

# 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")
install.packages("shinyvalidate")
```

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 trip 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 trip.

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)
path = paste0(getwd(), "/Input")
if (!dir.exists(path)){
  dir.create(path)}
zen4R::download_zenodo('10.5281/zenodo.6616046', path = path)
```

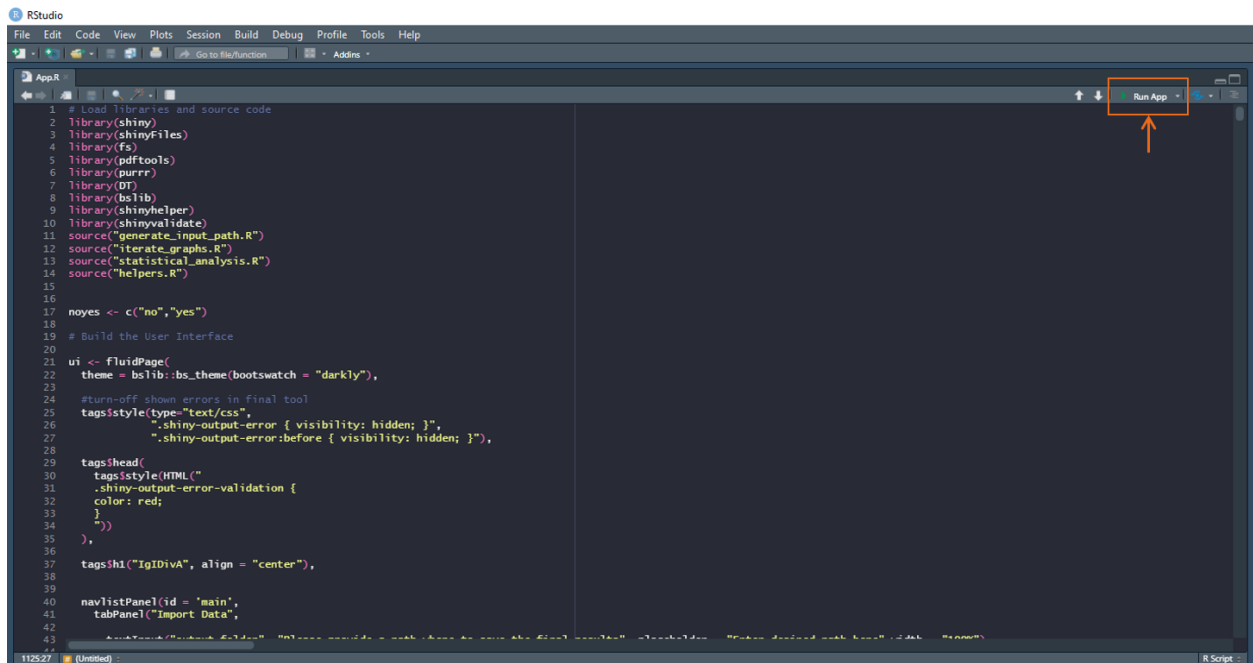
[The variable “path” can be changed with the location where the user wants to store the Input].

Note: warnings might appear in RStudio indicating that the downloaded length of some files != reported length. This means that not all the length of those files was downloaded [probably due to the Internet speed]. One solution is to increase the ‘downloading’ time in Rstudio, with this command:

```
options(timeout = max(600, getOption("timeout")))
```

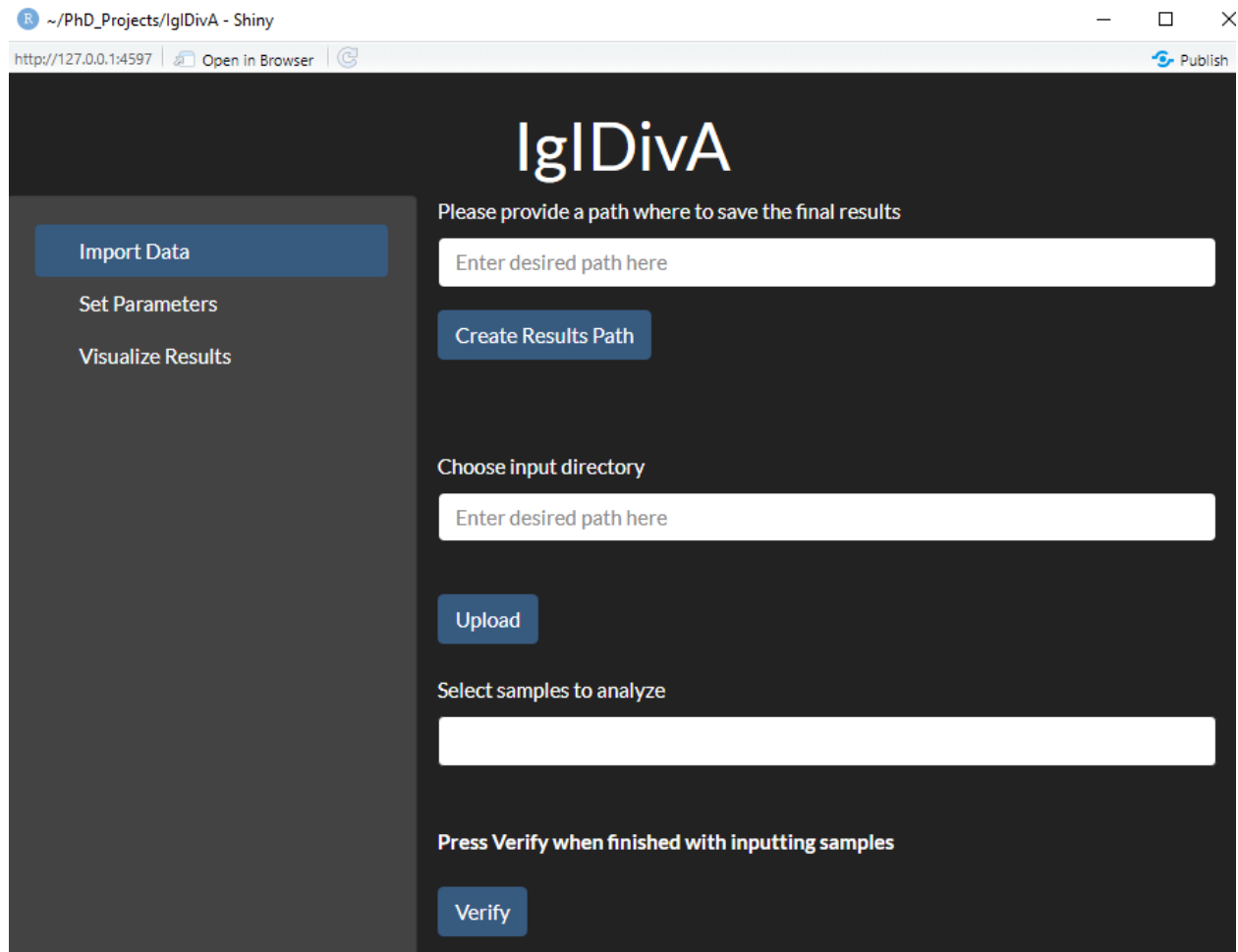
## Running IgIDivA as a shiny application

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.



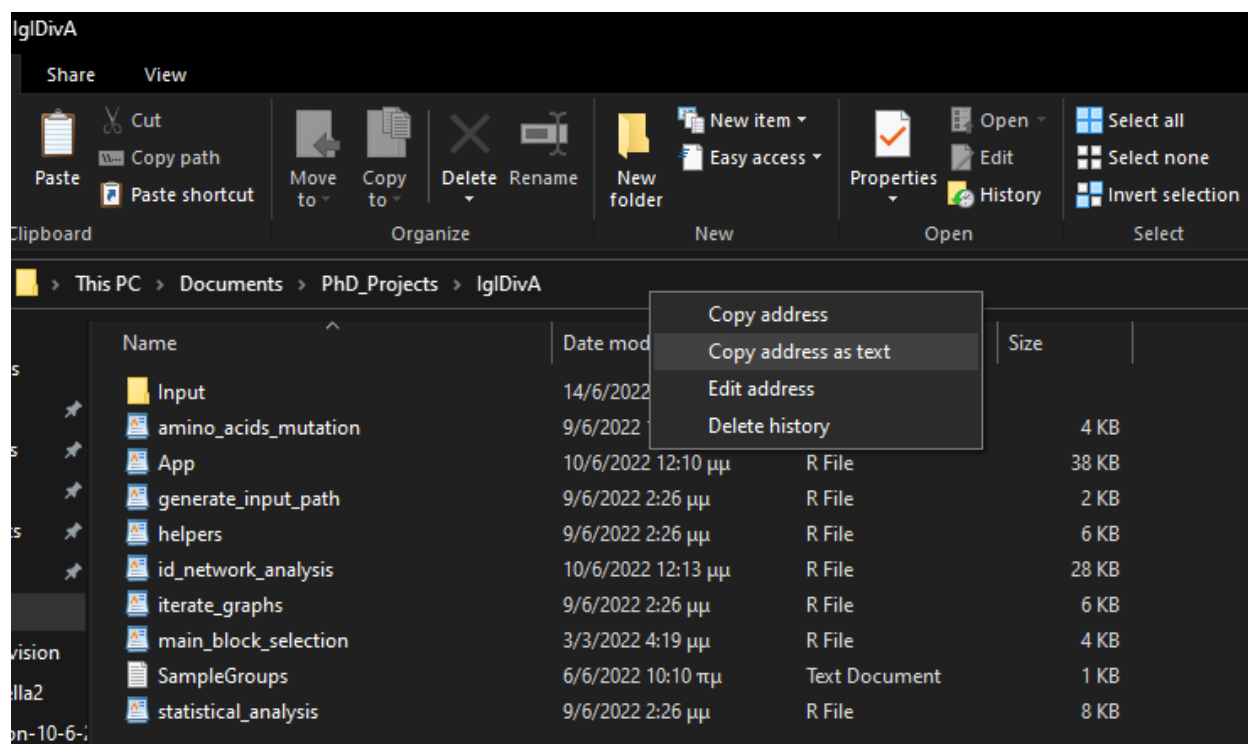
## Import data

In this tab users can create the folders where the results will be stored and import their data.

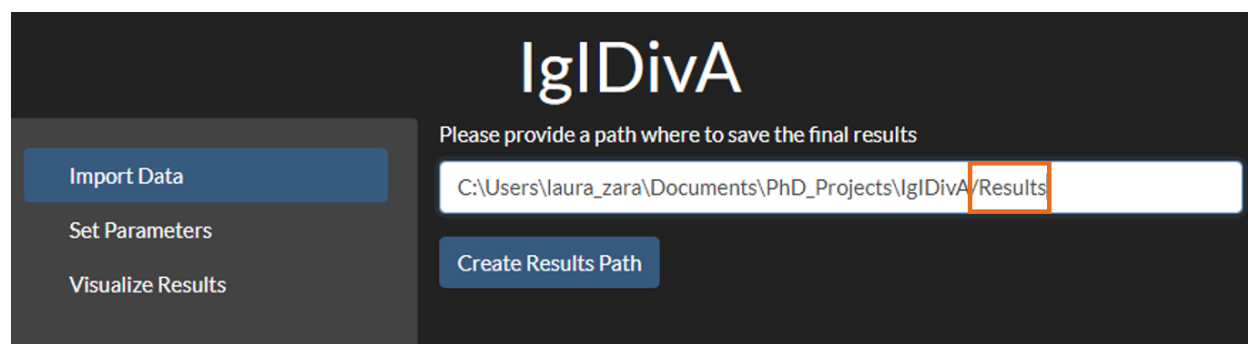


The screenshot shows a web browser window with the title "IgIDivA - Shiny". The address bar shows "http://127.0.0.1:4597". The browser has buttons for "Open in Browser" and "Publish". The application interface has a dark theme. On the left, there is a sidebar with three buttons: "Import Data" (highlighted in blue), "Set Parameters", and "Visualize Results". The main content area has the title "IgIDivA" in large white letters. Below the title, there is a section titled "Please provide a path where to save the final results". It contains a text input field with the placeholder "Enter desired path here" and a blue button labeled "Create Results Path". Below this, there is a section titled "Choose input directory" with another text input field with the placeholder "Enter desired path here". Below that is a blue button labeled "Upload". Then, there is a section titled "Select samples to analyze" with a text input field. At the bottom, there is a section titled "Press Verify when finished with inputting samples" and a blue button labeled "Verify".

**Import data: Select/Create Output folder** First, the user should specify the Results folder. For that, the user can go to the folder in their computer where they would like to store the output and press **copy address as text**:



Then, the copied path should be pasted in the area **Enter desired path here**, together with a “/” and the name of the Results folder that the user wants to create [e.g. “Results”]. If a folder with this name does not exist, it will be created:



Then, the **Create Results Path** should be pressed.

**Import data: Choose input directory** The following step consists on the selection of the Input folder. Following the same approach as for the output folder, the user will enter in the path where the Input files are stored. 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:

# IgIDivA

Please provide a path where to save the final results

C:\Users\laura\_zara\Documents\PhD\_Projects\IgIDivA\Results

Create Results Path

Choose input directory

C:\Users\laura\_zara\Documents\PhD\_Projects\IgIDivA\Input

Upload

Import Data

Set Parameters

Visualize Results

Once the Input folder address has been added, 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.

The screenshot shows the IgIDivA web interface. On the left is a dark sidebar with three buttons: 'Import Data' (highlighted in blue), 'Set Parameters', and 'Visualize Results'. The main area has a dark background. At the top, the 'IgIDivA' logo is displayed. Below it, a text prompt says 'Please provide a path where to save the final results'. A text input field contains the path 'C:\Users\laura\_zara\Documents\PhD\_Projects\IgIDivA\Results', and a blue 'Create Results Path' button is below it. Further down, another text prompt says 'Choose input directory'. A text input field contains the path 'C:\Users\laura\_zara\Documents\PhD\_Projects\IgIDivA\Input', and a blue 'Upload' button is below it. At the bottom, a text prompt says 'Select samples to analyze'. A text input field contains 'AMRMES COMJEA |'. Below this is a scrollable list of sample names: 'DUFJEA' (highlighted in blue), 'FAY', 'H15', 'H16', 'H18', 'H20', and 'H21'.

Users should subsequently verify the selection by pressing the button **Verify**. Please, mind the order of the steps. If the output folder is changed, it is necessary to press again the **Verify** button for the selected samples.

**Import data: 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. The file would look like this:

```

SampleGroups - Notepad
File Edit Format View Help
sample_id      group_name
AMRMES  CLL subset #169
COMJEA  CLL subset #2
DUFJEA  CLL subset #2
FAY     CLL subset #2
H15     CLL subset #2
H16     CLL subset #2
H18     CLL subset #2
H20     CLL subset #2
H21     CLL subset #2
H22     CLL subset #2
H23     CLL subset #2
H28     CLL subset #2
H29     CLL subset #2
H30     CLL subset #2
H31     CLL subset #2
H33     CLL subset #169
H34     CLL subset #169
H35     CLL subset #169
H36     CLL subset #169
H38     CLL subset #169
H39     CLL subset #169
HERVO   CLL subset #2
LIEBER  CLL subset #2
MAL     CLL subset #2
MENCHRCDNA  CLL subset #2
MORDO   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.

Create a .txt file determining the grouping of the chosen samples for comparisons. The file must contain two columns and the first row must contain the names of the columns.

First column contains the IDs of the samples to be analyzed. It must be named 'sample\_id'.

Second column includes the name of the group that each sample belongs to. Default name is 'group\_name' but it can be changed below.

The columns and their corresponding values must be separated by tabs.

Enter the name chosen for the second column:

Upload .txt file with grouped samples:

Upload complete

## 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.

The screenshot displays the IgIDivA web application interface. On the left, a dark sidebar contains three navigation links: 'Import Data', 'Set Parameters' (highlighted in blue), and 'Visualize Results'. The main content area has a dark background with the 'IgIDivA' logo at the top. Below the logo, there are five input sections, each with a label and a text box:

- Enter starting column [suggested: 5 (beginning of FR1 region), 23-59 (when using FR1 primers)]:** The text box contains the value '5'.
- Enter ending column [suggested: 313 (end of FR3 region)]:** The text box contains the value '313'.
- Enter threshold minimum reads for the nodes [suggested: 10]:** The text box contains the value '10'.
- Enter p-value threshold [suggested: 0.01 or 0.05]:** The text box contains the value '0,05'.
- Do you want the p-values to be adjusted?:** A dropdown menu shows the selected option 'no'.

- **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. The default is 1.



Clonotypes to be taken into account for the analysis: Choose the row from the `highly_sim_clonos_file` and enter its index [default = 1]

## Parameters: processing

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

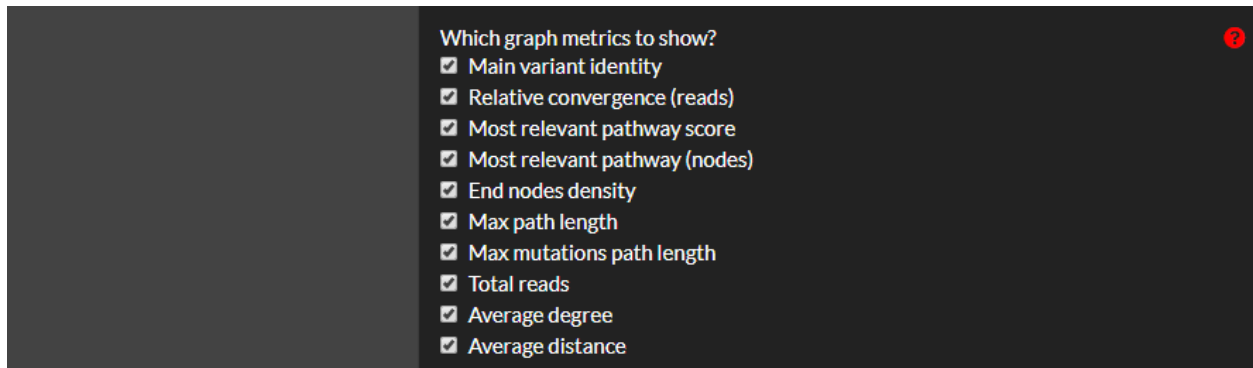
Which should be included from the following?

- ☒ Summary tables
- ☒ Jumps between non-adjacent nodes
- ☐ Separate graphs
- ☒ Amino-acid mutations
- ☒ Size scaling of nodes proportional to reads
- ☒ Graph metrics
- ☒ Graph networks
- ☒ Metrics comparisons

- **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.
- **Separate graphs:** If selected, the graph network of each sample will be separated into two different graphs: on the left, the main nt var and the nt vars with fewer SHMs than the main nt var [the “less mutations pathway”] and on the right the main nt var and the nt vars with additional mutations. The different levels of mutations are aligned in both graph networks. This parameter affects only the visualization. By default this parameter is “off”.
- **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.

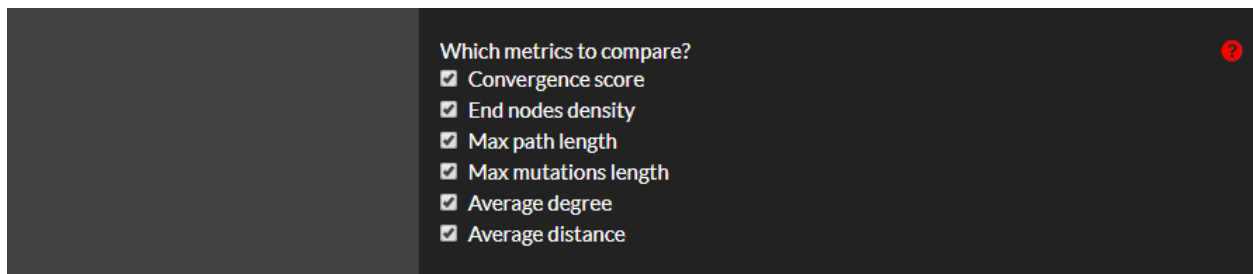
## Parameters: metrics

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

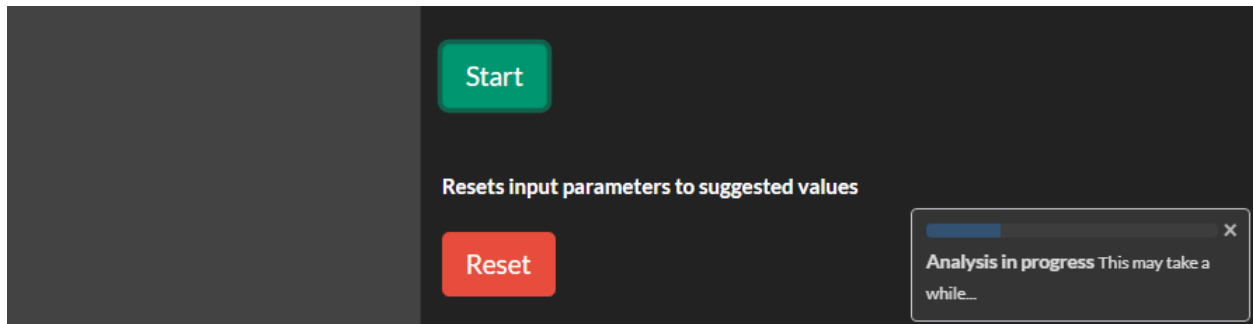


- **Main variant identity:** Percentage (%) of identity of the main nucleotide variant with its respective germline.
- **Relative convergence (reads):** Graph metric “convergence score”. Ratio of the number of sequences of the most relevant pathways to the number of sequences of the main nucleotide variant. 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.



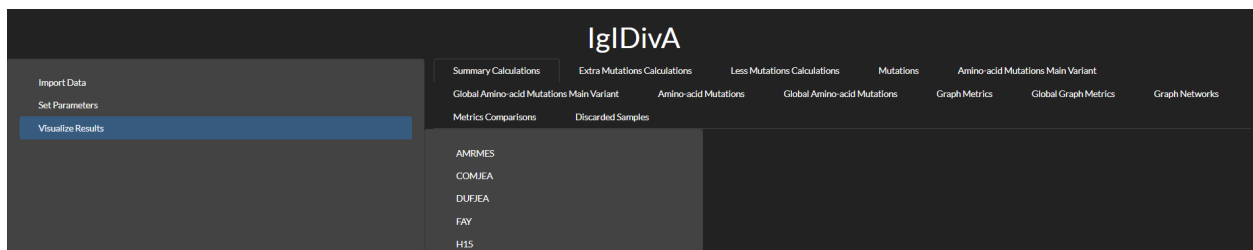
Once all the parameters have been selected, the button **Start** must be pressed. A bar will show how much of the analysis has been completed.



The button **Reset** can be used to start a new analysis, resetting the parameters [the output results will be reset when pressing the **Start** button].

When the analysis is finished, a notification will appear with the message 'File conversion in progress...'. This conversion is performed to allow the visualizations to be visible in the **Visualize results** tab. Once it is ready, the user will be automatically redirected to the **Visualize Results** tab.

## Visualize Results



This tab shows all the different output results and it offers the possibility of selecting them and choosing which sample to visualize. All the output results are saved locally in the user's previously selected output folder.

## Summary Calculations

Search: <input type="text"/>						
Number_related_clonos <span>↑↓</span>	Number_nt_vars <span>↑↓</span>	Total_seqs <span>↑↓</span>	Singletons <span>↑↓</span>	Expanded_nt_vars <span>↑↓</span>	Expanded_seqs <span>↑↓</span>	N_main_nt_var <span>↑↓</span>
1	52	13061	157356	9858	3203	147498
Summary Calculations H33						
Showing 1 to 1 of 1 entries						

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. [Example shown: sample H33].

## Extra Mutations Calculations

Search: <input type="text"/>			
	#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. [Example shown: sample H33].

## Less Mutations Calculations

Search: <input type="text"/>		
	#mutations ↑↓	N ↑↓
1	-3	360
2	-2	2186
3	-1	9067
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. [Example shown: sample H33].

## Mutations

Search: <input type="text"/>					
	Mutations	↑↓	suppl_mut ↑↓	N ↑↓	level ↑↓
1	c71t c74g c75- c191t c198t c265a g303a		-1	32	less
2	c71- c74g c75t c191t c198t c265a g303a		-1	22	less
3	c71t c74g c75t c191- c198t c265a g303a		-1	20	less
4	c71t c74g c75t c191t c198- c265a g303a		-1	12	less
5	c71t c74g c75t c191t c198t c265a g303a		0	8312	main
6	c108t		1	43	additional
7	c108t g188c g282a		3	19	additional

Mutations AMRMES

Showing 1 to 7 of 7 entries

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. [Example shown: sample AMRMES].

### Amino-acid Mutations Main Variant

Search: <input type="text"/>		
	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 carrying each mutation. [Example shown: sample H33].

## Global Amino-acid Mutations Main Variant

Search: <input type="text"/>					
	aa_muts	tl	seqs	tl	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. [Example shown: all samples from example dataset].

## Amino-acid Mutations

Search: <input type="text"/>		
	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]. [Example shown: sample H33].

## Global Amino-acid Mutations

Search: <input type="text"/>				
	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. [Example shown: all samples from example dataset].

## Graph Metrics

Search: <input type="text"/>								
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. [Example shown: sample H33].

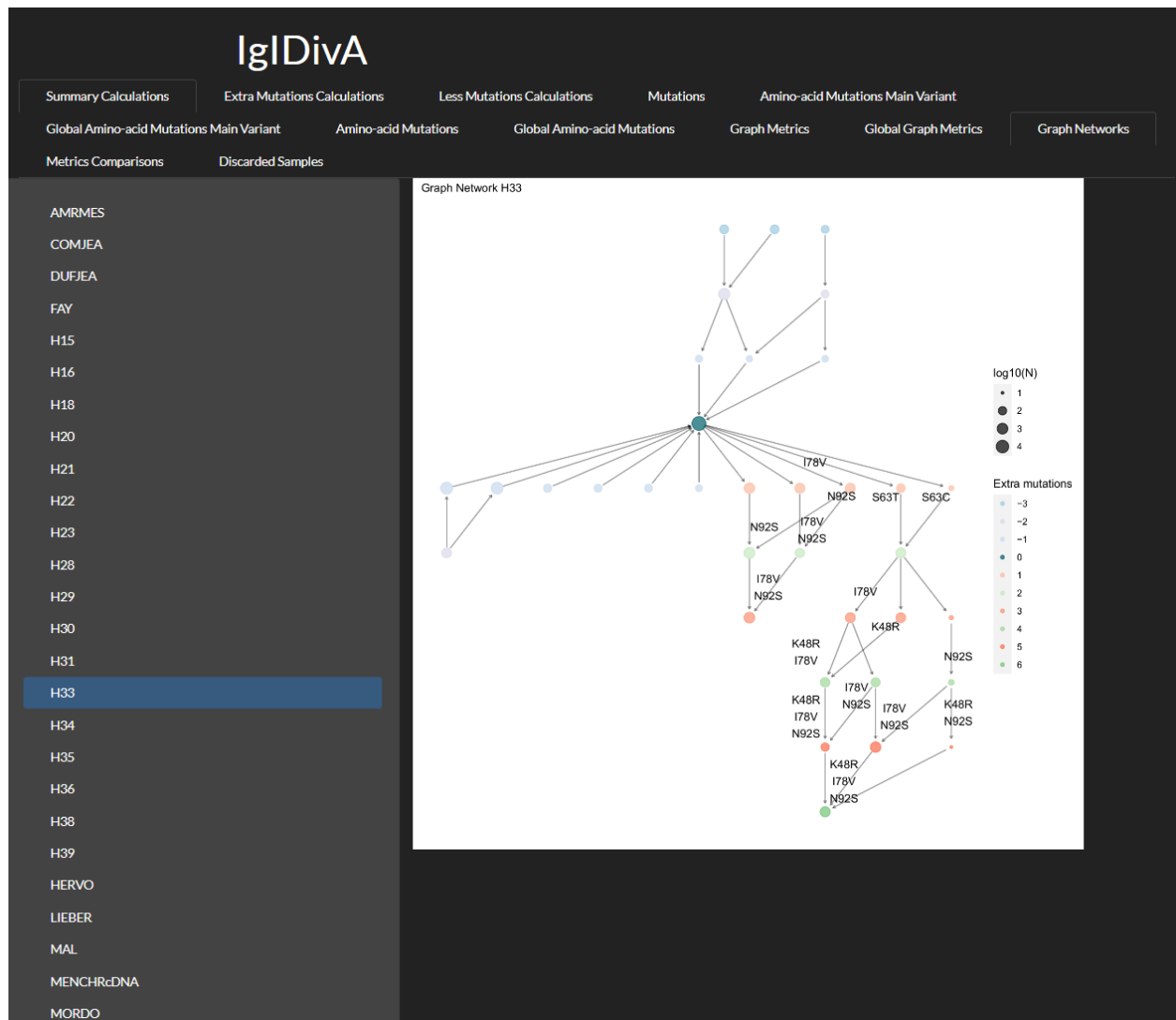
## Global Graph Metrics

Search: <input type="text"/>							
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	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	
9	H21	97.92	0.021	2138	101884	0.181	
10	H22	96.18	0.002	139	61580	0.535	

It shows the graph metrics values for all the samples. If a sample has been discarded, the cause is provided. [Example shown: all samples from example dataset].

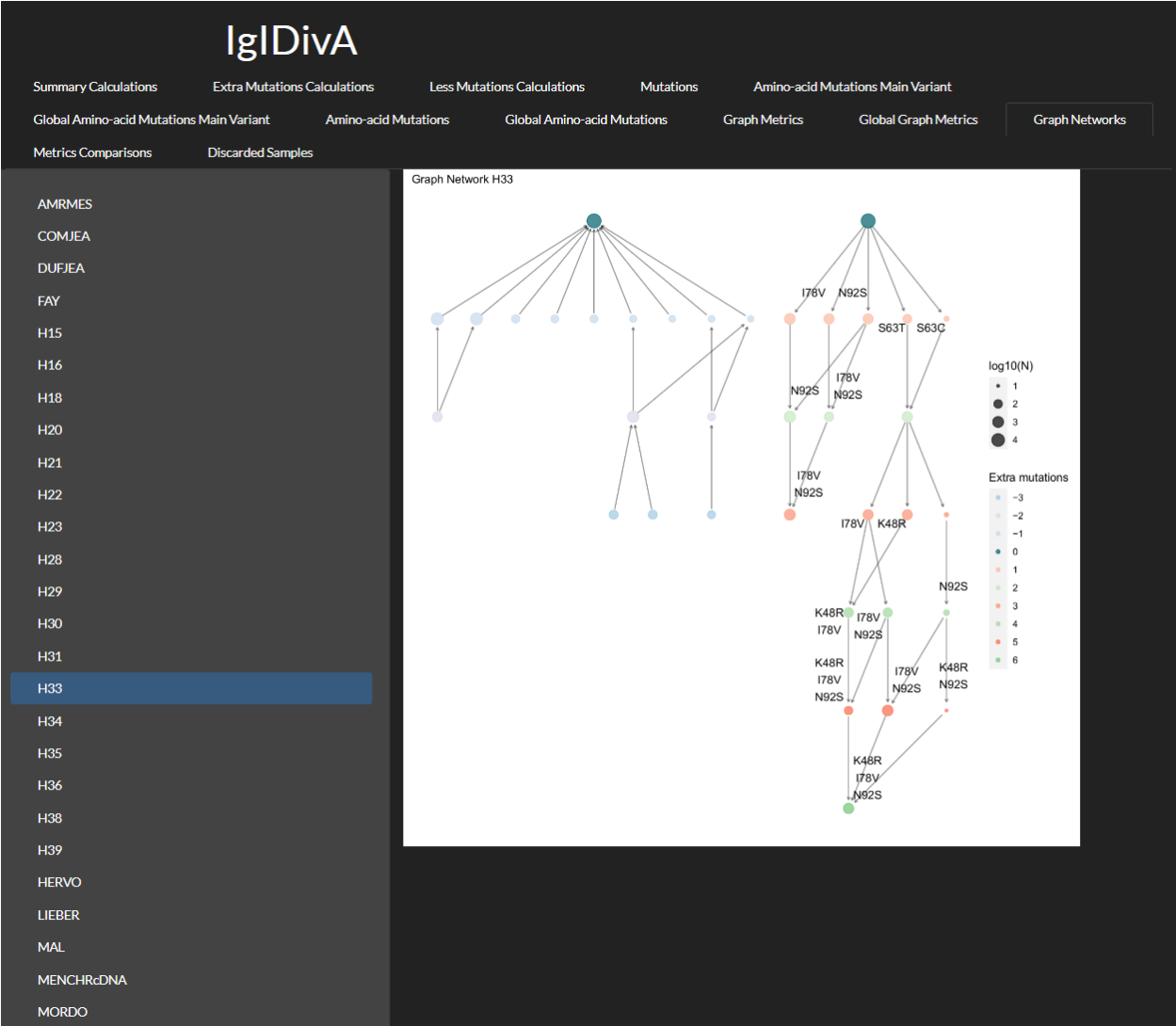
## Graph Networks





For each sample, it shows the graph network. [Example shown: sample H33].

If the parameter “Separate graphs” is selected, the graph network gets separated in two [nt vars with fewer SHMs than the main nt var on the left and nt vars with additional SHMs on the right]. For example [sample H33]:



# 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

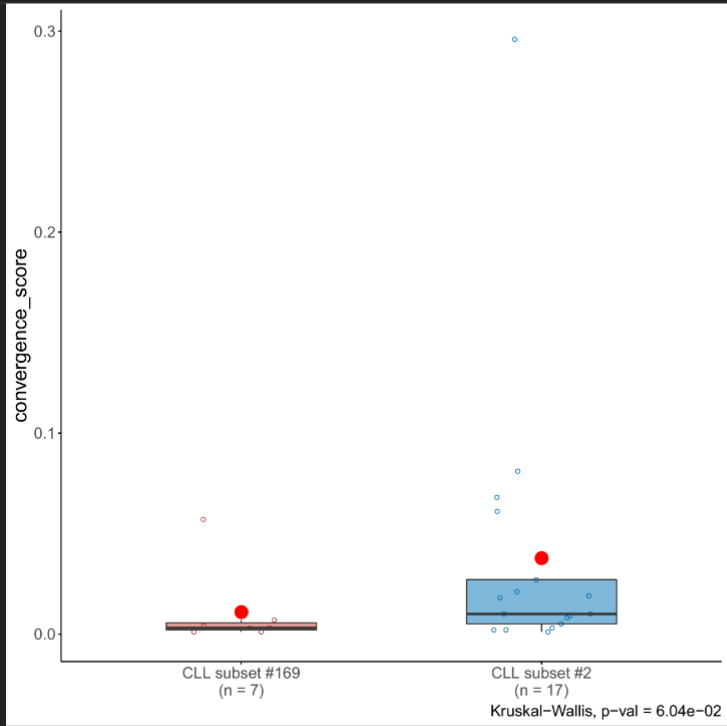
End Nodes Density

Max Path Length

Max Mutations Length

Average Degree

Average Distance





If samples are classified into groups, the tool performs pairwise comparisons for all groups. This is performed independently for each of the graph metrics. [Example shown: all samples from example dataset].

## Discarded Samples

Search:

	sample_id	
1	DUFJEA	11
2	FAY	

Discarded Samples

Showing 1 to 2 of 2 entries

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

That's all! If there is any issue, please feel free to open an issue in the [GitHub repository](#) of IgIDivA.

Thank you for using IgIDivA! Enjoy! :)