IgIDivA

Laura Zaragoza-Infante

20/5/2022

Introduction

`IgIDivA` (Immunoglobulin Intraclonal Diversification Analysis) is a purpose-built tool for the analysis of the intraclonal diversification process using high-throughput sequencing data.
It is written in shiny. Every step of the analysis can be performed interactively, thus not requiring any programming skills.
It takes as input the output files "clonotypes_computation" and "grouped_alignment_nt" from the `tripr` package.
Functions for an `R` command-line use are also available.

Installation

The `IgIDivA` scripts can be freely downloaded here. It requires `R` (version "4.1"), which can be installed on any operating system [e.g., Linux, Windows, MacOS] from CRAN. Installation with Docker will be available in the coming future.

All the packages that need to be installed in the `R` session are the following:

```
install.packages("shiny")
install.packages("shinyFiles")
install.packages("fs")
install.packages("pdftools")
install.packages("purrr")
install.packages("DT")
install.packages("bslib")
install.packages("shinyhelper")
install.packages("data.table")
install.packages("stringr")
install.packages("RGenetics")
install.packages("dplyr")
install.packages("ggsci")
install.packages("tidygraph")
install.packages("ggraph")
install.packages("igraph")
install.packages("ggplot2")
install.packages("ggpubr")
install.packages("rstatix")
```

All the scripts from `IgIDivA` need to be downloaded in the same folder. All the input files should also be stored in a different folder.

Download an example dataset as Input for `IgIDivA`

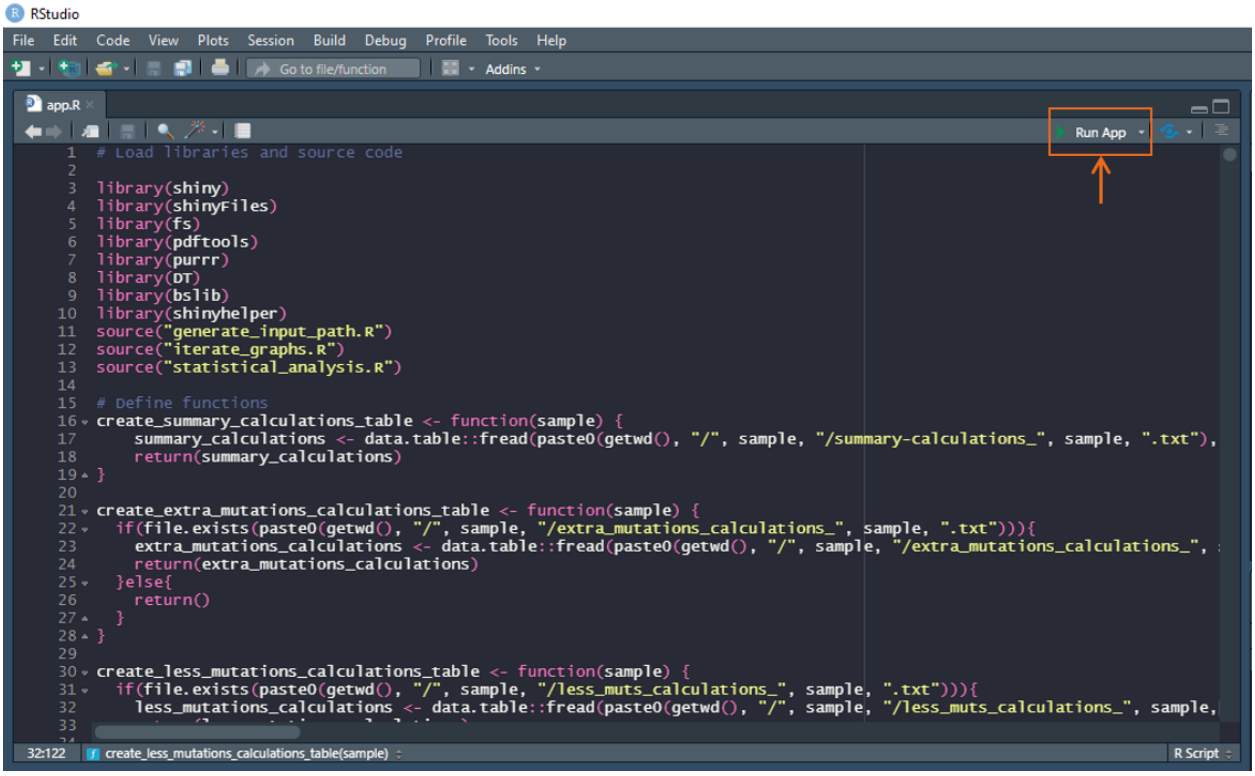
An example dataset to be used as Input for IgIDivA can be found here. The dataset comprises the tripr output files ["highly_sim_all_clonotypes" and "Grouped Alignment_nt] of 26 chronic lymphocytic leukemia (CLL) samples [19 CLL subset #2 samples and 7 CLL subset #169 samples]. The data was retrieved from ENA under the accession number PRJEB36589, and subsequently processed with IMGT/HighV-QUEST and tripr.
Each sample's data can be downloaded by pressing the button `Download`.
Alternatively, to download all the data at the same time, the following commands can be used in the `R` session:

```
install.packages("zen4R")
library(zen4R)
options(timeout = max(600, getOption("timeout")))
path = paste0(getwd(), "/Input")
if (!dir.exists(path)){
  dir.create(path)}
zen4R::download_zenodo('10.5281/zenodo.6616046',  path = path)
```

(The variable "path" can be changed with the location where the user wants to store the Input).

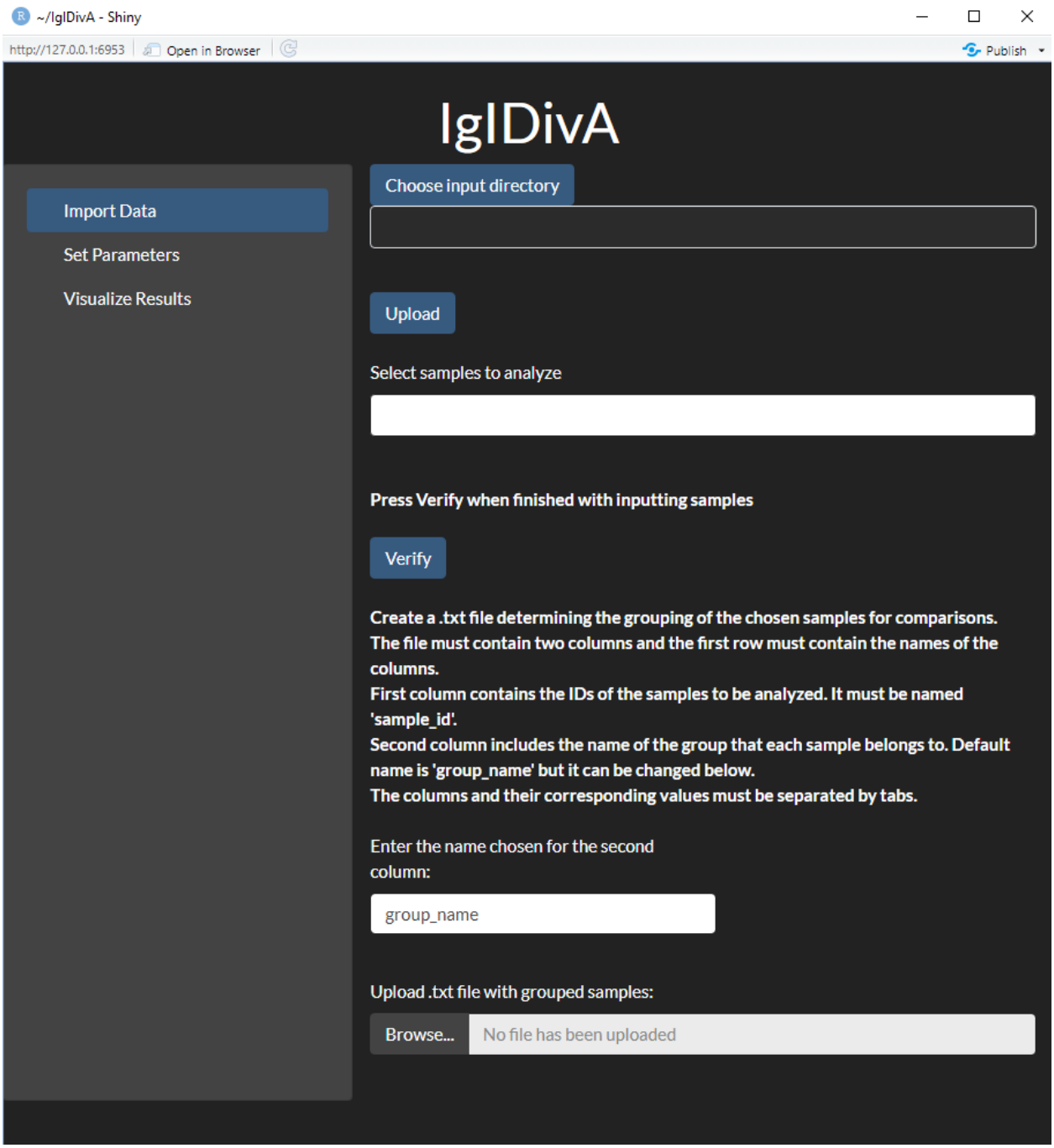
Running `IgIDivA` as a `shiny` application

In order to start the `shiny` app, the script `app.R` should be opened in the `R` session and the button `Run App` should be pressed.

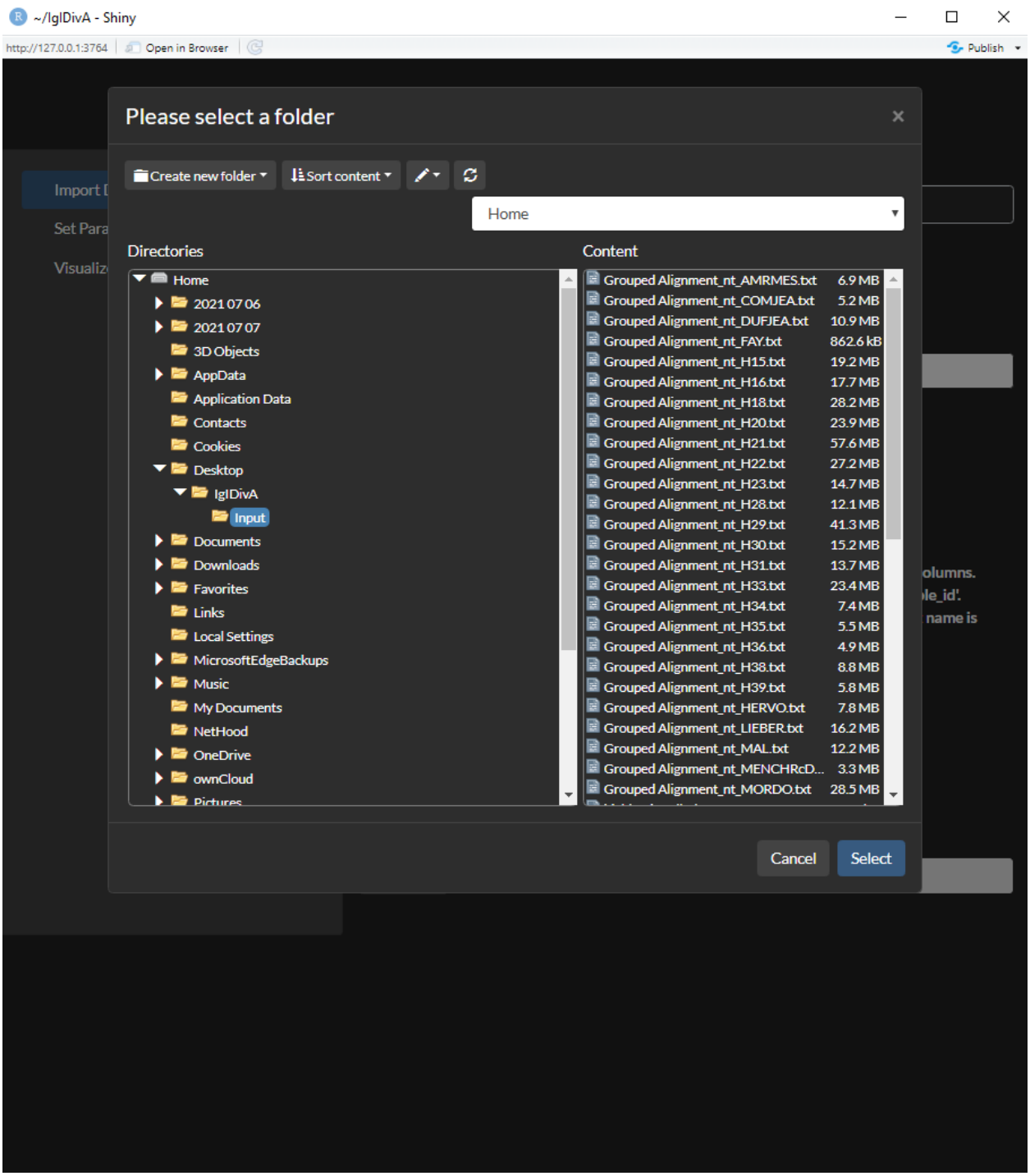


Import data

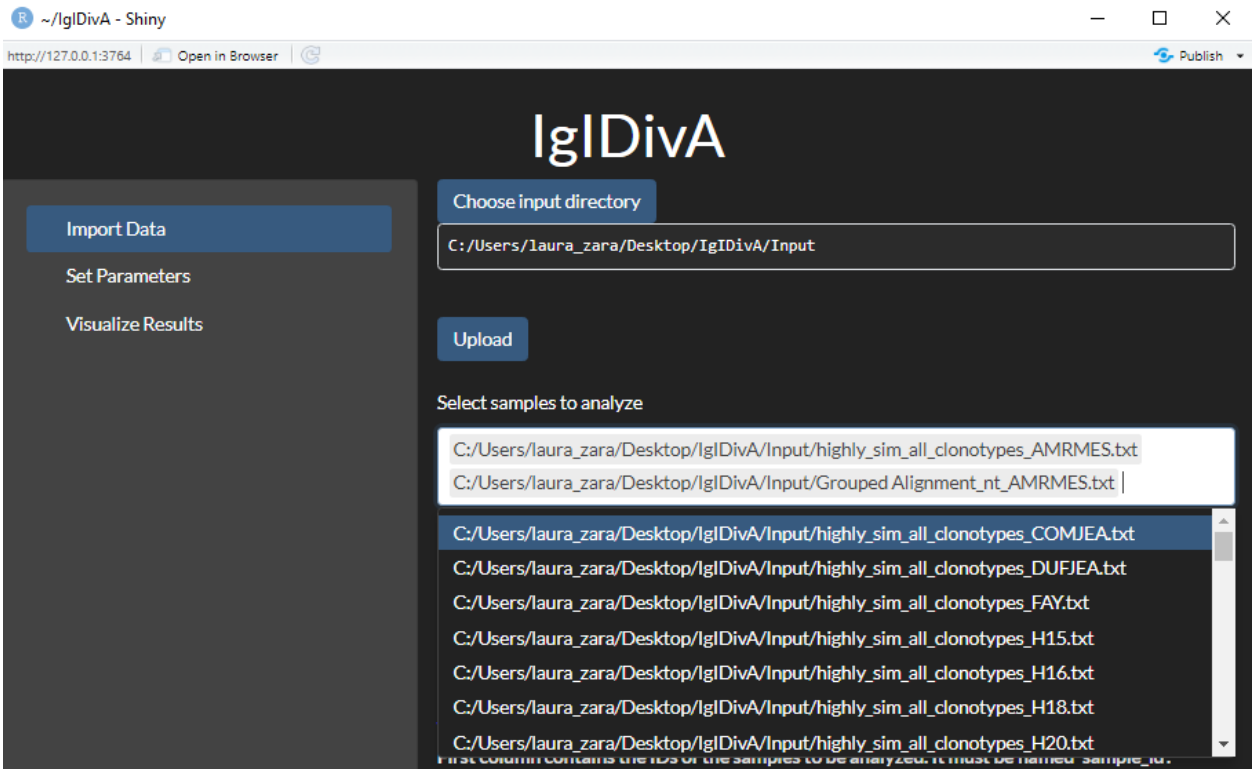
In this tab users can import their data by selecting the directory where the data is stored, by pressing the **Choose input directory** button.



The tool takes as input for each sample the `tripr` output files "highly similar clonotype computation" and "grouped alignment nt", in text format (.txt).
The input folder is selected:



Once the Input folder has been selected, users should verify it by pressing the button **Upload**. Then users can choose which samples from the Input folder they want to include in the analysis.



Users should subsequently verify it by pressing the button **Verify**.

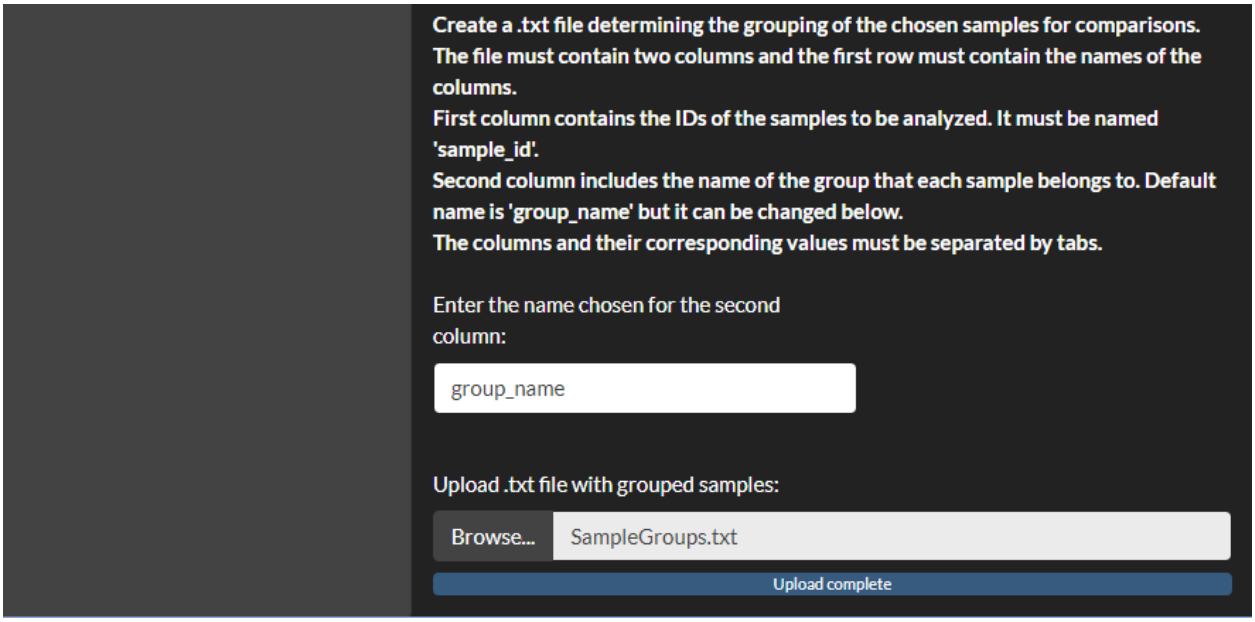
Including groups to compare (optional)

In order to make comparisons between groups of samples, the user needs to create a `tab-delimited` file with two columns.
The first column should be named "sample_id" and should include the names of the samples.
The second column should include the name of the group that each sample belongs to. By default the name of the column is "group_name", but it can be modified in the **Enter the name chosen for the second column** button.

```
groups = data.table::fread("SampleGroups.txt", header = TRUE,
                           sep = "\t",
                           stringsAsFactors = FALSE)
head(groups)
#>    sample_id      group_name
#> 1:    AMRMES CLL subset #169
#> 2:    COMJEA   CLL subset #2
#> 3:    DUFJEA   CLL subset #2
#> 4:       FAY   CLL subset #2
#> 5:       H15   CLL subset #2
#> 6:       H16   CLL subset #2
```

An example file can be found here as "SampleGroups.txt" (the samples correspond to the data mentioned before).

Once created, the file can be uploaded through the **Browse** button. When it is uploaded, a message "Upload completed" will appear. Then, the tab "Set Parameters" should be opened.



Set Parameters

There are different parameters that can be applied:

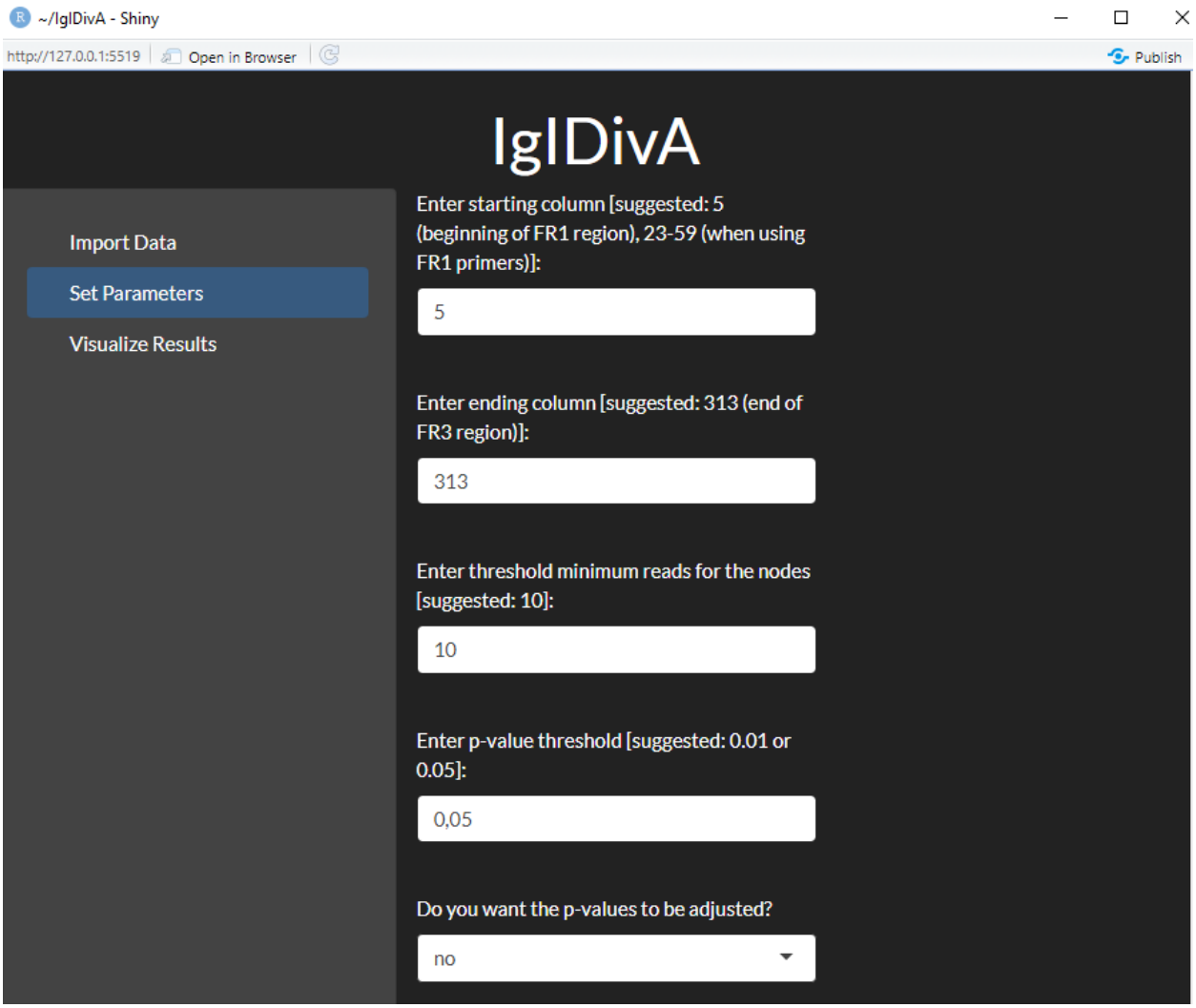
- **Enter starting column**: From the Grouped Alignment file, the user can choose which column corresponds to the beginning of the sequence. If the experimental procedure amplifies the whole immunoglobulin with, for example, leader primers, the starting column should be 5 [the initial 4 columns of the file contain additional information]. If the experimental procedure uses primers that bind in a more downstream position, the starting column should be changed [for example, for primers binding to the FR1 region of the immunoglobulin, the starting column position could be 23 or 59, for example, depending on the binding region]. The default is position 5.

- **Enter ending column**: From the Grouped Alignment file, the user can choose which column corresponds to the end of the sequence [the end of the FR3 region]. The default is position 313.

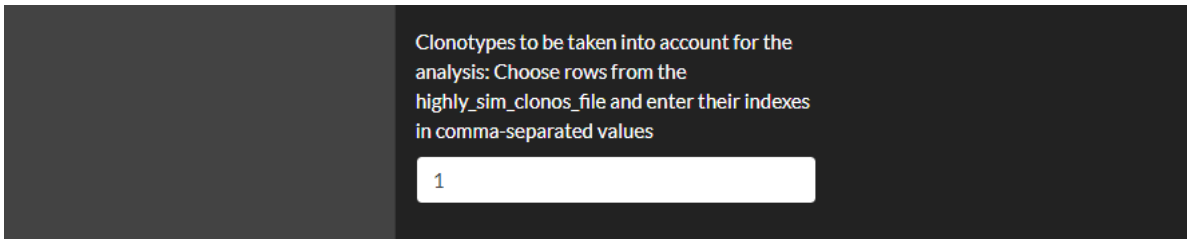
- **Enter threshold minimum reads for the nodes**: The user can choose the minimum number of reads that need to be part of a nucleotide variant (node) for it to be considered in the analysis. The default is 10.

- **Enter p-value threshold**: For the metrics comparison between groups of samples, the user can choose the p-value threshold for a comparison to be considered as statistically significant.The default is 0.05.

- **Do you want the p-values to be adjusted?**: The user can choose between p-value or adjusted p-value. The default is not-adjusted.

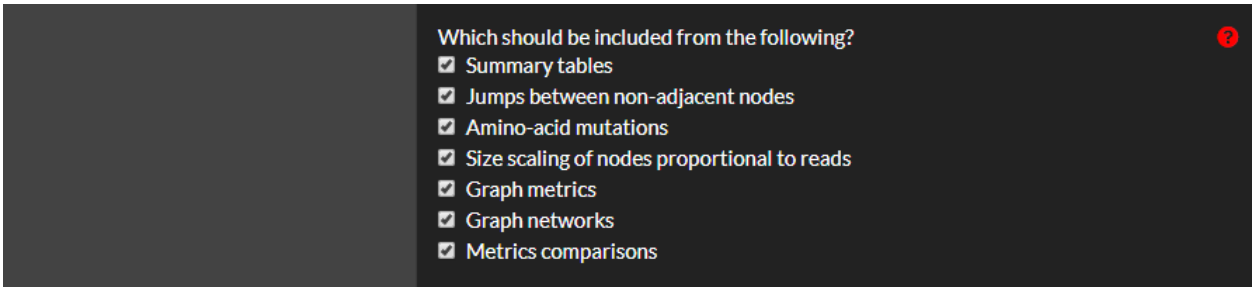


- **Clonotypes to be taken into account for the analysis**: Option for the user to choose the clonotypes to be included in the analysis. One approach would be, for example, to include the first [the most frequent] clonotype. If more than one clonotype is selected, their indexes have to be separated by comma. The default is 1.



Parameters: processing

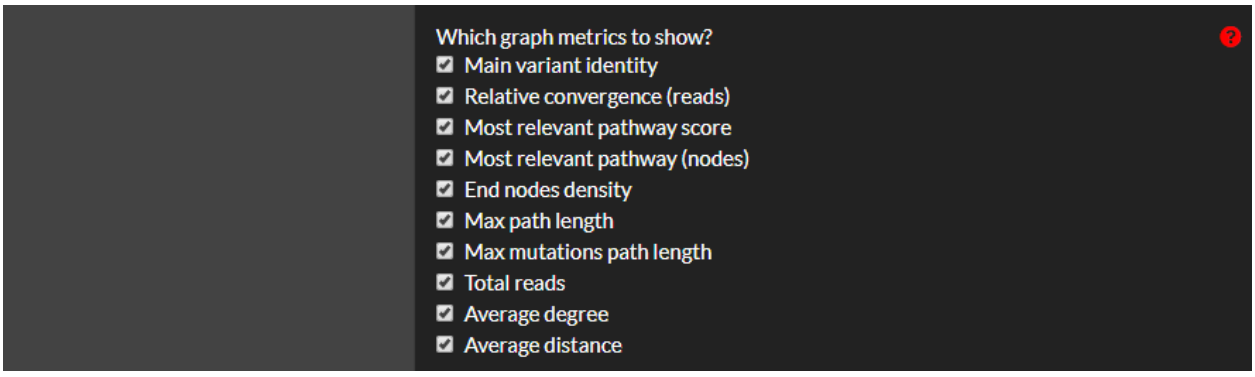
There are different options for the analysis that can be selected:



- **Summary tables**: Tables with summary information [regarding nucleotide variants, sequences, mutational level,...] will be produced throughout the process.
- **Jumps between non-adjacent nodes**: If selected, jumps are allowed and nt vars with common SHMs differing by two or more SHMs will be included.
- **Amino-acid mutations**: The analysis will include the analysis of SHMs at the amino acid level. Replacement mutations will be shown in the graph and tables with the replacement mutations will be produced.
- **Size scaling of nodes proportional to reads**: If selected, the size of the nodes of the graph networks will be proportional to the number of reads of the respective nucleotide variants.
- **Graph metrics**: For each sample, different graph metrics will be calculated (description of the metrics below).
- **Graph networks**: For each sample, a graph network representing the intraclonal diversification will be produced.
- **Metrics comparison**: If the above **Graph metrics** option is selected, there is the option of performing metrics comparison between different groups of samples.

Parameters: metrics

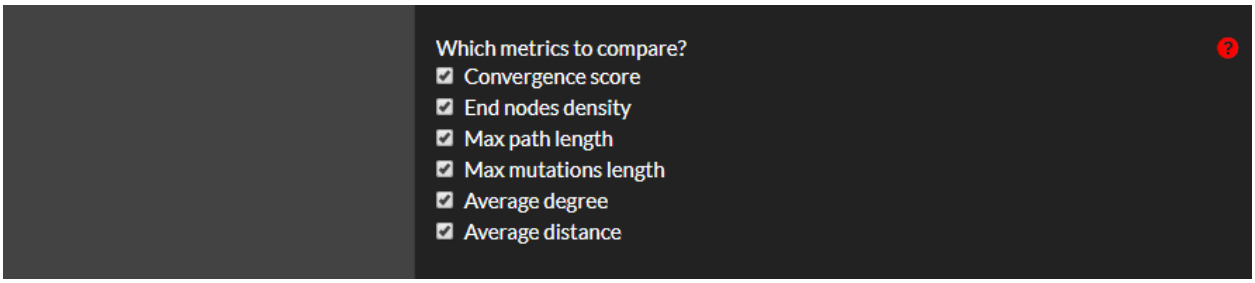
There are different metrics [or related calculations] that can be calculated for the description and determination of the intraclonal diversification level:



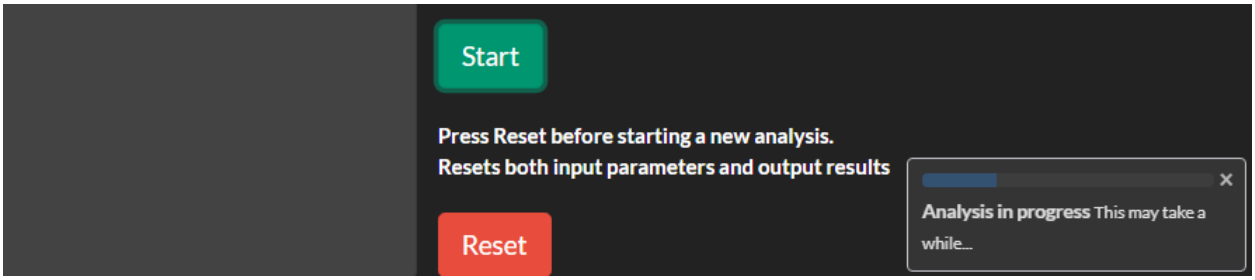
- **Main variant identity**: Percentage (%) of identity of the main nucleotide variant with its respective germline.
- **Relative convergence (reads)**: Graph metric "convergence score". Ratio of the number of sequences of the most relevant pathways to the number of sequences of the main nucleotide variant. It shows the tendency for the BcR IG sequences to accumulate in the main nt var or to acquire additional convergent SHMs.

- **Most relevant pathway score**: Each block of pathways that leads to a particular end node gets a score based on the ratio of the total number of sequences of the nodes forming that block of pathways to the total number of sequences of all the nodes of the network with more SHMs than the main nt var. The block with the highest score is the most relevant pathway, the one that will be used for the calculation of the relative convergence, the convergence score.
- **Most relevant pathway score (nodes)**: Number of nodes of the most relevant pathway.
- **End nodes density**: Graph metric. Ratio of the number of end nodes to the number of nucleotide variants with additional SHMs. It shows the randomness or specificity of the mutational path.
- **Max path length**: Graph metric. Number of levels of additional SHMs. It shows the complexity of the mutational pathways.
- **Max mutations path length**: Graph metric "maximal mutational length". Maximum level of additional SHMs. It shows the complexity of the mutational pathway, allowing non-consecutive SHMs.
- **Total reads**: Total number of reads of the sample.
- **Average degree**: Graph metric. Average total number of connections of each nucleotide variant. It shows the complexity and connectivity of the mutational pathways.
- **Average distance**: Graph metric. Average number of steps along the shortest pathways between each pair of nucleotide variants.

Then, it is possible to choose, among the graph metrics, which one(s) to use to perform comparisons between groups of samples.



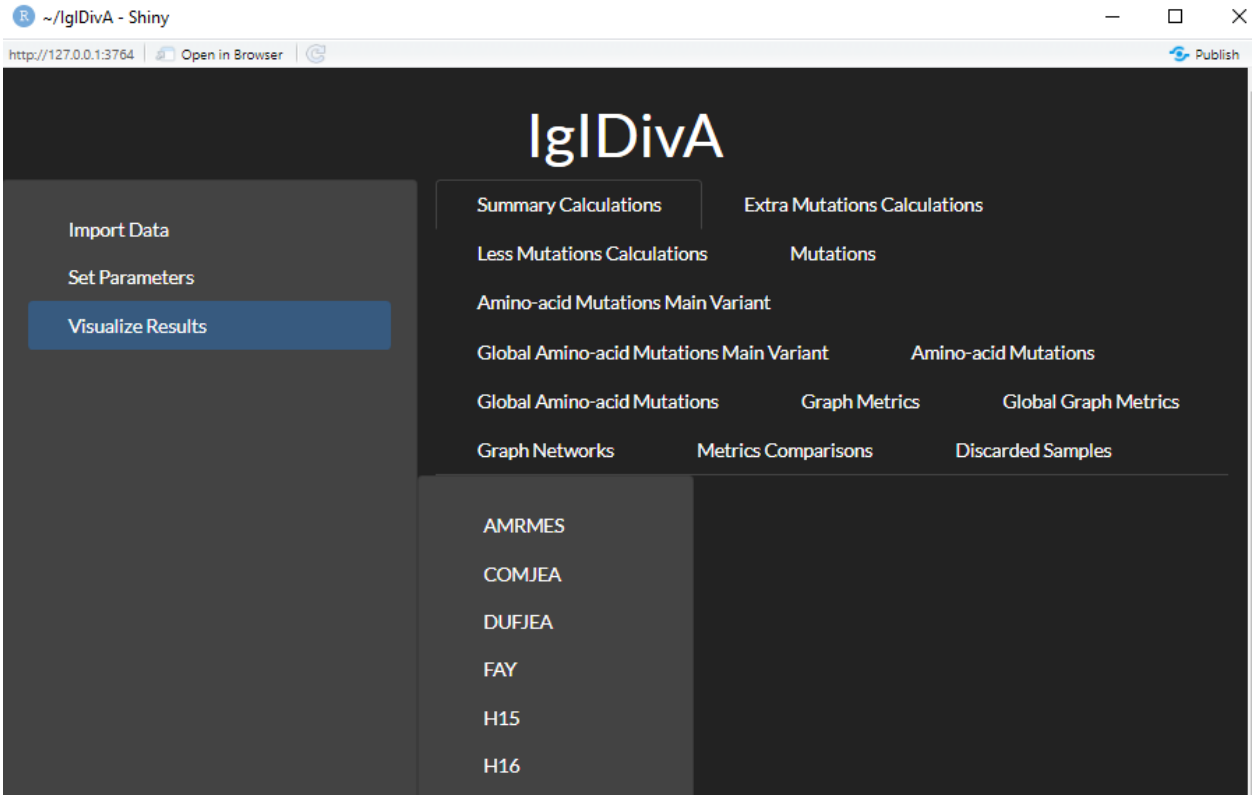
Once all the parameters have been selected, the button **Start** must be pressed. A bar will show how much of the analysis has been completed.



The button **Reset** can be used to start a new analysis, resetting the parameters and the output results.

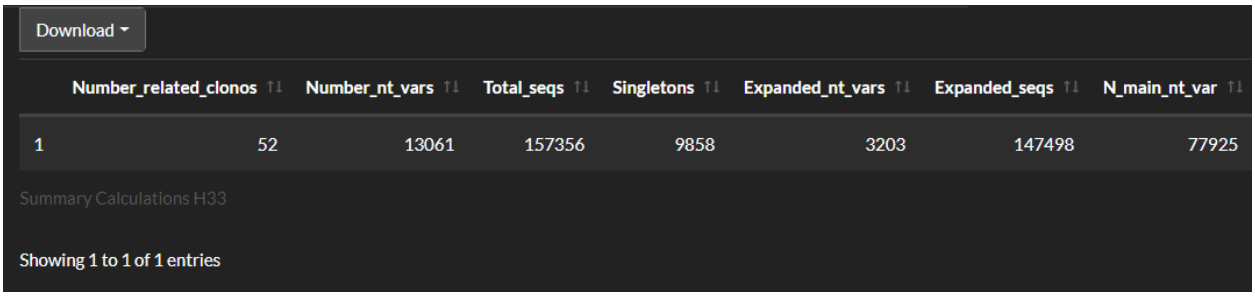
Visualize Results

When the analysis is finished, the **Visualize Results** tab will become active. It might take some time since the analysis is completed until the visualizations are visible.



It shows all the different output results and it offers the possibility of selecting them and choosing which sample to visualize. In each tab, a *Download* button allows for each output to be downloaded independently. Moreover, all the output is saved locally in the user's working directory.

Summary Calculations



For each sample, it shows the number of related clonotypes [clonotypes with the same IGV gene and very similar CDR3] considered for the analysis, the number of nucleotide variants included, the total number of sequences, the number of singletons [nucleotide variants constituted by only one sequence], number of expanded nucleotide variants [nucleotide variants constituted by more than 1 sequence], number of sequences belonging to expanded nucleotide variants, and the number of reads of the main nucleotide variant.

Extra Mutations Calculations

Download

	#mutations	nt_var	seqs
1	1	350	13887
2	2	6	2252
3	3	5	1790
4	4	3	458
5	5	3	854
6	6	1	565

Extra Mutations Calculations H33

Showing 1 to 6 of 6 entries

For each sample, it shows the number of nt vars with additional SHMs for each given number of SHMs, as well as the total number of sequences. It includes the total number of nt vars and sequences.

Less Mutations Calculations

Download

	#mutations	N
1	-3	360
2	-2	2186
3	-1	9089

Less Mutations Calculations H33

Showing 1 to 3 of 3 entries

For each sample, it shows the number of sequences lacking SHMs of the main nt var, for each different number of SHMs.

Mutations

Download

	Mutations	suppl_mut	N	level
18	a27g a106g g107a c108t g113a g151c a172g g173a a184g g185c c294t	0	77925	main
7	a27g a106g g107a c108t g113a g151c a172g g173a a184g g185c c294-	-1	4718	less
8	a27- a106g g107a c108t g113a g151c a172g g173a a184g g185c c294t	-1	4019	less
4	a27g a106- g107- c108t g113a g151c a172g g173a a184g g185c c294t	-2	1798	less
27	a275g g303c	2	1470	additional
33	a232g a275g g303c	3	1019	additional
19	g303c	1	806	additional

For each sample, it provides information for all unique SHMs or combinations of SHMs of all the nt vars that are part of the connected graph network. It also shows the number of SHMs in comparison to the germline, the number of sequences with those SHMs and the mutational level to which they belong. The mutational level is "less" if they have fewer SHMs than the main nt var, "main" for the SHMs of the main nt var, and "additional" for the cases with more SHMs than the main nt var.

Amino-acid Mutations Main Variant

Download

	aa_muts	seqs
1	S36D	77925
2	S38N	77925
3	E51Q	77925
4	S58D	77925
5	S62A	77925

Amino - acid Mutations (main) H33

Showing 1 to 5 of 5 entries

It provides information of the replacement SHMs in the main nt var of each sample, together with the number of sequences carring each mutation.

Gobal Amino-acid Mutations Main Variant

Download

	aa_muts	seqs	id
1	A24V	8312	AMRMES
20	A24V	154587	H16
89	A24V	31794	HERVO
2	A25G	8312	AMRMES
60	A25V	106060	H30
90	A25V	31794	HERVO
63	A68T	106060	H30
50	A68V	125029	H28
98	A83G	31794	HERVO
80	A83V	91010	H35
107	A96G	4792	LIEBER
29	A96V	110812	H18
45	A96V	61580	H22
47	A96V	122852	H23
59	A96V	29309	H29

It contains all identified replacement SHMs in the main nt var of all the samples. It can be used to identify mutational patterns among samples.

Amino-acid Mutations

Download

	aa_muts	seqs
1	N92S	4755
2	I78V	3958
3	K48R	1900
4	S63T	124
5	A24T	105
6	N82S	78
7	S63C	14
8	C23G	14

Amino - acid Mutations (rest) H33

Showing 1 to 8 of 8 entries

It contains all identified replacement SHMs in the nt vars with additional SHMs [excluding the ones of the main nt var].

Global Amino-acid Mutations

Download

	aa_muts	seqs	id
58	A24G	41	H18
169	A24P	11	H21
4	A24T	2065	H15
66	A24T	782	H20
190	A24T	105	H33
249	A24T	1398	LIEBER
36	A24V	130	H18
180	A24V	101	H29
14	A25I	723	H16
20	A25T	61	H16
13	A25V	1020	H16
50	A25V	69	H18

It contains all identified replacement SHMs in the nt vars with additional SHMs [excluding the ones of the main nt var] for all the samples. It can be used to identify mutational patterns among samples.

Graph Metrics

Download

	sample_id	main_nt_var_identity	convergence_score	nb_reads_most_relevant_pathway	nb_reads_main_nt_var	most_relevant_pathway_score	nb_nodes_most_relevant_pathway	max_path_length
1	H33	96.18	0.057	4411	77925	0.519	5	6

Graph Metrics H33

Showing 1 to 1 of 1 entries

For each sample, it contains the germline identity %, the values of the graph metrics as well as information related to those metrics.

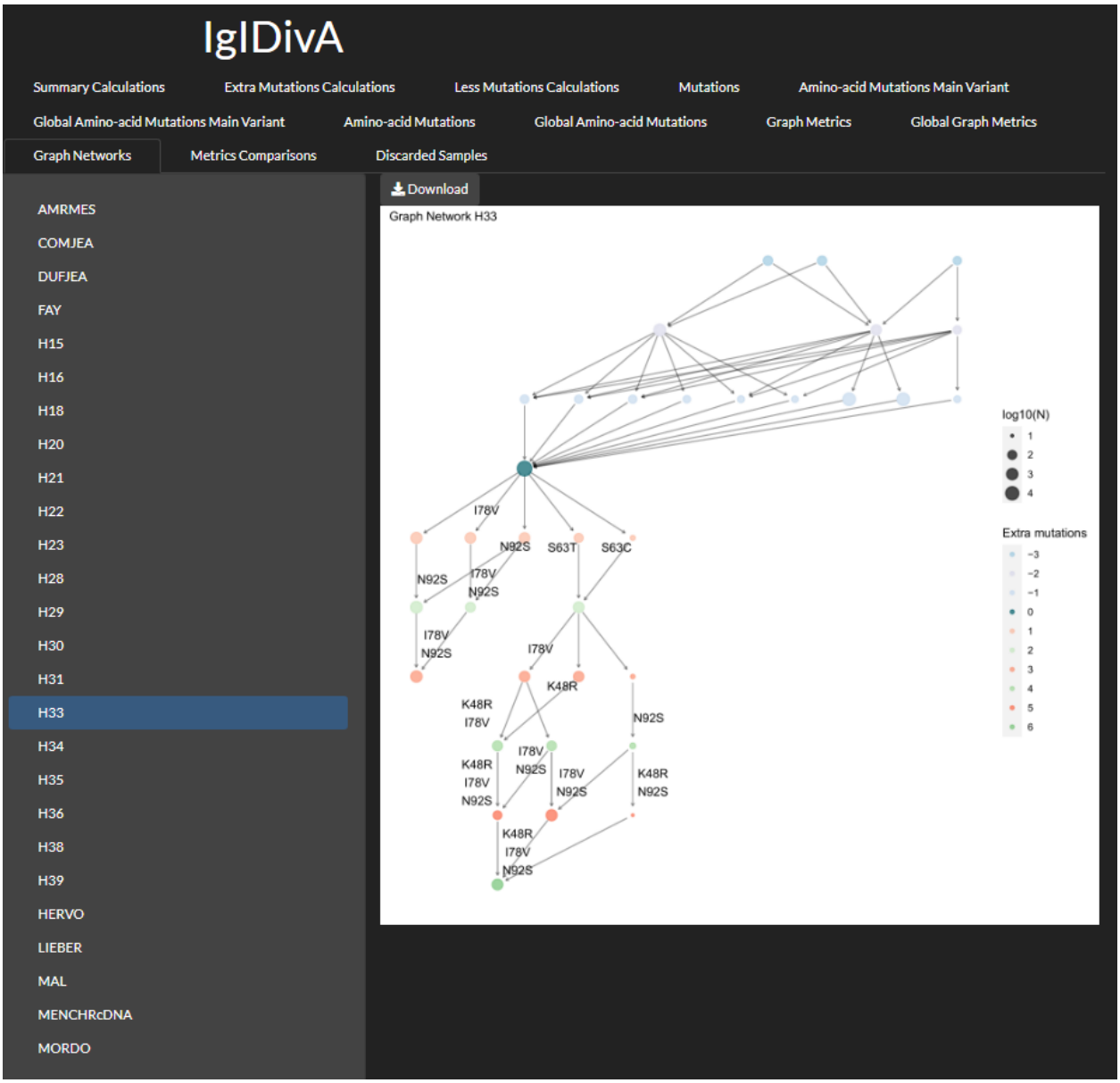
Global Graph Metrics

Download

	sample_id	main_nt_var_identity	convergence_score	nb_reads_most_relevant_pathway	nb_reads_main_nt_var	most_relevant_pathway_score
1	AMRMES	97.57	0.007	62	8312	1
2	COMJEA	97.22	0.008	126	15715	0.55
3	DUFJEA	98.26				
4	FAY	98.95				
5	H15	97.57	0.061	2670	43930	0.787
6	H16	96.18	0.009	1401	154587	0.403
7	H18	98.95	0.019	2121	110812	0.224
8	H20	97.92	0.018	2188	119140	0.231

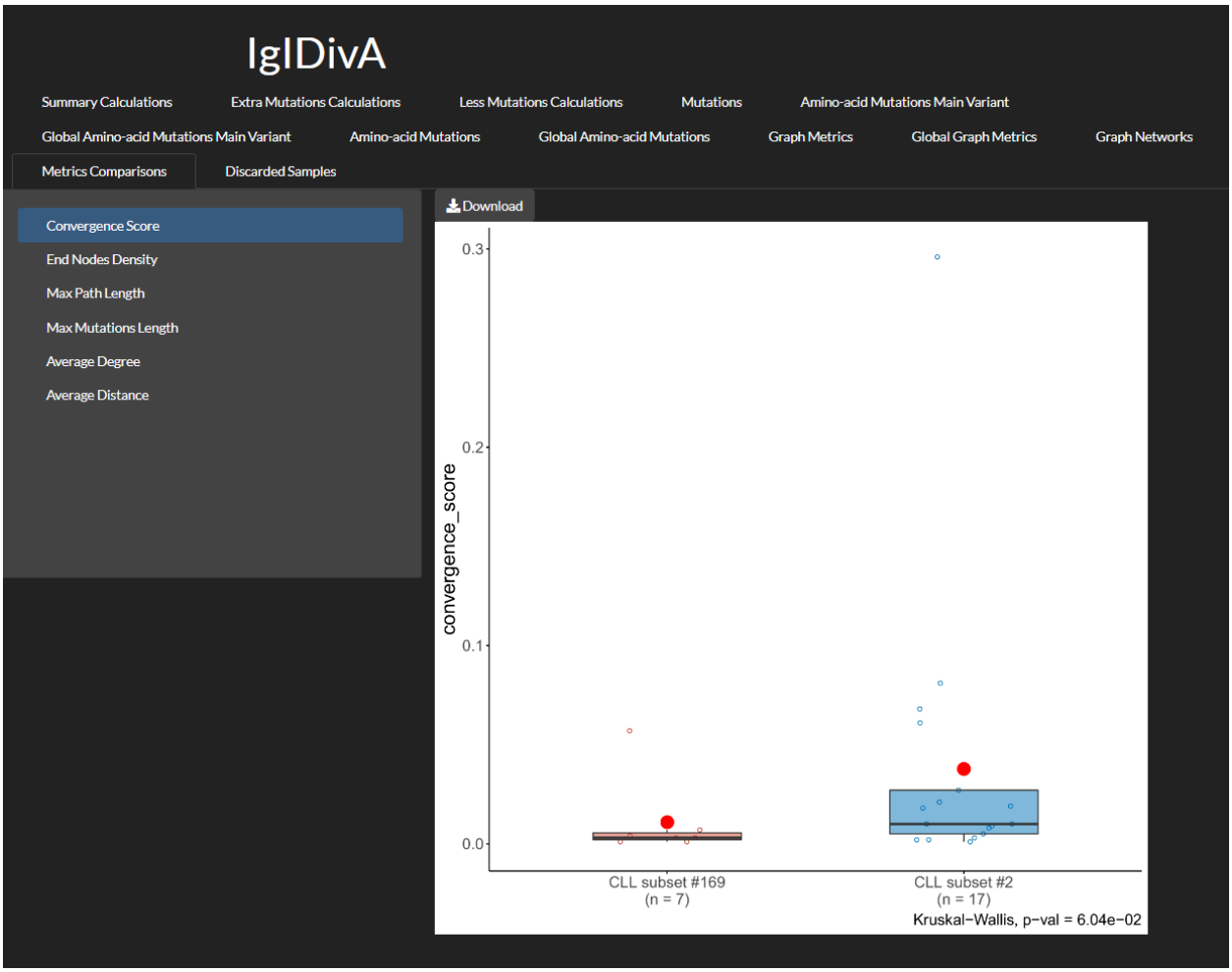
It shows the graph metrics values for all the samples. If a sample has been discarded, the cause is provided.

Graph Networks



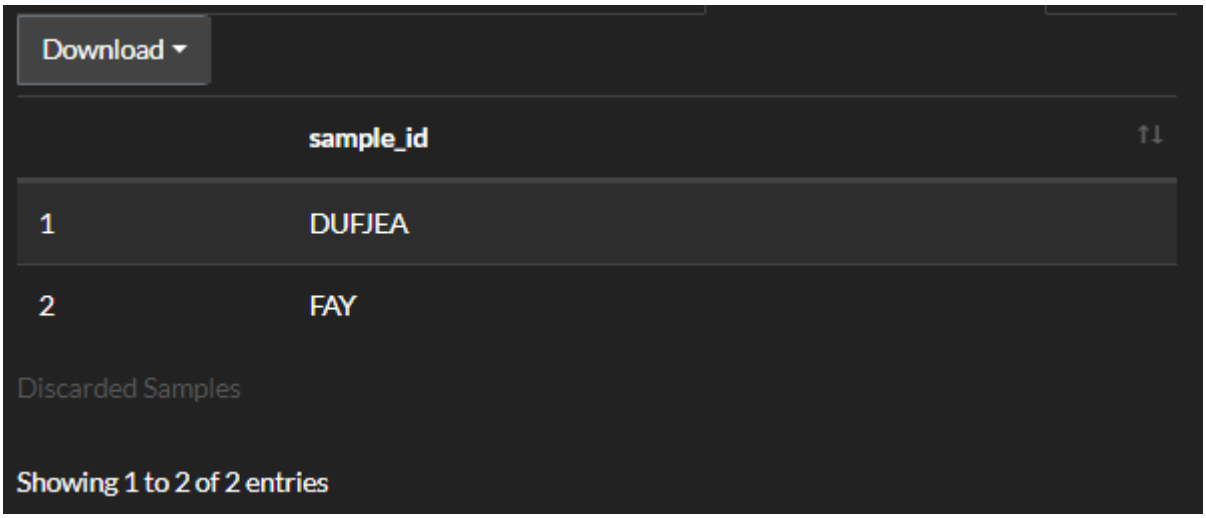
For each sample, it shows the graph network.

Metrics comparison


IgIDivA
Summary Calculations
Extra Mutations Calculations
Less Mutations Calculations
Mutations
Amino-acid Mutations Main Variant
Global Amino-acid Mutations Main Variant
Amino-acid Mutations
Global Amino-acid Mutations
Graph Metrics
Global Graph Metrics
Graph Networks
Metrics Comparisons
Discarded Samples
Convergence Score
End Nodes Density
Max Path Length
Max Mutations Length
Average Degree
Average Distance
Download
convergence_score
0.3
0.2
0.1
0.0
CLL subset #169
(n = 7)
CLL subset #2
(n = 17)
Kruskal−Wallis, p−val = 6.04e−02

If samples are classified into groups, the tool performs pairwise comparisons for all groups. This is performed independently for each of the graph metrics.

Discarded Samples



It provides the names of samples that have been discarded from the analysis [e.g. samples with no connections among nt vars].