion. Single chromosome deletion, deletion polymorphism, is inferred from RABHIT haplotype function, where the K threshold for an Unknown assignment of a gene in a given chromosome is larger than a certain threshold (by default 1000). Double chromosome deletion is homozygosity for the deletion event, this is inferred using RABHIT binomial test.

In haplotype graphs both deletion events are present, however, in genotype graphs only double chromosome deletion is present.

NR: A non reliable gene (NR) is defined when the ratio of the multiple assignments with a gene in a haplotype is below the threshold. We use the RabHIT ratio threshold.

NRA: A non-reliable allele is a term for an ambiguous call. We collapse ambiguous allele using the RabHIT reliability scores. Each allele for which more than 60% of alignments are ambiguous calls are marked as unreliable (NRA), and later collapsed. For example, if IGHV3-23 allele 01 frequently appeared in ambiguous calls with allele 02 in more than 60% of the sequences, then the alignments are said to be 01_02 allele. The annotation is ordered numerically, for example IGHV2-70*01_10_11_13_15 will not give credence to *10.

NRA marked genes are also designated '[*01]' in figures, and the full name is displayed at the bottom of the figure.

Certainty level of genotype and haplotype inference

Genotype Kdiff : The resulting genotypes include a measure of certainty of the genotype call for each gene (Kdiff). The larger the Kdiff, the greater the certainty of the allele chromosomal inference.

Haplotype Kdiff: The log of the Bayes factor (K) obtained from the haplotype Bayesian inference for each allele for a given gene. The larger the K, the greater the certainty of the haplotype inference.

VDJbase currently provides the following analysis outputs:

Genotype graph (Single or list)

Genotype displays a graphical analysis of the alleles observed in selected samples. The left blue column represents the certainty level in log scale ($\log(K)$). If gene does not appear in specific genotype samples, compared to other samples, their allele is set to “unknown”(Unk). ‘Del’ represents a double chromosome deletion of this gene, and ‘NRA’ represents non-reliable allele. Genotype interactive graph allows the users to modify parameters to explore the genotype data according to their interests. For instance, users can focus on specific alleles or screen the results by their certainty level(Kdiff parameter). The graph can visualize 1-20 genotypes to allow comparison between individuals. Each column in both panels represents a different sample.

Graphs
Advanced Filters ▾
Gene types: ▾
Genes: ▾
Kdiff: ▾
With pseudo genes: ▾

Genotype

Haplotypes

Allele distribution

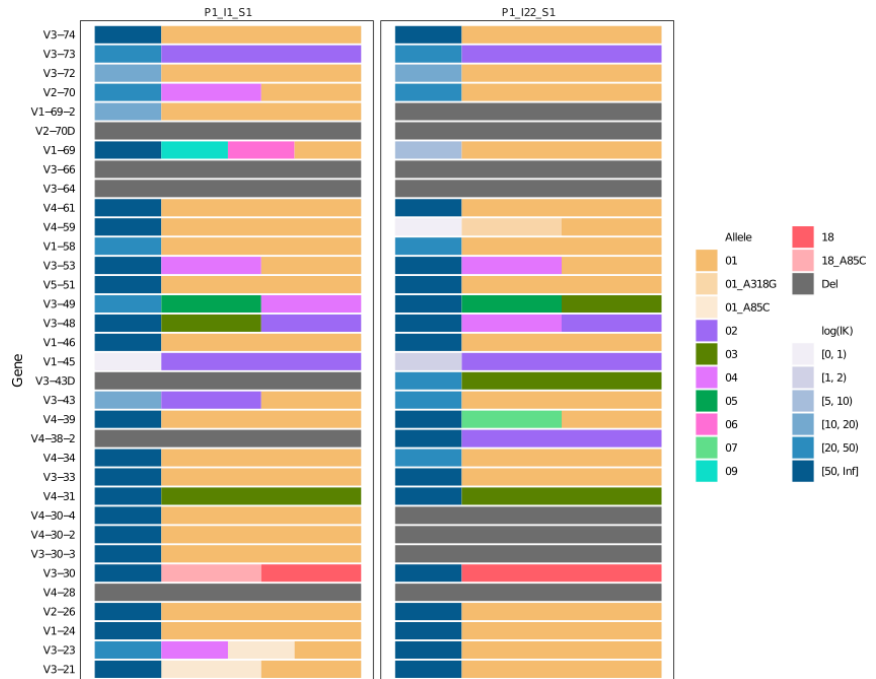
Heterozygous

Allele usage

Gene usage

Data Download
Download Selected

Support Forum



Compare multiple genotype (Heatmap)

Row: each row represents a different sample.

Column: each column represents a different IGH gene.

Colors: correspond to the different alleles.

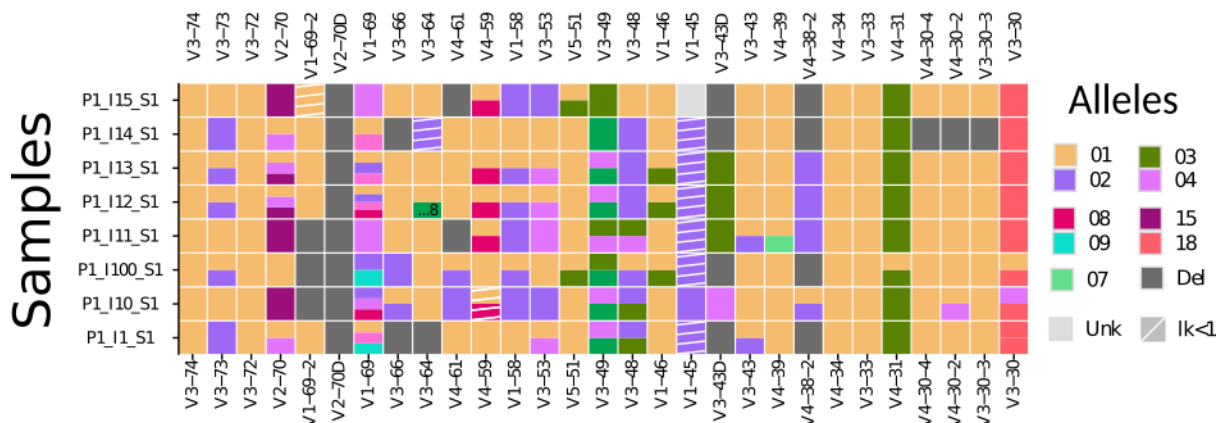
White lines: correspond to a low IK value ($IK < 1$) of the chromosome haplotype.

White ("NRA"): non-reliable allele (ambiguous allele)

Dark gray ("Del"): inferred double chromosome deletion

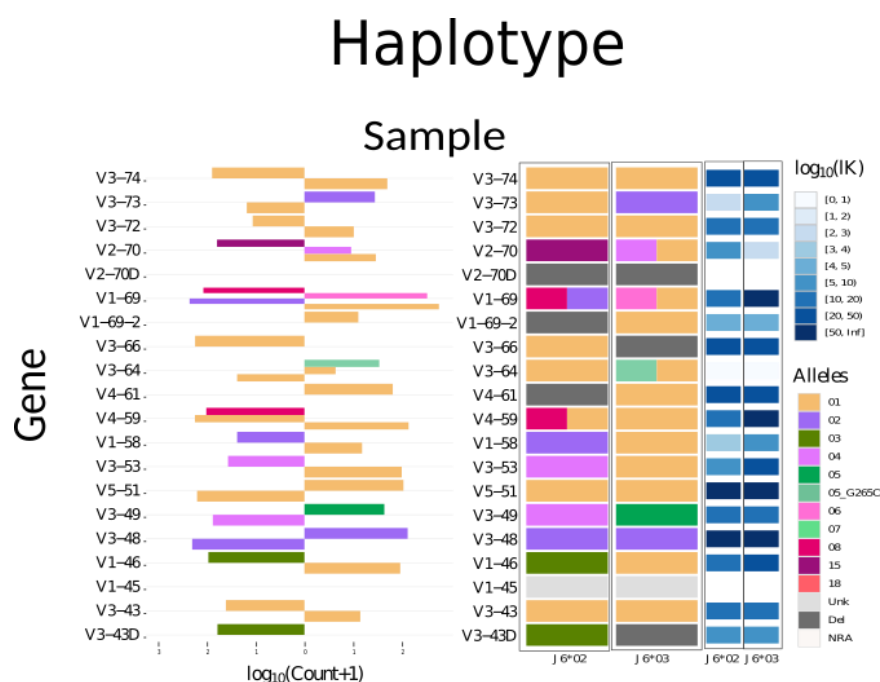
Light-gray("Unk"): unknown

Compare multiple genotypes



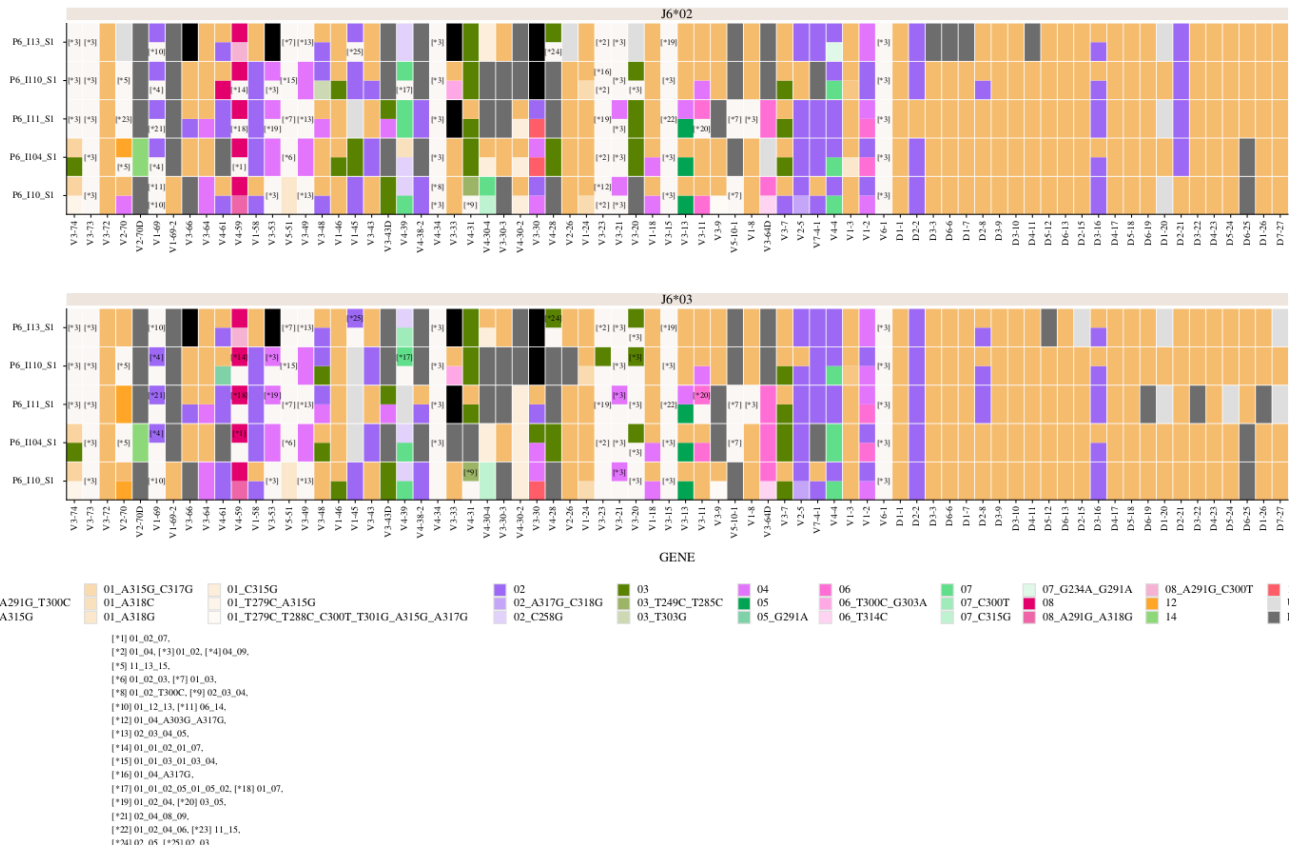
Haplotype (for individual)

Haplotype inference provides an allele map to the individual chromosomes, hence gene and allele assignments to a sequence are very important. The V haplotype by IGHJ6 for individual "sample". Each color represents a different allele, dark gray represent a gene deletion event on a certain chromosome, black represents a non reliable gene (NR), and off-white represents a non reliable allele (NRA). Users can generate individual haplotype by click on the value at the Haplotype column. For example, click on "IGHJ6" to generate haplotype graph for "Sample" by IGHJ6 anchor.



Compare multiple haplotypes

Haplotype must be according to the same anchor for comparison between samples. The upper panel corresponds to IGHJ6*02 and the lower to IGHJ6*03. Each row in both panels represents a different sample. If the evidence is not strong enough (K is lower than a certain threshold (set to 1000 by default)), haplotype inference for this allele is set to "unknown". The colors correspond to the different alleles, dark gray corresponds to inferred double chromosome deletion, black corresponds to non-reliable gene (NR), off-white to NRA, and white lines to a low IK value ($\text{IK} < 1$) of the chromosome haplotype. NRA marked also as e.g. "[*3]", full annotation is displayed at the bottom legends.



Allele distribution graph

This graph represents the allele distribution for each gene. Users can compare the distributions of alleles among different populations. The left Y axis represents the number of individuals, the right Y axis represents the frequency in selected population.



Graphs

Advanced Filters

Gene types

Genes

Kdiff

With pseudo genes

Genotype

Haplotypes

Allele distribution

Heterozygous

Allele usage

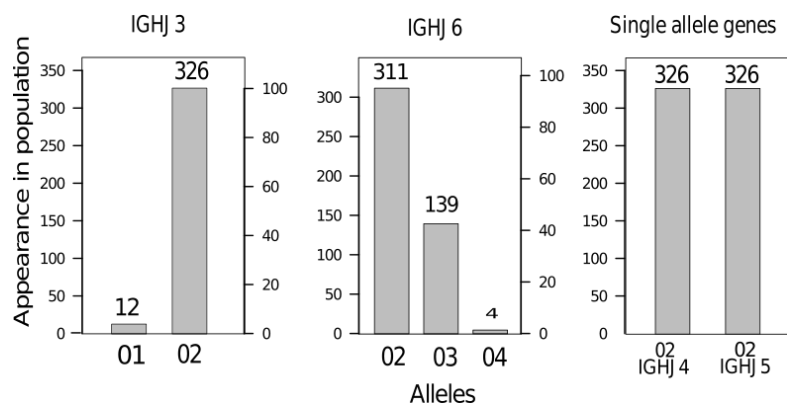
Gene usage

Data Download

Download Selected

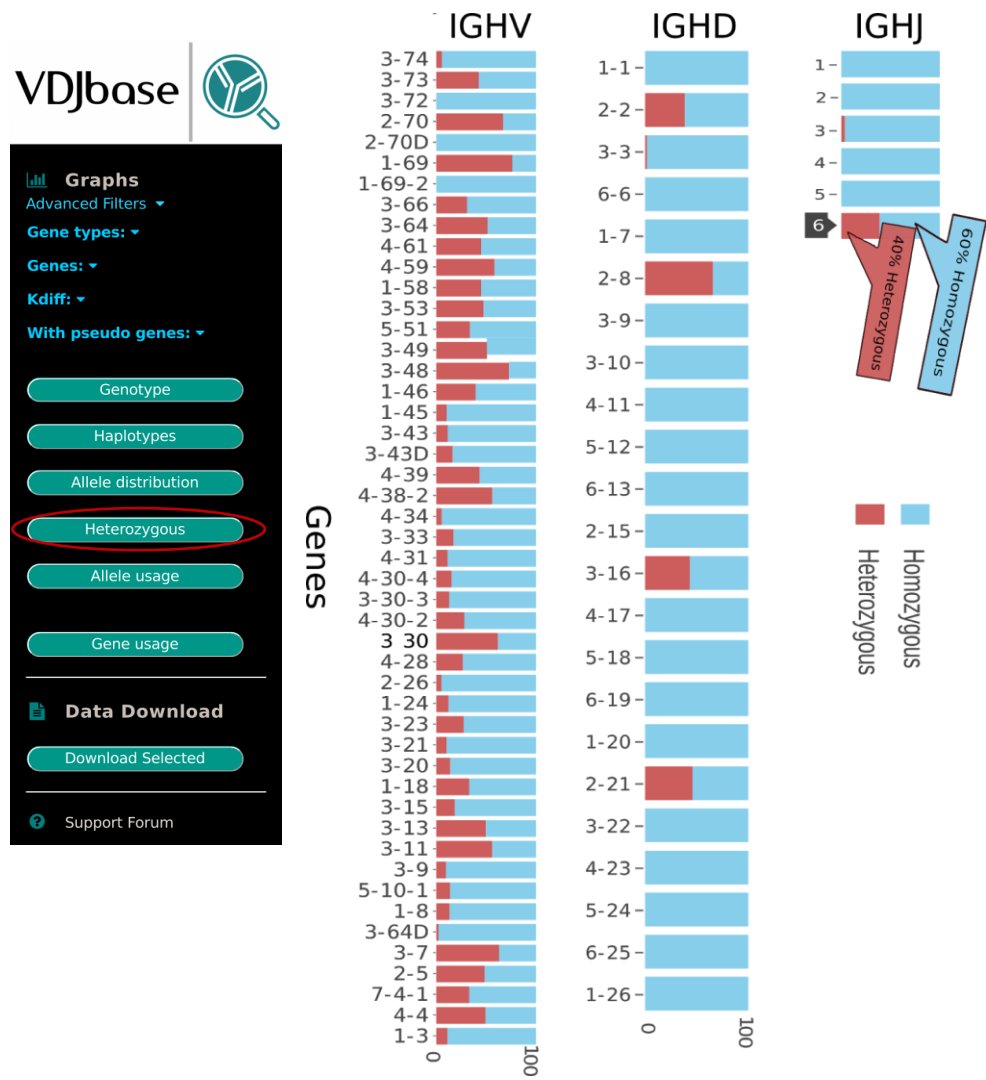
Support Forum

Allele Distribution



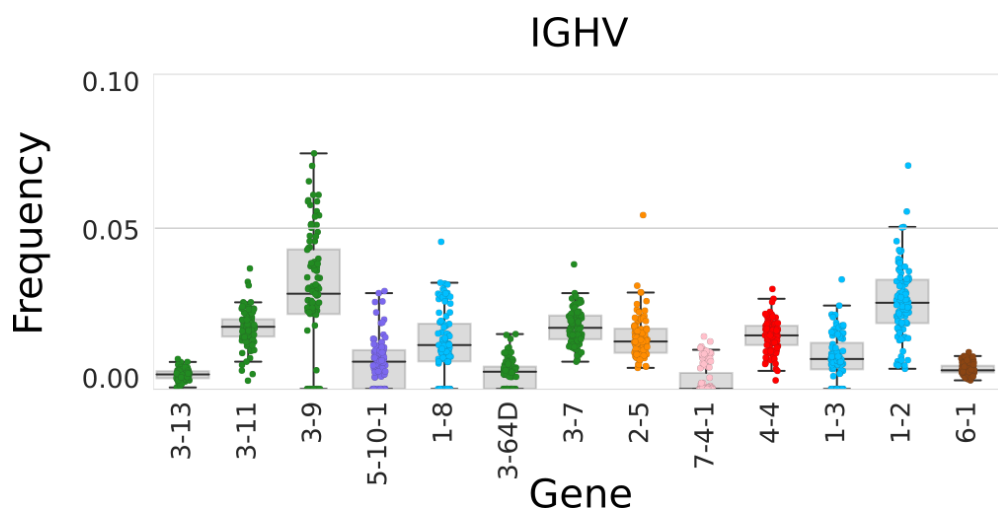
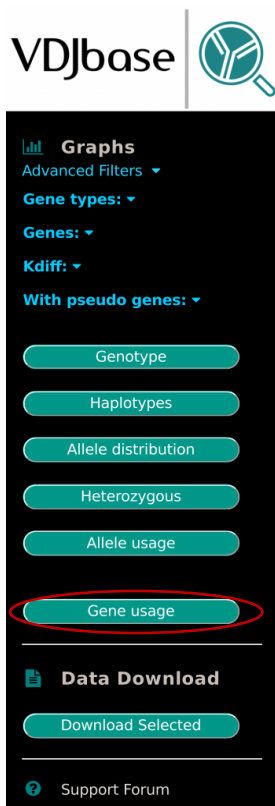
Heterozygote graph

Heterozygote graph allows the user to assess the level of homozygosity/heterozygosity for each gene in different populations. Frequency values appear as pop-ups when users move their mouse on the corresponding bar.



Gene usage

Gene usage provides a view on the gene expression in the population. Each point represents a patient, and color represents the gene's family. The genes ordered by their relative chromosome position.



Materials and methods

Inferred genotypes and haplotypes

We align each of the pre-processed datasets using the most recent version of the IgBlast aligner and the current IMGT germline reference set. We infer previously unknown alleles using TlgGERs inferGenotype function [1] with a modification to the position range input, which allows detection of novel alleles involving sequence variation at nucleotide positions beyond 312 (current default). The sequences are then aligned again, using IgBlast with a germline reference set that is extended to include any novel IGHV alleles inferred by inferGenotype. The output of IgBLAST is converted to the Change-O format [2] that is compliant with the MiAIRR standard [3]. To improve the subsequent quality of allele inference in samples containing highly mutated sequences, we infer clones using SCOPe [4] and choose a single representative, with the lowest number of mutations, for each clone. A genotype is then inferred using TlgGERs new inferGenotypeBayesian function [5], which can detect novel alleles at greater hamming distance from sequences in the provided reference set than previous versions of TlgGER, and assigns a probability, $K_{genotype}$, to each allele in the inferred genotype. The sequences are then aligned for a third time with IgBlast, using a germline reference set that contains only sequences of those alleles included in the personalised genotype created by inferGenotypeBayesian. Lastly, haplotypes are inferred for heterozygous individuals for genes IGHJ6/IGHD2-8/IGHD2-21 using RABHIT [6]. Samples with fewer than 2000 sequences either in the first IgBlast or after the collapsing of clones are excluded from further analysis, and data is not incorporated into VDJbase. For datasets with partial V-region coverage, some modifications (described below) are made to the processing protocol. We consistently update the datasets according to the latest versions of the above mentioned tools, and use version control for the website for reproducibility. Versions are displayed on the site.

IGHV annotation for datasets with partial V-region coverage

Aligners are more likely to make ambiguous calls (for example 'IGHV3-23*01 or IGHV3-23*02') when aligning sequences with partial V-region coverage, and this can influence downstream analyses. To resolve this situation, in these datasets, we collapse ambiguous allele assignments using the RAbHIT reliability scores. Each allele for which more than 60% of alignments are ambiguous calls are marked as non-reliable alleles (NRA), and later collapsed. Where ambiguous calls remain, such as 'IGHV3-23*01 or IGHV3- 23*02', this is indicated by allele designations such as 01 02. We screen for reliable alleles prior to inferring novel alleles and genotypes. We also modify the numbering of the starting position of partial novel inferences to correspond to the implied nucleotide numbers of full length sequences.

Reference

1. Gadala-Maria, D., Yaari, G., Uduman, M., and Kleinstein, S. H. (2015) Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences*, 112(8), E862–E870.
2. Gupta, N. T., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Yaari, G., and Kleinstein, S. H. (06, 2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20), 3356–3358.
3. Breden, F., Luning Prak, E. T., Peters, B., Rubelt, F., Schramm, C. A., Busse, C. E., Vander Heiden, J. A., Christley, S., Bukhari, S. A. C., Thorogood, A., Matsen IV, F. A., Wine, Y., Laserson, U., Klatzmann, D., Douek, D. C., Lefranc, M.-P., Collins, A. M., Bubela, T., Kleinstein, S. H., Watson, C. T., Cowell, L. G., Scott, J. K., and Kepler, T. B. (2017) Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Frontiers in Immunology*, 8, 1418.
4. Nouri, N. and Kleinstein, S. H. (2018) A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. *Bioinformatics*, 34(13), i341–i349.
5. Gadala-Maria, D., Gidoni, M., Marquez, S., Vander Heiden, J. A., Kos, J. T., Watson, C. T., OConnor, K., Yaari, G., and Kleinstein, S. H. (2019) Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Frontiers in immunology*, 10, 129.
6. Peres, A., Gidoni, M., Polak, P., and Yaari, G. (2019) RAbHIT: R Antibody Haplotype Inference Tool. *Bioinformatics*,.