vDiveR

Release 0.1

Pendy Tok, Li Chuin Chong, Evgeniia Chikina

Nov 12, 2022

CONTENTS

Note: This project is under active development.

vDiveR is a DiMA wrapper implemented web-based application, hosted on R Shiny server (https://protocol-viral-diversity.shinyapps.io/DiveR), to ease the visualization of outputs from Diversity Motif Analyser (DiMA; https://github.com/PU-SDS/DiMA). vDiveR allows visualization of the diversity motifs (index, major, minor and unique) for elucidation of the underlying inherent dynamics (Figure. 1).

Additionally, the R source code is publicly accessible from the GitHub repository at https://github.com/pendy05/DiveR (distributed under the MIT license).

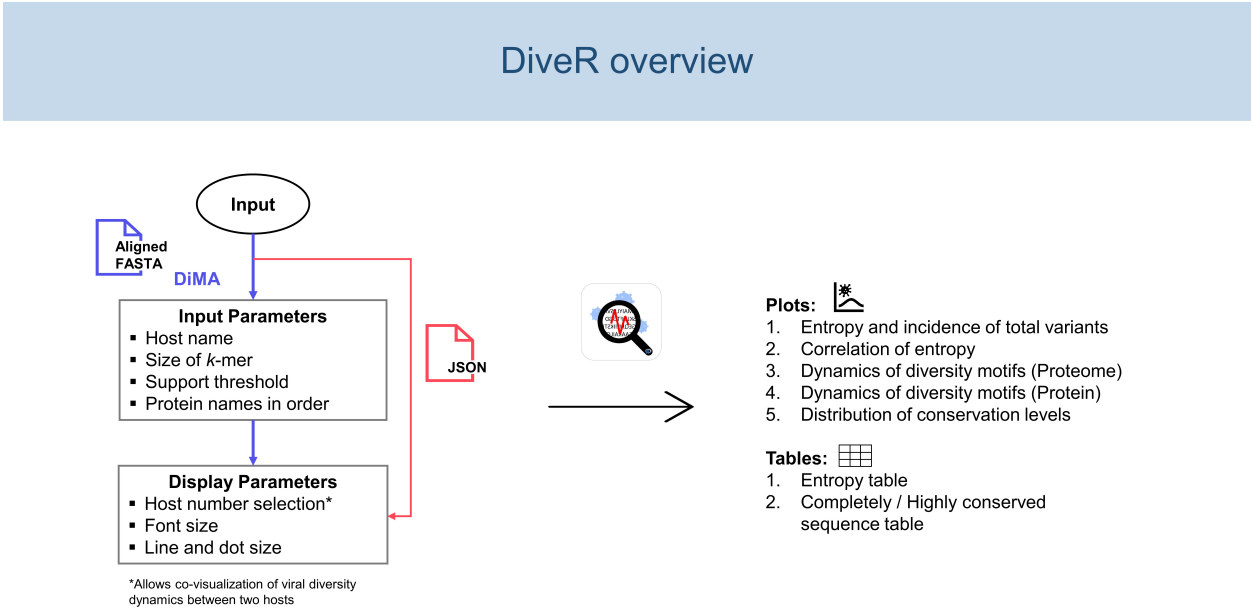


Figure 1: vDiveR overview.

Hint: A demonstration video on how to use vDiveR R Shiny App is available here!

CHAPTER

ONE

CONTENTS

1.1 About

Viruses are one of the main contributors to the global burden of infectious-related mortality and disability. Sequence diversity, as a result of various evolutionary forces, can expand host repertoire or enhance infective ability of viruses, resulting in immune escape. This poses a challenge to the design of diagnostic, prophylactic, and therapeutic interventions against viruses. Thus, it is crucial to understand the diversity and the dynamics of viral sequence change to aid in the design of vaccines or development of therapeutics and diagnostics against a virus. The publicly available tool, Diversity Motif Analyser (DiMA; https://github.com/PU-SDS/DiMA) was developed to facilitate the dissection of sequence diversity dynamics for viruses. DiMA quantifies the sequence diversity using Shannon's entropy for each aligned overlapping *k*-mer positions, distributes the *k*-mers into four diversity motifs (index, major, minor and unique) and stores this information in JSON format. However, interpretation and analysis of data stored in JSON data might be a challenging task to biologists who have limited or no knowledge of bio-informatics or programming background.

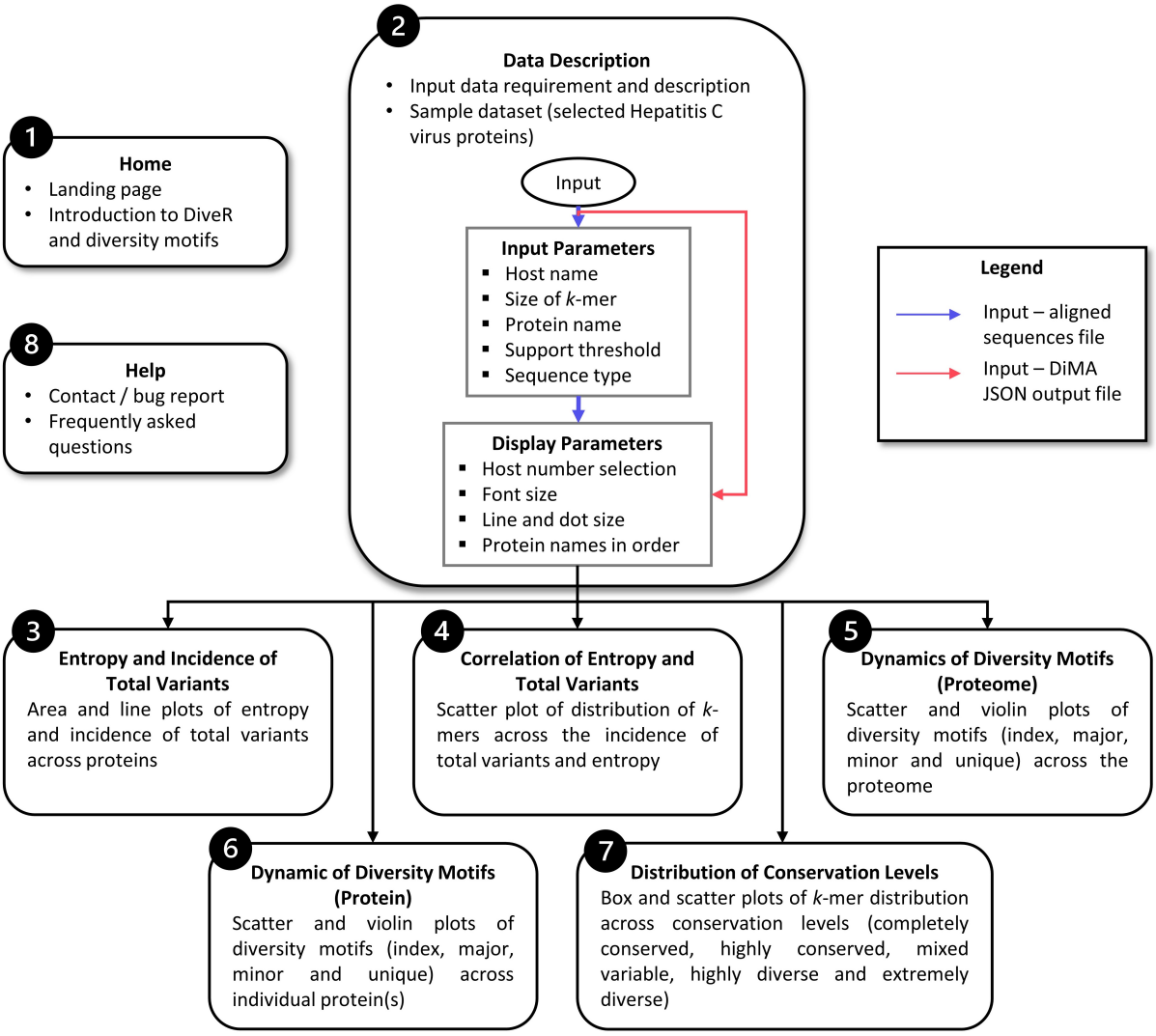
Herein, we present vDiveR, a DiMA wrapper implemented as a web-based application, hosted on R Shiny server (https://protocol-viral-diversity.shinyapps.io/DiveR/), to ease the visualization of outputs from DiMA. vDiveR allows visualization of the diversity motifs (index, major, minor and unique) for elucidation of the underlying inherent dynamics. The sequence with the highest incidence at a given *k*-mer position in a protein alignment is the index, while all the others at the position are variants to the index. Major variant is the predominant sequence amongst the variants, while minor variants are distinct sequences with frequency lesser than the major variant, but occur more than once. Unique variants are distinct sequences that occur only once.

vDiveR presents a total of eight tabs: 1) homepage, 2) data description, with tabs 3) to 7) presenting five plots depicting sequence variability dynamics and lastly 8) help page tab (Figure. 2). vDiveR generates five plots for *k*-mer positions of a viral protein/proteome:

1. entropy and incidence of total variants,
2. relationship between entropy and total variants,
3. dynamics of diversity motifs for the collective proteome,
4. dynamics of diversity motifs for the individual proteins, and
5. distribution of conservation levels (completely conserved, highly conserved, mixed variable, highly diverse, and extremely diverse).

In summary, the simplicity of vDiveR makes the study of viral protein sequence diversity dynamics more accessible to a wider community of researchers. This should help better understand the dynamics of sequence change among viruses and further explore its effects on intervention strategies.

Figure 2: Flowchart of vDiveR.


1
Home
• Landing page
• Introduction to DiveR and diversity motifs
8
Help
• Contact / bug report
• Frequently asked questions
2
Data Description
• Input data requirement and description
• Sample dataset (selected Hepatitis C virus proteins)
Input
Input Parameters
▪ Host name
▪ Size of k-mer
▪ Protein name
▪ Support threshold
▪ Sequence type
Display Parameters
▪ Host number selection
▪ Font size
▪ Line and dot size
▪ Protein names in order
Legend
Input – aligned sequences file
Input – DiMA JSON output file
3
Entropy and Incidence of Total Variants
Area and line plots of entropy and incidence of total variants across proteins
4
Correlation of Entropy and Total Variants
Scatter plot of distribution of k-mers across the incidence of total variants and entropy
5
Dynamics of Diversity Motifs (Proteome)
Scatter and violin plots of diversity motifs (index, major, minor and unique) across the proteome
6
Dynamic of Diversity Motifs (Protein)
Scatter and violin plots of diversity motifs (index, major, minor and unique) across individual protein(s)
7
Distribution of Conservation Levels
Box and scatter plots of k-mer distribution across conservation levels (completely conserved, highly conserved, mixed variable, highly diverse and extremely diverse)

1.2 Terminology

1.2.1 Diversity Motifs

The figure below (Figure. 3) depicts the definition of diversity motifs via a aligned nonamer.

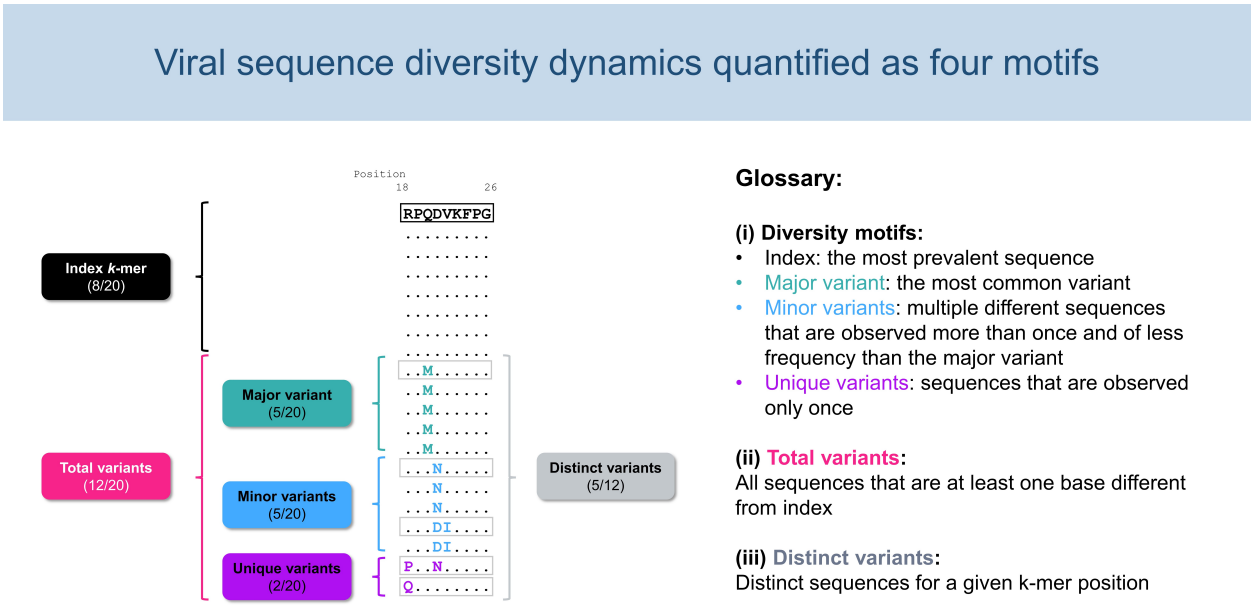


Figure 3. Definitions of diversity motifs.

1.2.2 Conservation Levels

k-mers are distributed into any of these conservation levels based on their index incidence value:

- completely conserved (black) (index incidence = 100%),
- highly conserved (blue) (90% index incidence < 100%),
- mixed variable (green) (20% < index incidence 90%),
- highly diverse (purple) (10% < index incidence 20%) and
- extremely diverse (pink) (index incidence 10%).

1.3 vDiveR Usage

1.3.1 R Shiny App

Hint: You may watch the demonstration video on how to utilize vDiveR R Shiny App here!

Upload your aligned FASTA / DiMA (v4.1.1) JSON / JSON-converted CSV output file(s) at the *Input Data Description* tab of vDiveR. There are **five input parameters** (Figure. 4):

- **Host Name**: Species name of the organism host to the studied virus.
- **Size of k-mer**: *k*-mer, a window with size of *k*, gives us the overview, overall diversity of that particular window. By default, DiMA uses *k*-mer size of nine to evaluate the viral diversity, with respect to cellular immune response.
- **Protein Name(s)**: Name of the protein.
- **Support Threshold**: Support is defined as the number of sequences at a given *k*-mer position that are free of gaps, unknown or ambiguous nucleotide bases, and amino acid residues. Positions with less than 30 sequences (default) are defined as of low support.
- **Sequence Type**: Nucleotide or amino acid (default) sequence.

Other than that, vDiveR allows user to manipulate **display parameters** (Figure. 4), such as:

- **Host Number Selection**: Select the number of host studied (one (default) or two hosts). vDiveR supports co-visualization of viral diversity dynamics between two hosts.
- **Font Size**: Font size displayed on the plots.
- **Line and Dot Size**: Line and dot size displayed on the plots.
- **Protein Names in Order**: Determine the order of proteins displayed on plot (Please ensure the protein names provided are the same as the one used in input run!).

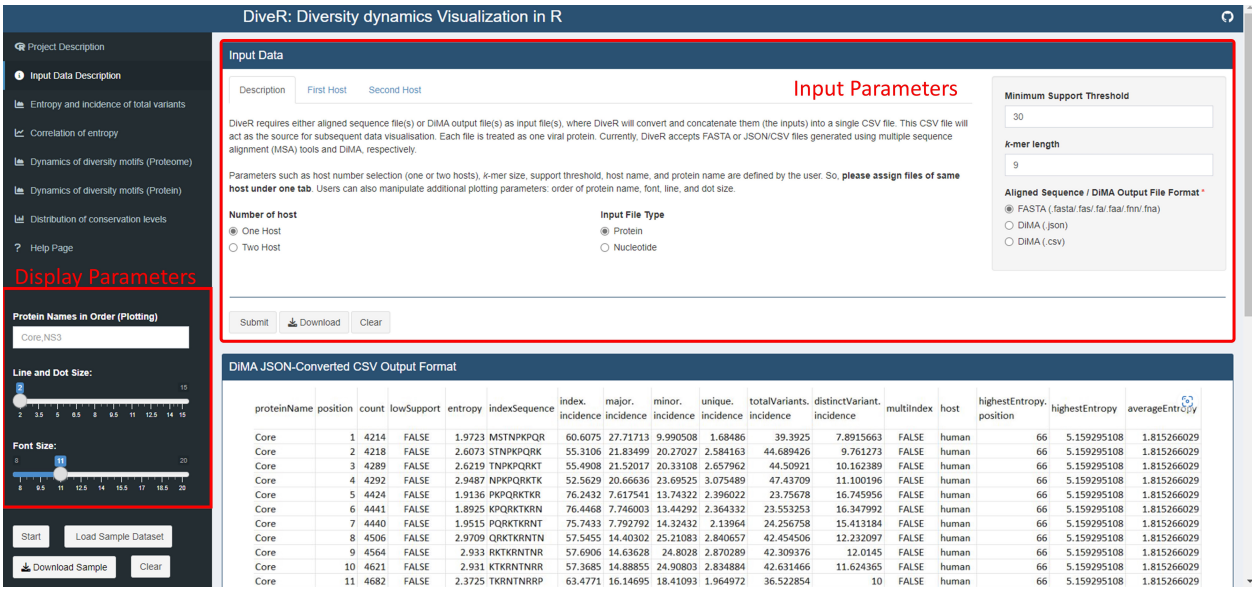


Figure 4. Location of the input and display parameters at vDiveR R Shiny App.

1.3.2 Bioconductor Package

There are seven functions provided:

1. **json2csv()**: convert DiMA (v4.1.1) JSON output to JSON-converted CSV dataframe, which will act as the data source for other functions in vDiveR.
2. **plot_incidence()**: plot entropy and total variant incidence.
3. **plot_entropy()**: plot entropy.
4. **plot_correlation()**: plot correlation between entropy and total variant incidence.
5. **plot_dynamics_proteome()**: plot dynamics of diversity motifs at proteome level (not recommended if the studied proteins do not represent the entire proteome).
6. **plot_dynamics_protein()**: plot dynamics of diversity motifs at protein level.
7. **plot_conservationLevel(): plot conservation levels distribution of k-mer positions, which consists of:**
    - completely conserved (index incidence = 100%; black),
    - highly conserved (90% index incidence < 100%; blue),
    - mixed variable (20% < index incidence 90%; green),
    - highly diverse (10% < index incidence 20%; purple), and
    - extremely diverse (index incidence 10%; pink).
8. **concat_conserved_kmer()**: concatenate completely/highly conserved k-mer positions that overlapped at least one k-mer position or are adjacent to each other and generate the output in dataframe that suits either CSV or FASTA format.

1.3.2.1 Usage

1.json2csv()

```
#default arguments
json2csv(json_data, hostName = "unknown host", proteinName = "unknown protein")
#example
inputdf<-json2csv(JSONsample)
```

Arguments:

- json_data: DiMA JSON output dataframe
- hostName: name of the host species
- proteinName: name of the protein

2.plot_incidence()

```
#default arguments
plot_incidence(df,host = 1,proteinOrder = "",kmer_size = 9,ymax = 10,line_dot_size = 2,
↪wordsize = 8)

#example 1 (1 host)
plot_incidence(proteins_1host)
#example 2 (2 hosts)
plot_incidence(protein_2hosts, host = 2)
```

Arguments:

- df: DiMA JSON converted csv file data
- host: number of host (1/2)
- proteinOrder: order of proteins displayed in plot
- kmer_size: size of the *k*-mer window
- ymax: maximum y-axis
- line_dot_size: size of the line and dot in plot
- wordsize: size of the wordings in plot

2.plot_entropy()

```
#default arguments
plot_entropy(df,host = 1,proteinOrder = "",kmer_size = 9,ymax = 10,line_dot_size = 2,
↪wordsize = 8)

#example 1 (1 host)
plot_entropy(proteins_1host)
#example 2 (2 hosts)
plot_entropy(protein_2hosts, host = 2)
```

Arguments:

- df: DiMA JSON converted csv file data
- host: number of host (1/2)
- proteinOrder: order of proteins displayed in plot
- kmer_size: size of the *k*-mer window
- ymax: maximum y-axis
- line_dot_size: size of the line and dot in plot
- wordsize: size of the wordings in plot

3.plot_correlation()

```
#default arguments
plot_correlation(df,host = 1,alpha = 1/3,size = 3,ylabel = "k-mer entropy (bits)\n",
↪xlabel = "\nTotal variants (%)",ymax = ceiling(max(df$entropy)),ybreak = 0.5)

#example 1 (1 host)
plot_correlation(proteins_1host)
#example 2 (2 hosts)
plot_correlation(protein_2hosts, size = 2, ybreak=1, ymax=10, host = 2)
```

Arguments:

- df: DiMA JSON converted csv file data
- host: number of host (1/2)
- alpha: any number from 0 (transparent) to 1 (opaque)
- size: dot size in scatter plot
- ylabel: y-axis label

- xlabel: x-axis label
- ymax: maximum y-axis
- ybreak: y-axis breaks

4.plot_dynamics_proteome()

```
#default arguments
plot_dynamics_proteome(df,host = 1,dot_size = 2,word_size = 15,alpha = 1/3)

#example 1 (1 host)
plot_dynamics_proteome(proteins_1host)
#example 2 (2 hosts)
plot_dynamics_proteome(protein_2hosts, host = 2)
```

Arguments:

- df: DiMA JSON converted csv file data
- host: number of host (1/2)
- dot_size: dot size in scatter plot
- word_size: word size in plot
- alpha: any number from 0 (transparent) to 1 (opaque)

5.plot_dynamics_protein()

```
#default arguments
plot_dynamics_protein(df,host = 1,proteinOrder = "",base_size = 8,alpha = 1/3,dot_size =␣
↪3)

#example 1 (1 host)
plot_dynamics_protein(proteins_1host)
#example 2 (2 hosts)
plot_dynamics_protein(protein_2hosts, host = 2)
```

Arguments:

- df: DiMA JSON converted csv file data
- host: number of host (1/2)
- proteinOrder: order of proteins displayed in plot
- base_size: base font size in plot
- alpha: any number from 0 (transparent) to 1 (opaque)
- dot_size: dot size in scatter plot

6.plot_conservationLevel()

```
#default arguments
plot_conservationLevel(df,proteinOrder = "",conservationLabel = 1,host = 1,base_size =␣
↪11,label_size = 2.6,alpha = 0.6)

#example 1 (1 host)
plot_conservationLevel(proteins_1host, conservationLabel = 1,alpha=0.8, base_size = 15)
```

(continues on next page)

(continued from previous page)

```
#example 2 (2 hosts)
plot_conservationLevel(protein_2hosts, conservationLabel = 0, host=2)
```

Arguments:

- df: DiMA JSON converted csv file data
- proteinOrder: order of proteins displayed in plot
- conservationLabel: 0 (partial; show present conservation labels only) or 1 (full; show ALL conservation labels) in plot
- host: number of host (1/2)
- base_size: base font size in plot
- label_size: conservation labels font size
- alpha: any number from 0 (transparent) to 1 (opaque)

7.concat_conserved_kmer()

```
#default arguments
concat_conserved_kmer(data,conservationLevel = "HCS",kmer = 9,output_type = "csv")

#example 1 (1 host and store the output in csv format)
csv<-concat_conserved_kmer(proteins_1host)
#example 1 (1 host and store the HCS output in FASTA format)
fasta <- concat_conserved_kmer(protein_2hosts, output_type = "fasta", conservationLevel␣
↪= "HCS")
#example 2 (2 hosts)
csv_2hosts<-concat_conserved_kmer(protein_2hosts, conservationLevel = "CCS")
```

Arguments:

- data: DiMA JSON converted csv file data
- conservationLevel: CCS (completely conserved) / HCS (highly conserved)
- kmer: size of the *k*-mer window
- output_type: type of the output; “csv” or “fasta”

1.4 R Shiny App

1.4.1 Input File

vDiveR requires either aligned sequence file(s) or DiMA output file(s) (JSON format) as input file(s), where DiveR will convert and concatenate them (the inputs) into a single CSV file (Figure. 5), which will act as the source for subsequent data visualisation. Each aligned sequence / DiMA output file is treated as one viral protein. Currently, vDiveR accepts aligned FASTA or JSON files generated using multiple sequence alignment (MSA) tools and DiMA, respectively.

Figure 5. Input CSV file format.

1. **proteinName**: name of the protein.
2. **position**: starting position of the aligned, overlapping *k*-mer window.
3. **count**: number of *k*-mer sequences at the given position.

proteinName	position	count	lowSupport	entropy	indexSequence	index. incidence	major. incidence	minor. incidence	unique. incidence	totalVariants. incidence	distinctVariant. incidence	multiIndex	host	highestEntropy. position	highestEntropy	averageEntropy
Core	1	4214	FALSE	1.9723	MSTNPKPQR	60.6075	27.71713	9.990508	1.68486	39.3925	7.8915663	FALSE	human	66	5.159295108	1.815266029
Core	2	4218	FALSE	2.6073	STNPKPQRK	55.3106	21.83499	20.27027	2.584163	44.689426	9.761273	FALSE	human	66	5.159295108	1.815266029
Core	3	4289	FALSE	2.6219	TNPKPQRKT	55.4908	21.52017	20.33108	2.657962	44.50921	10.162389	FALSE	human	66	5.159295108	1.815266029
Core	4	4292	FALSE	2.9487	NPKPQRKTK	52.5629	20.66636	23.69525	3.075489	47.43709	11.100196	FALSE	human	66	5.159295108	1.815266029
Core	5	4424	FALSE	1.9136	PKPQRKTKR	76.2432	7.617541	13.74322	2.396022	23.75678	16.745956	FALSE	human	66	5.159295108	1.815266029
Core	6	4441	FALSE	1.8925	KPQRKTKRN	76.4468	7.746003	13.44292	2.364332	23.553253	16.347992	FALSE	human	66	5.159295108	1.815266029
Core	7	4440	FALSE	1.9515	PQRKTKRNT	75.7433	7.792792	14.32432	2.13964	24.256758	15.413184	FALSE	human	66	5.159295108	1.815266029
Core	8	4506	FALSE	2.9709	QRKTKRNTN	57.5455	14.40302	25.21083	2.840657	42.454506	12.232097	FALSE	human	66	5.159295108	1.815266029
Core	9	4564	FALSE	2.933	RKTKRNTNR	57.6906	14.63628	24.8028	2.870289	42.309376	12.0145	FALSE	human	66	5.159295108	1.815266029
Core	10	4621	FALSE	2.931	KTKRNTNRR	57.3685	14.88855	24.90803	2.834884	42.631466	11.624365	FALSE	human	66	5.159295108	1.815266029
Core	11	4682	FALSE	2.3725	TKRNTNRRP	63.4771	16.14695	18.41093	1.964972	36.522854	10	FALSE	human	66	5.159295108	1.815266029
Core	12	4748	FALSE	2.8811	KRNTNRRPQ	55.1179	15.33277	27.4642	2.085088	44.882057	9.103707	FALSE	human	66	5.159295108	1.815266029
Core	13	4772	FALSE	2.6451	RNTNRRPQD	58.1727	14.60604	25.54485	1.676446	41.827328	8.316633	FALSE	human	66	5.159295108	1.815266029
Core	14	4878	FALSE	2.6588	NTNRRPQDV	57.9336	13.71464	26.67077	1.681017	42.06642	7.94347	FALSE	human	66	5.159295108	1.815266029
Core	15	5103	FALSE	2.7021	TNRRPQDVK	57.8875	13.03155	27.31726	1.763668	42.11248	7.910656	FALSE	human	66	5.159295108	1.815266029
Core	16	5127	FALSE	2.5627	NRRPQDVKF	58.8648	13.71172	25.92159	1.501853	41.13517	6.9701276	FALSE	human	66	5.159295108	1.815266029
Core	17	5207	FALSE	1.5631	RRPQDVKFP	77.0117	9.141541	12.80968	1.037065	22.988285	8.103592	FALSE	human	66	5.159295108	1.815266029
Core	18	5421	FALSE	1.5407	RPQDVKFPG	77.1075	8.891349	13.07877	0.922339	22.892456	7.8968577	FALSE	human	66	5.159295108	1.815266029
Core	19	5440	FALSE	1.5524	PQDVKFPGG	77.0588	8.860293	13.16176	0.919118	22.941177	7.852564	FALSE	human	66	5.159295108	1.815266029
Core	20	5442	FALSE	1.597	QDVKFPGGG	76.9019	8.857038	13.1018	1.139287	23.098125	8.67144	FALSE	human	66	5.159295108	1.815266029
Core	21	5457	FALSE	0.9726	DVKFPGGGQ	87.9971	5.2593	5.515851	1.227781	12.002933	16.183207	FALSE	human	66	5.159295108	1.815266029
Core	22	5465	FALSE	0.7102	VKFPGGGQI	92.882	1.591949	4.336688	1.189387	7.118024	26.73522	FALSE	human	66	5.159295108	1.815266029
Core	23	5474	FALSE	0.608	KFPGGGQIV	94.4282	0.858604	3.544026	1.169163	5.5717936	34.42623	FALSE	human	66	5.159295108	1.815266029
Core	24	5564	FALSE	0.5325	FPGGGQIVG	95.3451	0.305536	3.199137	1.150252	4.6549244	39.76834	FALSE	human	66	5.159295108	1.815266029

4. **lowSupport**: *k*-mer position with sequences lesser than the minimum support threshold (TRUE) are considered of low support, in terms of sample size.
5. **entropy**: level of variability at the *k*-mer position, with zero representing completely conserved.
6. **indexSequence**: the predominant sequence (index motif) at the given *k*-mer position.
7. **index.incidence**: the fraction (in percentage) of the index sequences at the *k*-mer position.
8. **major.incidence**: the fraction (in percentage) of the major sequence (the predominant variant to the index) at the *k*-mer position.
9. **minor.incidence**: the fraction (in percentage) of minor sequences (of frequency lesser than the major variant, but not singletons) at the *k*-mer position.
10. **unique.incidence**: the fraction (in percentage) of unique sequences (singletons, observed only once) at the *k*-mer position.
11. **totalVariants.incidence**: the fraction (in percentage) of sequences at the *k*-mer position that are variants to the index (includes: major, minor and unique variants).
12. **distinctVariant.incidence**: incidence of the distinct *k*-mer peptides at the *k*-mer position.
13. **multiIndex**: presence of more than one index sequence of equal incidence.
14. **host**: species name of the organism host to the virus.
15. **highestEntropy.position**: *k*-mer position that has the highest entropy value.
16. **highestEntropy**: highest entropy values observed in the studied protein.
17. **averageEntropy**: average entropy values across all the *k*-mer positions.

1.4.2 Sample Results

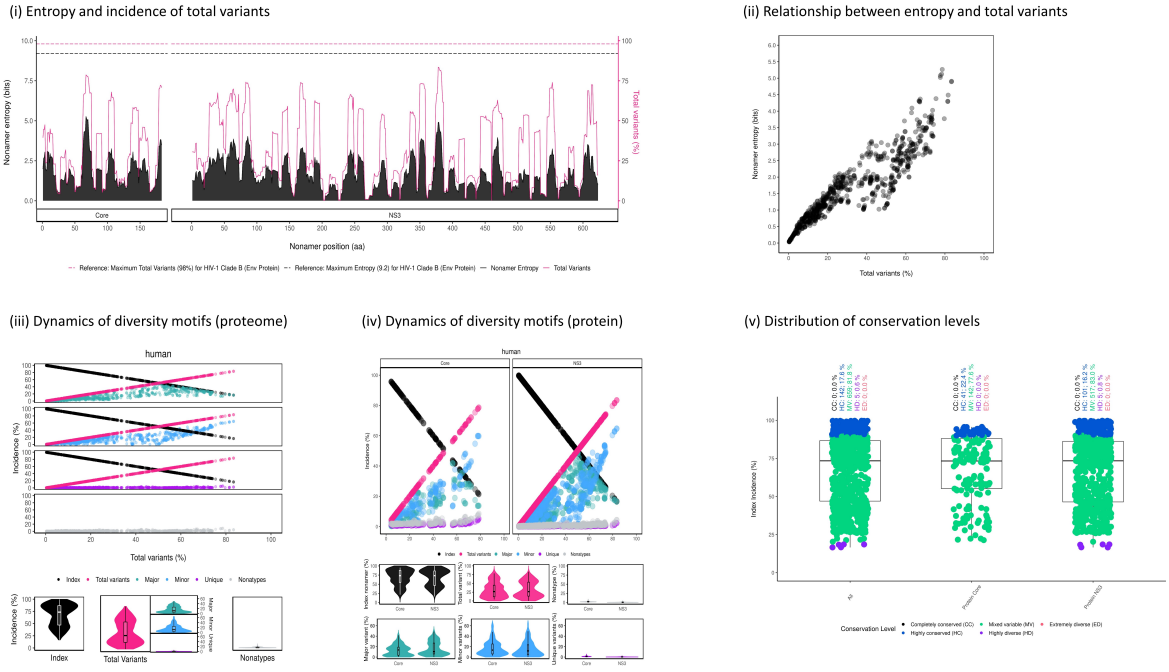


Figure 6: An example of vDiveR output, comprising of five plots for sample HCV proteins (Core and NS3). (i) Entropy and incidence of total variants for each aligned nonamer (*k*-mer) position of sample proteins. Entropy (black) and incidence of total variants (pink) were measured for each aligned nonamer (nine amino acids; 9-mer) position (1-9, 2-10, etc.) of the sample proteins. The entropy values indicate the level of variability at the corresponding nonamer positions, with zero representing completely conserved positions (total variants incidence of 0%). (ii) Relationship between incidence of total variants and entropy for HCV proteome nonamer positions (both Core and NS3 collectively). A positive correlation was observed. The nonamer entropy increased as the total variants' incidence increased. No completely conserved (entropy and total variant incidence of zero) nonamer position was observed. (iii) and (iv) Peptides at each of the 9-mer positions were classified into four different motifs, namely index, major, minor and unique, based on their incidences. Nonatypes is defined as the fraction of distinct sequences among the variants at a given position. The diversity spectrum of the *k*-mer position was depicted by the decline of the index incidences (black) and the increase of total variants incidences (pink). (v) Conservation levels of HCV nonamer positions for each individual protein and across the proteome. Both Core and NS3 proteins exhibited highly conserved (90% index incidence < 100%; blue) and mixed variable (20% index incidence < 90%; green) nonamer positions, while NS3 also included the highly diverse positions (10% index incidence < 20%; purple). No completely conserved (index incidence = 100%; black) and extremely diverse (index incidence < 10%; pink) nonamer positions were observed, indicating that the two proteins are of mixed variability.

1.4.2.1 Test Data

To demonstrate the functionality of vDiveR, the Core and NS3 proteins of Hepatitis C virus (HCV) were selected and used as sample datasets (Figure. 6). The human host HCV viral protein sequences were retrieved from the publicly available database, National Center for Biotechnology Information (NCBI) Virus (Hatcher et al., 2017). Subsequently, the data was deduplicated using Cluster Database at High Identity with Tolerance (CD-HIT) (Li & Godzik, 2006) and aligned using Multiple Alignment using Fast Fourier Transform (MAFFT) (Katoh et al., 2002). The HCV sample datasets are provided for users to download and run the visualization of sequence change dynamics in vDiveR.

Hint: Sample result is accessible on vDiveR R Shiny App via the **"Load Sample Dataset"** and **"Download Sample**

Dataset" buttons on its side panel.

1.4.3 Output Summary

In vDiveR R Shiny App, after providing either aligned sequence file(s) or DiMA JSON output file(s) in tab 2, visualization of dynamics in sequence change in the form of plots will be presented in tabs 3 to 7, with a brief description of the implemented functionalities (Figure 1).

- Tab 3: Entropy and Incidence of Total Variants
- Tab 4: Correlation of Entropy and Total Variants
- Tab 5: Dynamics of Diversity Motifs (Proteome)
- Tab 6: Dynamics of Diversity Motifs (Proteins)
- Tab 7: Distribution of Conservation Levels

Note: If there is only one protein input, no plot is shown in Tab 5.

1.4.3.1 Output (Plots and Tables)

1. Entropy and Incidence of Total Variants

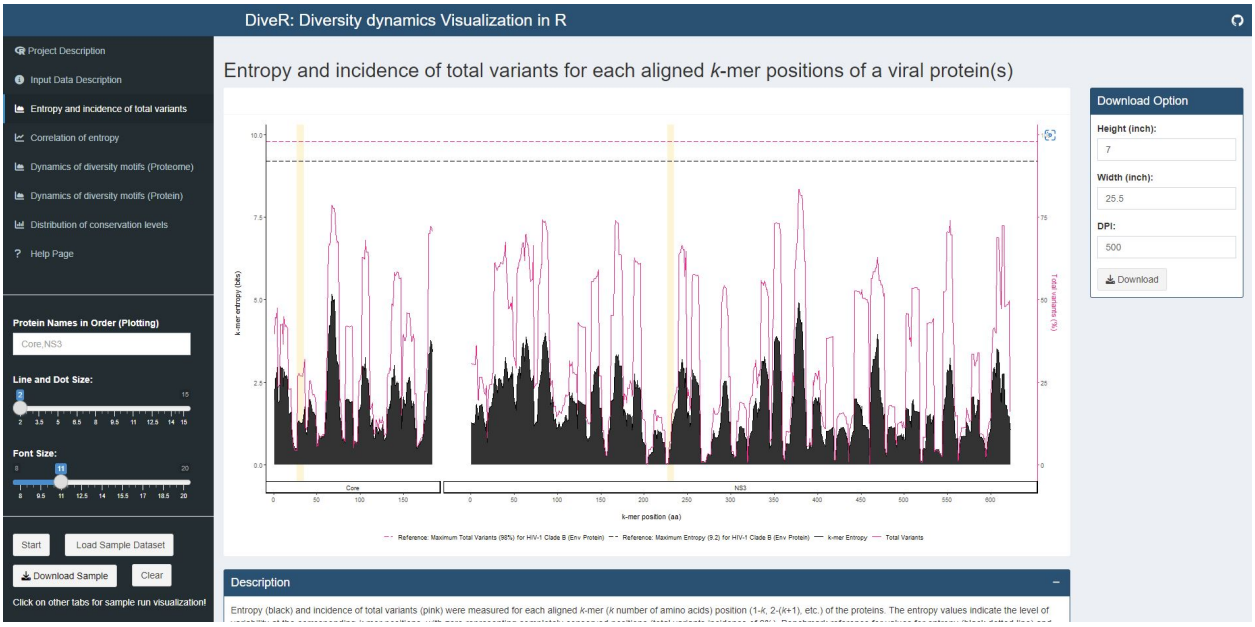


Figure 7.1. Entropy and incidence of total variants plot for sample HCV proteins (Core and NS3).

Entropy (black) and incidence of total variants (pink) were measured for each aligned *k*-mer position (1- *k*, 2- *k* +1, etc.) of the proteins. The entropy values indicate the level of variability at the corresponding *k*-mer positions, with zero representing completely conserved positions (total variants incidence of 0%). Benchmark reference for entropy (black dotted line; 9.2) and total variants (pink dotted line; 98%) from HIV-1 clade B envelope protein (Hu et al., 2013) are provided. For both individual protein and across proteome, the minimum entropy value is zero. The regions highlighted in yellow are *k*-mer positions with zero entropy value.

Figure 7.2. Entropy table for sample HCV proteins (Core and NS3).

A table with minimum and maximum values of entropy and total variants(%) values are provided for each studied protein. Positions that have the minimum entropy values are also provided.

2. Entropy

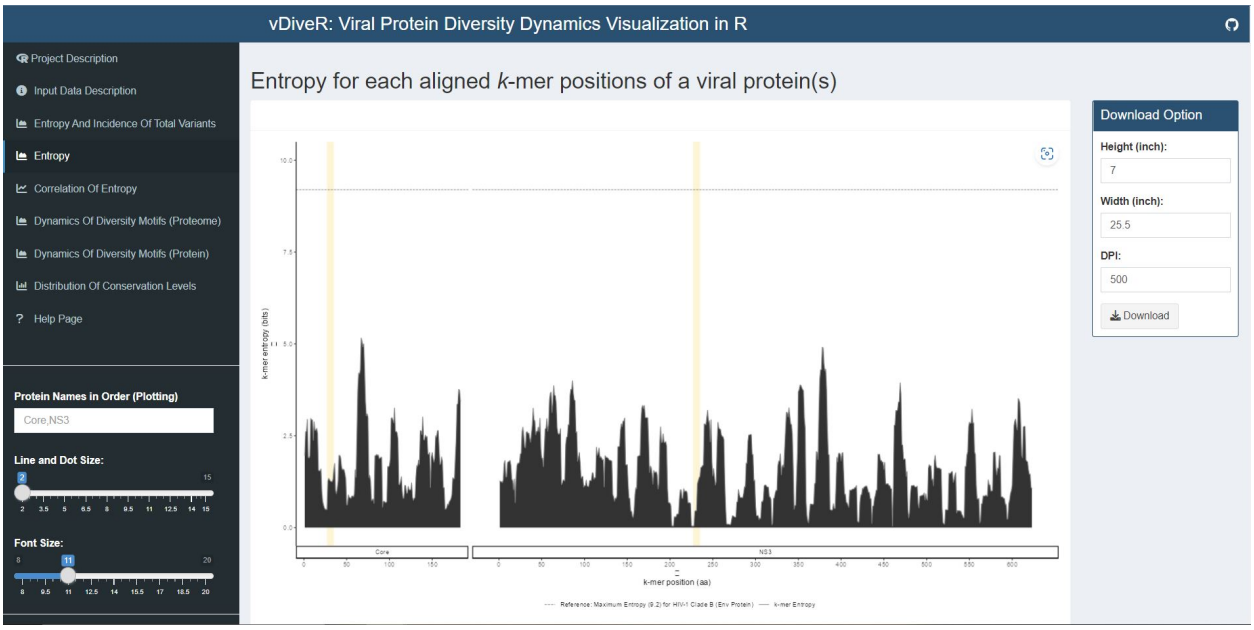


Figure 7.3. Entropy plot for sample HCV proteins (Core and NS3).

Entropy (black) was measured for each aligned k-mer position (1- k, 2- k +1, etc.) of the proteins. The entropy values indicate the level of variability at the corresponding k-mer positions, with zero representing completely conserved positions (total variants incidence of 0%). Benchmark reference for entropy (black dotted line; 9.2) from HIV-1 clade B envelope protein (Hu et al., 2013) is provided. For both individual protein and across proteome, the minimum entropy value is zero. The regions highlighted in yellow are k-mer positions with zero entropy value.

3. Correlation of Entropy

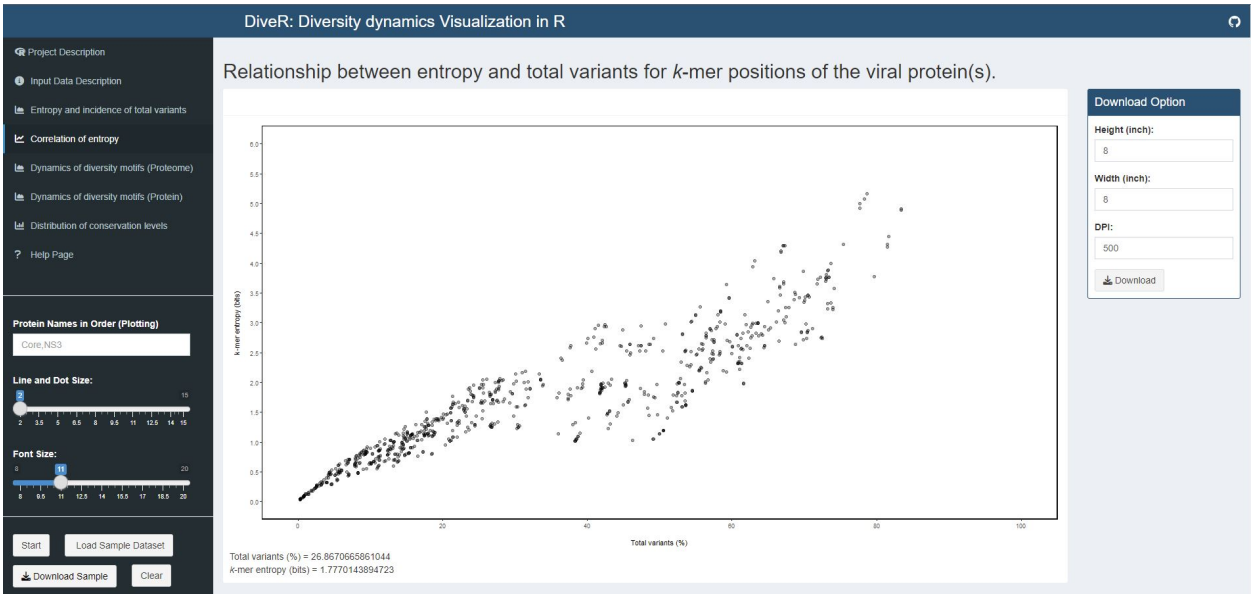
Figure 7.4. Correlation of entropy and total variants scatter plot for sample HCV proteins (Core and NS3).

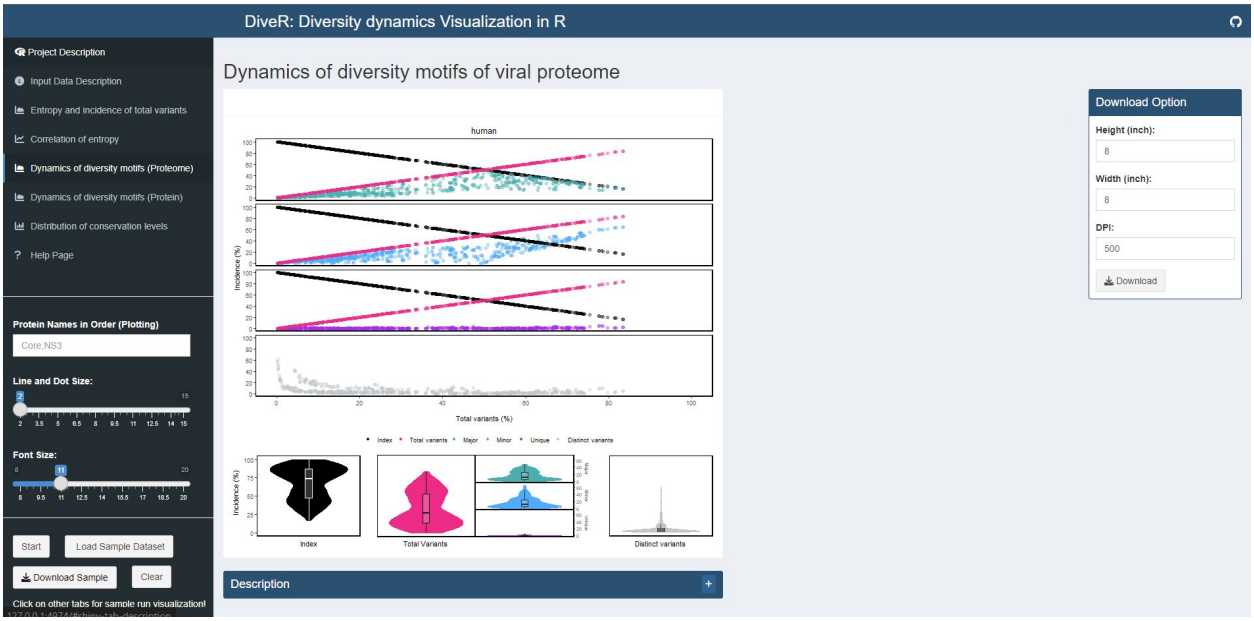
Relationship between incidence of total variants and entropy for viral proteome nonamer positions. At y-axis, the minimum entropy value is zero while the maximum entropy value is obtained by rounding the highest entropy encountered up to integer.

4. Dynamics of Diversity Motifs (Proteome)

Figure 7.5. Dynamics of diversity motifs (proteome) plot for sample HCV proteins (Core and NS3).

k-mers are classified into four different motifs, namely index, major, minor and unique, based on their incidences. Distinct variants is defined as distinct sequence for a given k-mer position. The above dot plot showcases the relationshop between the distribution of four distinct motifs and mutations. The diversity of the position is depicted by the decline


DiveR: Diversity dynamics Visualization in R
Project Description
Input Data Description
Entropy and incidence of total variants
Correlation of entropy
Dynamics of diversity motifs (Proteome)
Dynamics of diversity motifs (Protein)
Distribution of conservation levels
Help Page
Protein Names in Order (Plotting)
Core,NS3
Line and Dot Size:
Font Size:
Start
Load Sample Dataset
Download Sample
Clear
Relationship between entropy and total variants for k-mer positions of the viral protein(s).
k-mer entropy (bits)
Total variants (%)
Total variants (%) = 26.8670665861044
k-mer entropy (bits) = 1.7770143894723
Download Option
Height (inch):
8
Width (inch):
8
DPI:
500
Download


DiveR: Diversity dynamics Visualization in R
Project Description
Input Data Description
Entropy and incidence of total variants
Correlation of entropy
Dynamics of diversity motifs (Proteome)
Dynamics of diversity motifs (Protein)
Distribution of conservation levels
Help Page
Protein Names in Order (Plotting)
Core,NS3
Line and Dot Size:
Font Size:
Start
Load Sample Dataset
Download Sample
Clear
Click on other tabs for sample run visualization!
Dynamics of diversity motifs of viral proteome
human
Incidence (%)
Total variants (%)
Index
Total variants
Major
Minor
Unique
Distinct variants
Index
Total Variants
Distinct variants
Description
Download Option
Height (inch):
8
Width (inch):
8
DPI:
500
Download

of the index incidences (black), the increase of total variant incidences (pink) and corresponding individual patterns of the major, minor, unique motifs and distinct variants. The below violin plot demonstrates the frequency distribution of the motifs. The width of the plot (x-axis) represents the frequency distribution of a given incidence of the indicated motif. The black thick horizontal line of box plot in the middle represents the median incidence value.

5. Dynamics of Diversity Motifs (Protein(s))

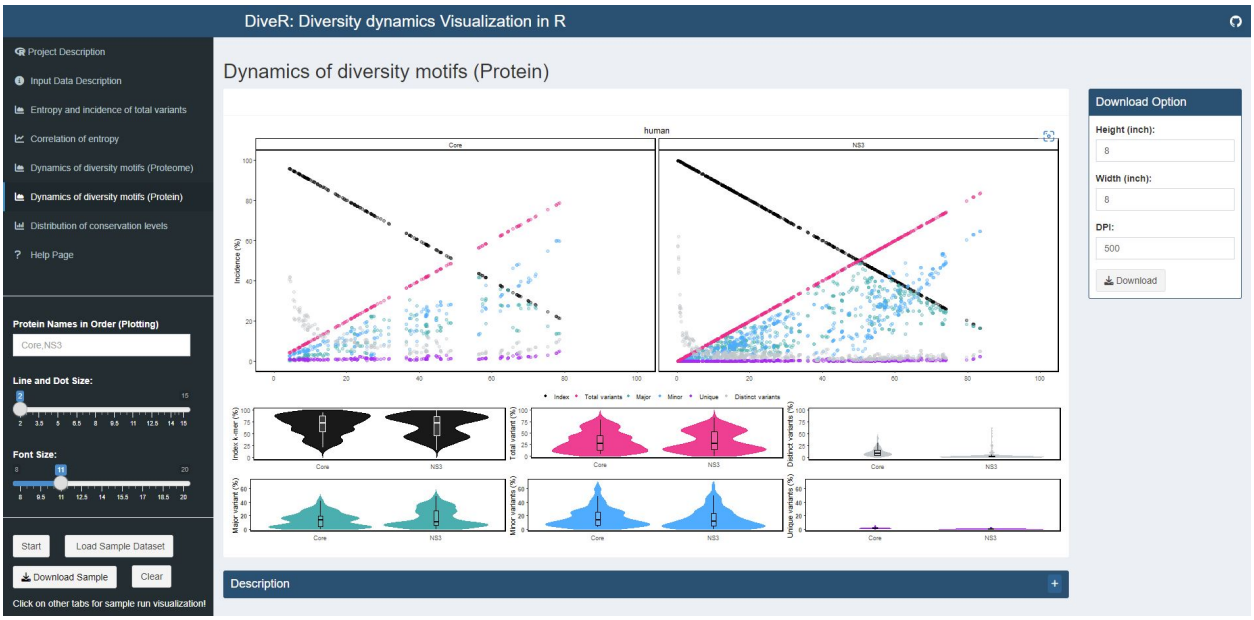


Figure 7.6. Dynamics of diversity motifs (proteins) plot for sample HCV proteins (Core and NS3).

k-mers are classified into four different motifs, namely index, major, minor and unique, based on their incidences. Distinct variants is defined as distinct sequence for a given *k*-mer position. The above dot plot showcases the relationshop between the distribution of four distinct motifs and mutations. The diversity of the position is depicted by the decline of the index incidences (black), the increase of total variant incidences (pink) and corresponding individual patterns of the major, minor, unique motifs and distinct variants. The below violin plot demonstrates the frequency distribution of the motifs. The width of the plot (x-axis) represents the frequency distribution of a given incidence of the indicated motif. The black thick horizontal line of box plot in the middle represents the median incidence value.

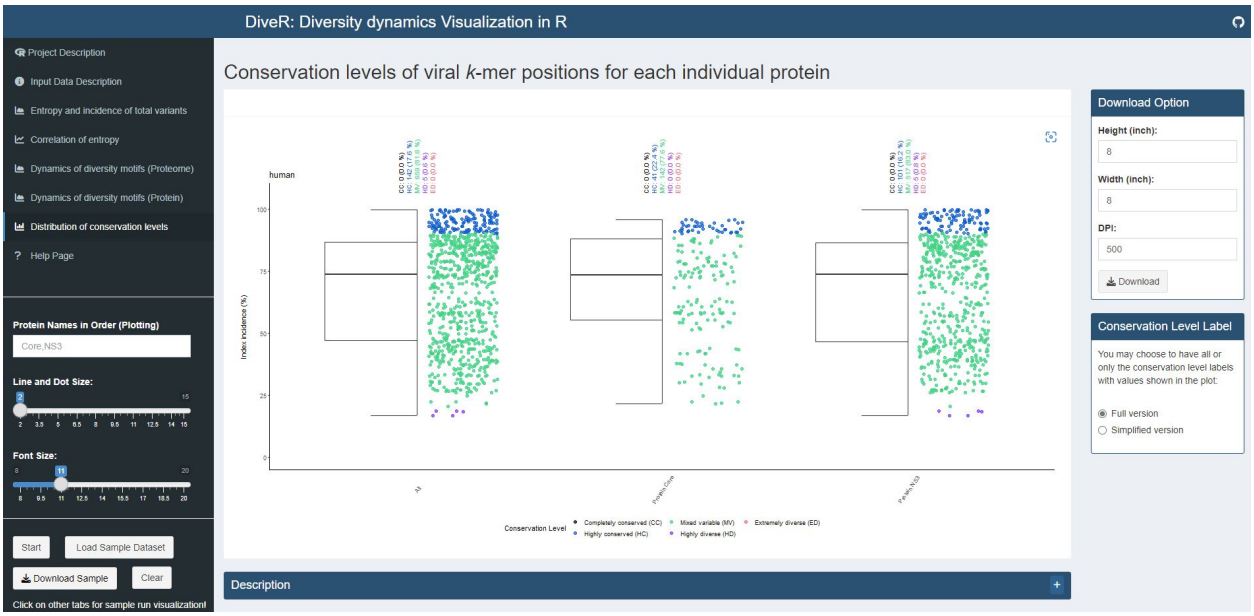
6. Distribution of Conservation Levels

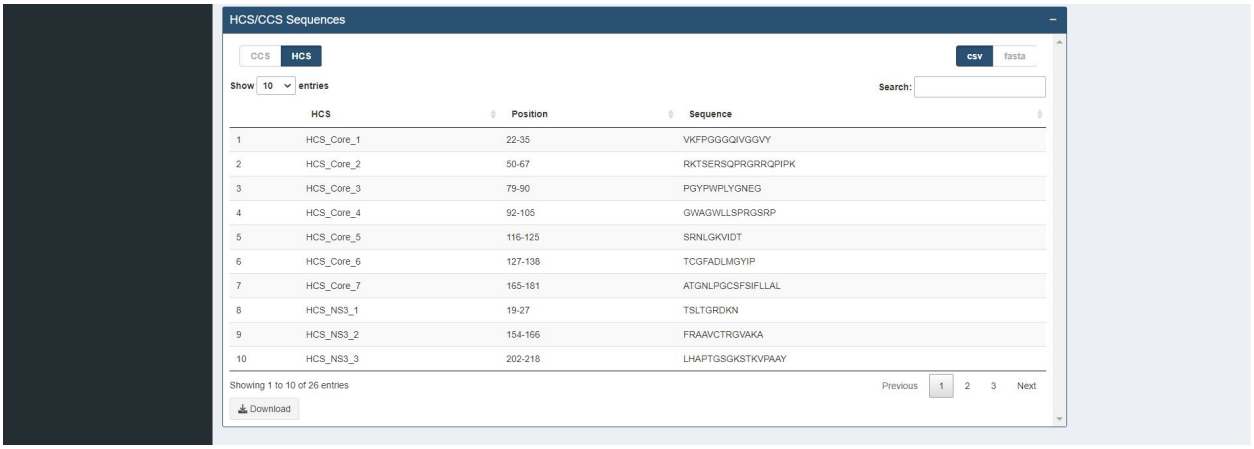
Figure 7.7. Distribution of conservation levels plot for sample HCV proteins (Core and NS3).

The *k*-mer positions of the proteome and the individual proteins were defined as completely conserved (black) (index incidence = 100%), highly conserved (blue) (90% index incidence < 100%), mixed variable (green) (20% < index incidence 90%), highly diverse (purple) (10% < index incidence 20%) and extremely diverse (pink) (index incidence 10%).

Figure 7.8. Identification of completely (CCS) / highly conserved (HCS) sequences table for sample HCV proteins (Core and NS3).

The *k*-mer positions that overlapped at least one *k*-mer position or are adjacent to each other are concatenated and displayed in table format. The concatenated sequences can be used for further immune relevance analysis via the usage of the Immune Epitope Database and Analysis Resource (IEDB) (Vita et al., 2019).


DiveR: Diversity dynamics Visualization in R
Project Description
Input Data Description
Entropy and incidence of total variants
Correlation of entropy
Dynamics of diversity motifs (Proteome)
Dynamics of diversity motifs (Protein)
Distribution of conservation levels
Help Page
Protein Names in Order (Plotting)
Core;NS3
Line and Dot Size:
Font Size:
Start
Load Sample Dataset
Download Sample
Clear
Click on other tabs for sample run visualization!
Conservation levels of viral k-mer positions for each individual protein
human
Index Incidence (%)
Conservation Level
Completely conserved (CC)
Mixed variable (MV)
Extremely diverse (ED)
Highly conserved (HC)
Highly diverse (HD)
Download Option
Height (inch):
8
Width (inch):
8
DPI:
500
Download
Conservation Level Label
You may choose to have all or only the conservation level labels with values shown in the plot:
Full version
Simplified version
Description


HCS/CCS Sequences
CCS
HCS
csv
fasta
Show 10 entries
Search:
HCS
Position
Sequence
1 HCS_Core_1 22-35 VKFPGGGQIVGGVY
2 HCS_Core_2 50-67 RKTSERSQPRGRRQPIPK
3 HCS_Core_3 79-90 PGYPWPLYGNEG
4 HCS_Core_4 92-105 GWAGWLLSPRGSRP
5 HCS_Core_5 116-125 SRNLGKVIDT
6 HCS_Core_6 127-138 TCGFADLMGYIP
7 HCS_Core_7 165-181 ATGNLPGCSFSIFLLAL
8 HCS_NS3_1 19-27 TSLTGRDKN
9 HCS_NS3_2 154-166 FRAAVCTRGVAKA
10 HCS_NS3_3 202-218 LHAPTGSGKSTKVPAAY
Showing 1 to 10 of 26 entries
Previous
1
2
3
Next
Download

1.5 Bioconductor Package

1.5.1 Input

vDiveR Bioconductor package functions require a DiMA(v4.1.1) JSON-converted CSV dataframe as input. As DiMA stores its output in JSON format, vDiveR has also provided a JSON2CSV() function to assist users in converting DiMA output from JSON to CSV format. Each DiMA JSON output file is treated as one viral protein. If there is more than one protein to be visualized simultaneously, users are required to concatenate the CSV dataframes (Figure. 8) into one, which will eventually act as the source for subsequent data visualisation and *k*-mer concatenation.

proteinName	position	count	lowSupport	entropy	indexSequence	index. incidence	major. incidence	minor. incidence	unique. incidence	totalVariants. incidence	distinctVariant. incidence	multiIndex	host	highestEntropy. position	highestEntropy	averageEntropy
Core	1	4214	FALSE	1.9723	MSTNPKPQR	60.6075	27.71713	9.990508	1.68486	39.3925	7.8915663	FALSE	human	66	5.159295108	1.815266029
Core	2	4218	FALSE	2.6073	STNPKPQRK	55.3106	21.83499	20.27027	2.584163	44.689426	9.761273	FALSE	human	66	5.159295108	1.815266029
Core	3	4289	FALSE	2.6219	TNPKPQRKT	55.4908	21.52017	20.33108	2.657962	44.50921	10.162389	FALSE	human	66	5.159295108	1.815266029
Core	4	4292	FALSE	2.9487	NPKPQRKTK	52.5629	20.66636	23.69525	3.075489	47.43709	11.100196	FALSE	human	66	5.159295108	1.815266029
Core	5	4424	FALSE	1.9136	PKPQRKTKR	76.2432	7.617541	13.74322	2.396022	23.75678	16.745956	FALSE	human	66	5.159295108	1.815266029
Core	6	4441	FALSE	1.8925	KPQRKTKRN	76.4468	7.746003	13.44292	2.364332	23.553253	16.347992	FALSE	human	66	5.159295108	1.815266029
Core	7	4440	FALSE	1.9515	PQRKTKRNT	75.7433	7.792792	14.32432	2.13964	24.256758	15.413184	FALSE	human	66	5.159295108	1.815266029
Core	8	4506	FALSE	2.9709	QRKTKRNTN	57.5455	14.40302	25.21083	2.840657	42.454506	12.232097	FALSE	human	66	5.159295108	1.815266029
Core	9	4564	FALSE	2.933	RKTKRNTNR	57.6906	14.63628	24.8028	2.870289	42.309376	12.0145	FALSE	human	66	5.159295108	1.815266029
Core	10	4621	FALSE	2.931	KTKRNTNRR	57.3685	14.88855	24.90803	2.834884	42.631466	11.624365	FALSE	human	66	5.159295108	1.815266029
Core	11	4682	FALSE	2.3725	TKRNTNRRP	63.4771	16.14695	18.41093	1.964972	36.522854	10	FALSE	human	66	5.159295108	1.815266029
Core	12	4748	FALSE	2.8811	KRNTNRRPQ	55.1179	15.33277	27.4642	2.085088	44.882057	9.103707	FALSE	human	66	5.159295108	1.815266029
Core	13	4772	FALSE	2.6451	RNTNRRPQD	58.1727	14.60604	25.54485	1.676446	41.827328	8.316633	FALSE	human	66	5.159295108	1.815266029
Core	14	4878	FALSE	2.6588	NTNRRPQDV	57.9336	13.71464	26.67077	1.681017	42.06642	7.94347	FALSE	human	66	5.159295108	1.815266029
Core	15	5103	FALSE	2.7021	TNRRPQDVK	57.8875	13.03155	27.31726	1.763668	42.11248	7.910656	FALSE	human	66	5.159295108	1.815266029
Core	16	5127	FALSE	2.5627	NRRPQDVKF	58.8648	13.71172	25.92159	1.501853	41.13517	6.9701276	FALSE	human	66	5.159295108	1.815266029
Core	17	5207	FALSE	1.5631	RRPQDVKFP	77.0117	9.141541	12.80968	1.037065	22.988285	8.103592	FALSE	human	66	5.159295108	1.815266029
Core	18	5421	FALSE	1.5407	RPQDVKFPG	77.1075	8.891349	13.07877	0.922339	22.892456	7.8968577	FALSE	human	66	5.159295108	1.815266029
Core	19	5440	FALSE	1.5524	PQDVKFPGG	77.0588	8.860293	13.16176	0.919118	22.941177	7.852564	FALSE	human	66	5.159295108	1.815266029
Core	20	5442	FALSE	1.597	QDVKFPGGG	76.9019	8.857038	13.1018	1.139287	23.098125	8.67144	FALSE	human	66	5.159295108	1.815266029
Core	21	5457	FALSE	0.9726	DVKFPGGGQ	87.9971	5.2593	5.515851	1.227781	12.002933	16.183207	FALSE	human	66	5.159295108	1.815266029
Core	22	5465	FALSE	0.7102	VKFPGGGQI	92.882	1.591949	4.336688	1.189387	7.118024	26.73522	FALSE	human	66	5.159295108	1.815266029
Core	23	5474	FALSE	0.608	KFPGGGQIV	94.4282	0.858604	3.544026	1.169163	5.5717936	34.42623	FALSE	human	66	5.159295108	1.815266029
Core	24	5564	FALSE	0.5325	FPGGGQIVG	95.3451	0.305536	3.199137	1.150252	4.6549244	39.76834	FALSE	human	66	5.159295108	1.815266029

Figure 8. DiMA JSON-converted CSV dataframe format.

1. **proteinName**: name of the protein.
2. **position**: starting position of the aligned, overlapping *k*-mer window.
3. **count**: number of *k*-mer sequences at the given position.
4. **lowSupport**: *k*-mer position with sequences lesser than the minimum support threshold (TRUE) are considered of low support, in terms of sample size.
5. **entropy**: level of variability at the *k*-mer position, with zero representing completely conserved.
6. **indexSequence**: the predominant sequence (index motif) at the given *k*-mer position.
7. **index.incidence**: the fraction (in percentage) of the index sequences at the *k*-mer position.
8. **major.incidence**: the fraction (in percentage) of the major sequence (the predominant variant to the index) at the *k*-mer position.
9. **minor.incidence**: the fraction (in percentage) of minor sequences (of frequency lesser than the major variant, but not singletons) at the *k*-mer position.
10. **unique.incidence**: the fraction (in percentage) of unique sequences (singletons, observed only once) at the *k*-mer position.
11. **totalVariants.incidence**: the fraction (in percentage) of sequences at the *k*-mer position that are variants to the index (includes: major, minor and unique variants).
12. **distinctVariant.incidence**: incidence of the distinct *k*-mer peptides at the *k*-mer position.
13. **multiIndex**: presence of more than one index sequence of equal incidence.

14. **host**: species name of the organism host to the virus.
15. **highestEntropy.position**: *k*-mer position that has the highest entropy value.
16. **highestEntropy**: highest entropy values observed in the studied protein.
17. **averageEntropy**: average entropy values across all the *k*-mer positions.

1.5.1.1 Sample Dataset

To demonstrate the functionality of vDiveR, three sample datasets (JSONsample, proteins_1host, protein_2hosts) are provided where:

1. **JSONsample**: a DiMA JSON output file which acts as the input for JSON2CSV(),
2. **proteins_1host** (consists of protein A and B from human host) and **protein_2hosts** (consists of protein A from human and bat hosts): input for remaining functions.

1.5.1.2 Sample Output

Note: Please refer 'section 4. R Shiny App (Output)' for sample output.

1.6 FAQS

Note: For technical assistance or bug report, please reach us out via GitHub (https://github.com/pendy05/DiveR). For the general correspondence, please email Dr. Asif M. Khan (asif@perdanauniversity.edu.my, makhan@bezmialem.edu.tr).

1. What can I do if the elements in the plot appear to be overlapping each other due to the displayed plot size?

- You may want to increase the height and/or width of the plot offered in the download option, based on your need before downloading the plot.

2. Where can I get the source code of these R plots if I would like to modify the code based on my need?

- You may visit this GitHub repository (https://github.com/pendy05/DiveR) to get the corresponding source codes.

3. What is the maximum image size (in inches) that can be downloaded via vDiveR R Shiny App deployed on Shiny server?

- Maximum 50 (H) x50 (W) inches to prevent the common error of specifying dimensions in pixels encountered in R ggsave() function.

4. I encountered *'Error in x$clone: attempt to apply non-function'* in plot 'entropy and incidence of total variants' when I submit files for two hosts. Other plots work fine though. Why does this happen?

- vDiveR expects the proteins with same protein name have same length (number of positions) across both the hosts to carry out the comparison plot.

5. What should I do if I would like to run each function in vDiveR separately and locally?

- vDiveR will be soon released as a Bioconductor package. Tentatively, you may visit our GitHub repository (https://github.com/pendy05/DiveR) for its source code.