Immunome Browser maps and visualizes queried in IEDB linear peptidic epitopes along the length of a protein sequence (a target, or reference, protein). The purpose of the tool is to allow users exploring how often each protein region has been studied in immune assays and in how many assays the immune response was positive or negative.

Immunome Browser provide a summarized data by reference antigens; and the main reasons for doing that are the following: (i) epitopes reported in IEDB were identified for different strains and protein isoforms – mapping to the reference protein allows to visualize and study such epitopes as they would have the same antigen; (2) different mutant variants of the same epitope were tested and reported; and (3) immune response varies among studies and assays due to heterogeneity of samples and complexity of immune response.

Immunone Browser is accessible in IEDB via two entry points (or tabs): either the Antigen or Epitope navigation paths *[the latter is subject of release by 2015 and will provide functionality of changing mapping parameters and target proteins].*

To get started, run the following query:

and click on the Immunome Browser icon:

Each selected antigen is considered to be a target protein, and the tool by default maps and visualizes all epitopes in the query result that are from the source antigen which is the target protein or any of its taxonomic child. The mapping is done using the alignment between the source antigen and the target protein and matching the epitope positions in the source antigen with the residue positions in the target protein (* -- the default mapping criteria are provided below). **Note**, by default, epitopes that were tested as negative for the correspondent query will be mapped as well, even if the user queried only for positive assays. Also, the epitopes that failed to satisfy the default mapping criteria won't be mapped and visualized.
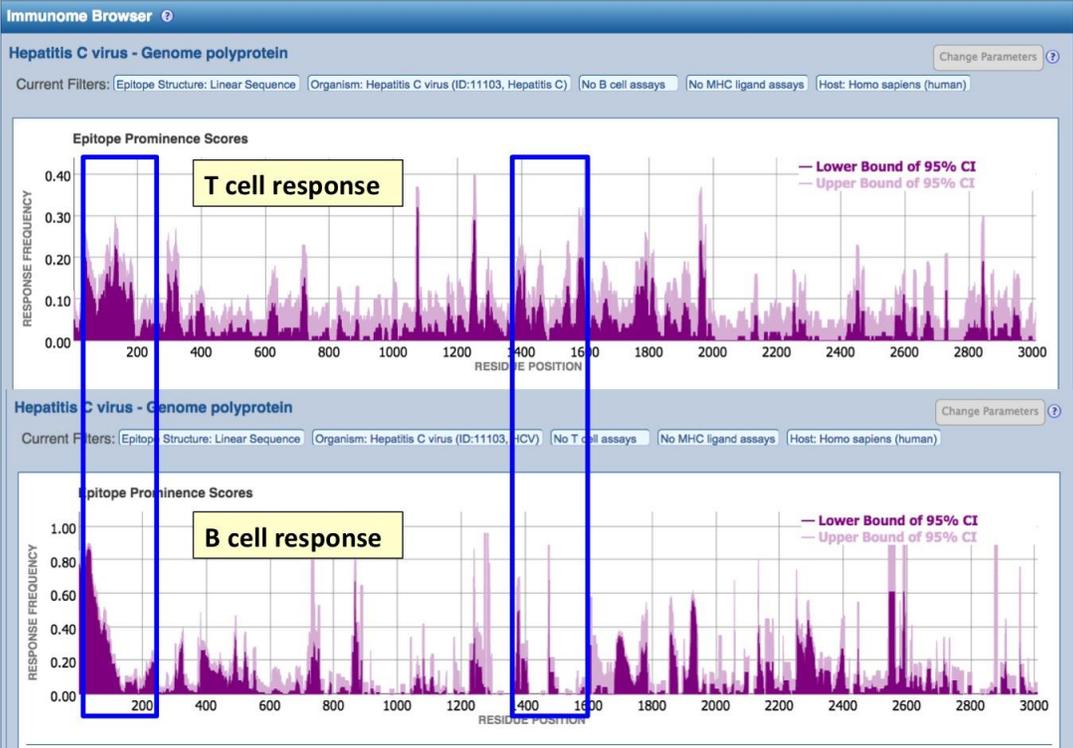
Immunome Browser outputs two plots and the table.

The tool retrieves negative epitopes for the antigen automatically even if the user queried for "positive assays only"
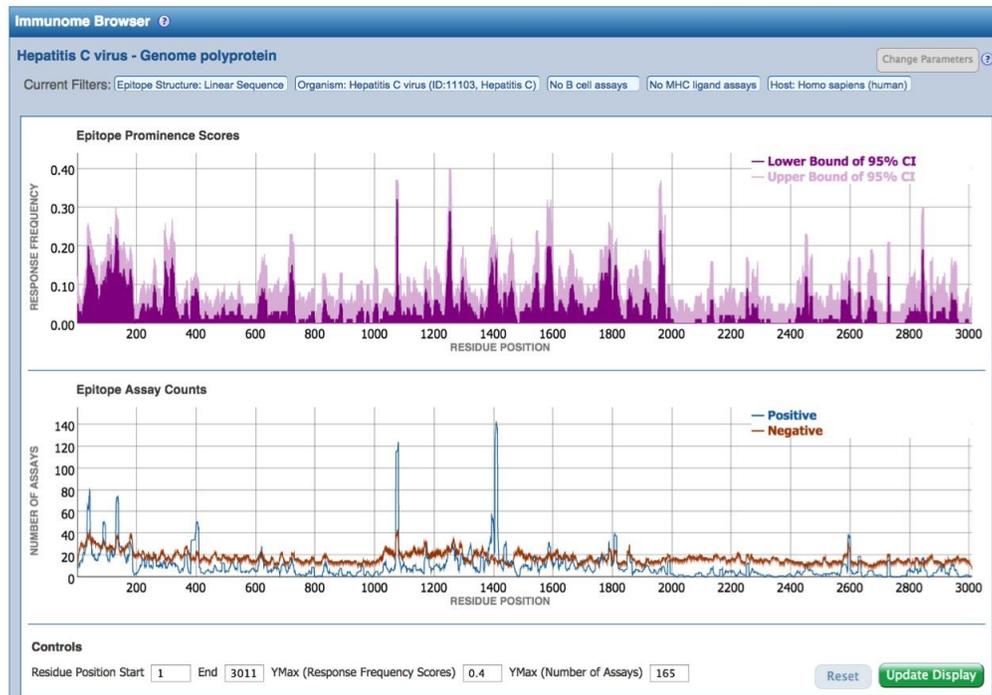
The results are query-specific! See the following figures for examples.

The results are query-specific: T cell vs. B cell response

T cell response: chronic vs. acute Hepatitis C
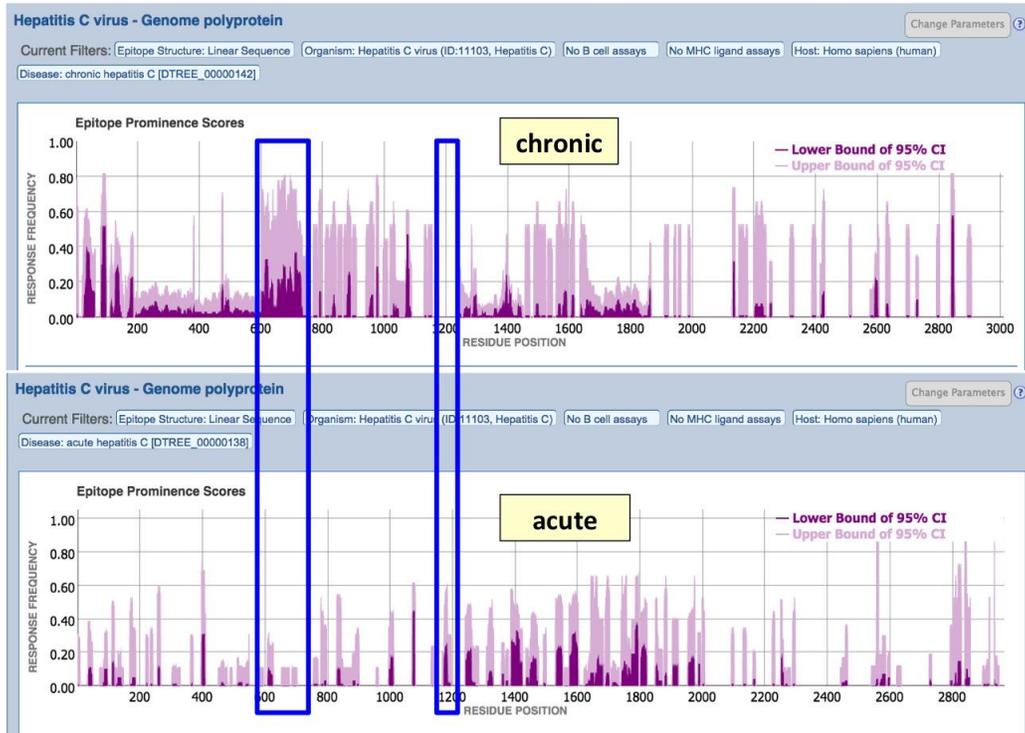
The upper plot renders the lower (dark color) and upper (light color) bounds of the **95% confidence interval (CI) of the response frequency (RF)** for each **position** target protein position, averaged over all epitopes mapped to that position. RF is calculated as the number of positively responded subjects (or individuals in this case) relative to the total number of those tested summed up by mapped epitopes. **Note**, however, that the graph displays only the upper and lower bounds of the 95% confidence interval (CI) for the RF score and not the RF score itself.

Values on the plot: summing up across epitopes

Epitope Prominence Scores

N= 170   R = 6   →   RF = R/N = 0.04, CI = [0.02; 0.07]

80:
Lower Bound of 95% CI: 0.02
Upper Bound of 95% CI: 0.07

Epitope ID	Epitope Sequence	Mapped Position	Identity	Subjects Tested	Subjects Responded	Assays Positive	Assays Negative	Response Freq. (95% CI)
74440	YKDVDKPPFSGM	66-77	100%			1	1	0.06 (0.00:0.16)
127443	YKDVDKPPFSGMTGC	66-80	100%	26	2	2	0	0.08 (0.01:0.17)
67999	VDKPPFSGMTGC	69-80	100%	17	1	1	1	0.06 (0.00:0.16)
126461	KPPFSGMTGCGNTPI	71-85	100%	26	3	2	0	0.12 (0.03:0.22)
48777	PPFSGMTGCGNT	72-83	100%	17	0	0	2	0.00 (0.00:0.00)
58142	SGMTGCGNTPIF	75-86	100%	17	0	0	2	0.00 (0.00:0.00)
126156	GMTGCGNTPIFKSGR	76-90	100%	50	0	0	2	0.00 (0.00:0.00)
63797	TGCGNTPIFKSG	78-89	100%	17	0	0	2	0.00 (0.00:0.00)
21534	GNTPIFKSGRGC	81-92	100%	17	1	1	1	0.06 (0.00:0.16)
126159	GNTPIFKSGRGCGSC	81-95	100%	50	0	0	2	0.00 (0.00:0.00)
47896	PIFKSGRGCGSC	84-95	100%	17	1	1	1	0.06 (0.00:0.16)

The bottom plot shows the number of positive and negative assays averaged over epitopes mapped to each position of the target protein.

Values on the plot: summing up across epitopes

Epitope ID	Epitope Sequence	Mapped Position	Identity	Subjects Tested	Subjects Responded	Assays Positive	Assays Negative	Response Freq. (95% CI)
74440	YKDVDKPPFSGM	66-77	100%	17	1	1	1	0.06 (0.00:0.16)
127443	YKDVDKPPFSGMTGC	66-80	100%	26	2	2	0	0.08 (0.01:0.17)
67999	VDKPPFSGMTGC	69-80	100%	17	1	1	1	0.06 (0.00:0.16)
126461	KPPFSGMTGCGNTPI	71-85	100%	26	3	2	0	0.12 (0.03:0.22)
48777	PPFSGMTGCGNT	72-83	100%	17	0	0	2	0.00 (0.00:0.00)
58142	SGMTGCGNTPIF	75-86	100%	17	0	0	2	0.00 (0.00:0.00)
126156	GMTGCGNTPIFKSGR	76-90	100%	50	0	0	2	0.00 (0.00:0.00)
63797	TGCGNTPIFKSG	78-89	100%	17	0	0	2	0.00 (0.00:0.00)
21534	GNTPIFKSGRGC	81-92	100%	17	1	1	1	0.06 (0.00:0.16)
126159	GNTPIFKSGRGCGSC	81-95	100%	50	0	0	2	0.00 (0.00:0.00)
47896	PIFKSGRGCGSC	84-95	100%	17	1	1	1	0.06 (0.00:0.16)

The table presents results **for each epitope**. The figures below provide explanation how values in the table are calculated.

How values in the table are calculated

- For each reference, **N (total tested subjects)** and **R (number of responded subjects)** are assigned the highest values from any positive assay (if there is no positive assay, *N* comes from a negative assay with the highest *N*).

- For each epitope, *N* and *R* values are summed up by references.

Epitope ID	Epitope Sequence	Mapped Position	Identity	Subjects Tested	Subjects Responded	Assays Positive	Assays Negative	Response Freq. (95% CI)
126664	MASSSSVLLVVALFA	1-15	100%	50	0	0	2	0.00 (0.00:0.00)
126665	MASSSSVLLVVVLFA	1-15	100%	50	0	0	2	0.00 (0.00:0.00)
74440	YKDVDKPPFSGM	66-77	100%	17	1	1	1	0.06 (0.00:0.16)
127443	YKDVDKPPFSGMTGC	66-80	100%	26	2	2	0	0.08 (0.01:0.17)
67999	VDKPPFSGMTGC	69-80	100%	17	1	1	1	0.06 (0.00:0.16)
126461	KPPFSGMTGCGNTPI	71-85	100%	26	3	2	0	0.12 (0.03:0.22)

Results Returned: 167 | Displaying: 167   Display Graphed Residue Positions   Export Results

 # Tested Subjects
N = 26

 # Responded Subjects
R = 2

RF = R/N =
2/26 = 0.08
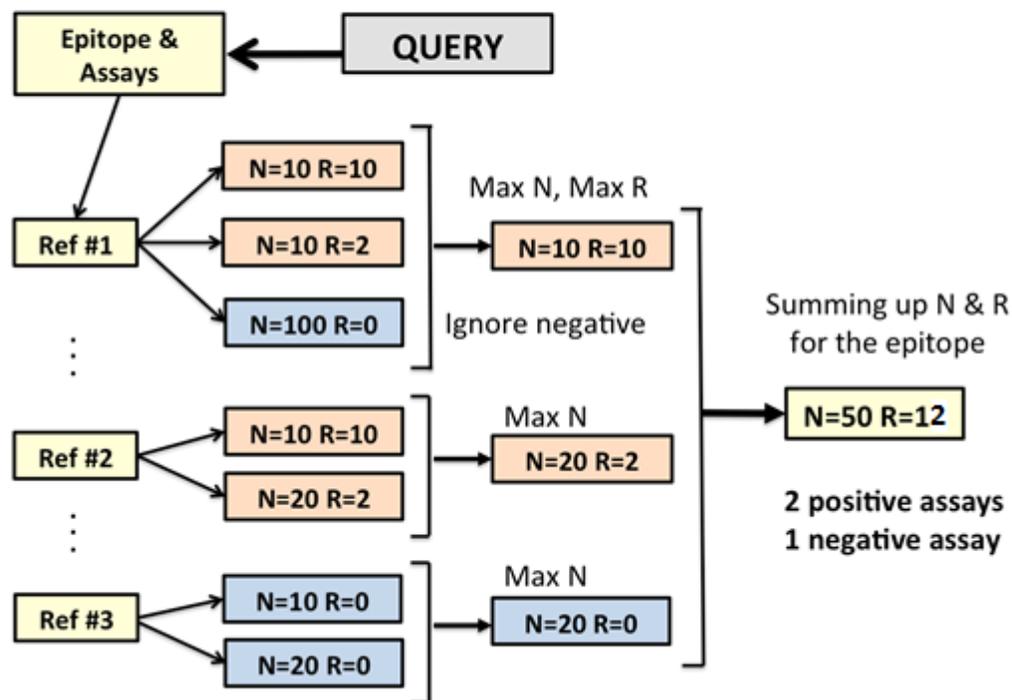
How values in the table are calculated (cont.)

- The statistical distribution of outcomes of either a B- or T-cell assay, or RF score, can be approximated by the binomial distribution.

- For each epitope, given the observed RF score, it is possible to estimate the 95% confidence interval for the true RF.

- This interval is a range of possible values that will contain the true RF value 95% of the times if the experiment is repeated an infinite number of times under the same conditions.

Epitope ID	Epitope Sequence	Mapped Position	Identity	Subjects Tested	Subjects Responded	Assays Positive	Assays Negative	Response Freq. (95% CI)
126664	MASSSSVLLVVALFA	1-15	100%	50	0	0	2	0.00 (0.00:0.00)
126665	MASSSSVLLVVVLFA	1-15	100%	50	0	0	2	0.00 (0.00:0.00)
74440	YKDVDKPPFSGM	66-77	100%	17	1	1	1	0.06 (0.00:0.16)
127443	YKDVDKPPFSGMTGC	66-80	100%	26	2	2	0	0.08 (0.01:0.17)
67999	VDKPPFSGMTGC	69-80	100%	17	1	1	1	0.06 (0.00:0.16)
126461	KPPFSGMTGCGNTPI	71-85	100%	26	3	2	0	0.12 (0.03:0.22)

Results Returned: 167 | Displaying: 167   Display Graphed Residue Positions   Export Results

95% Confidence Interval for RF score estimate

Values for each epitope are summed up by references

Note the essential difference: the plots depict values defined <u>for each residue position</u> in the target protein; while the table displays the values of the RF score and CI <u>for each epitope</u>.

The user can interactively zoom in and out the plots to a specific protein region and, if the check box for the table to correspond the plots is selected, the table is automatically updated to show only epitope from the selected region.

---

* Default mapping criteria:

- o Source protein and target alignment: NCBI Blast is used with default parameters. If less than a half of the length source protein were aligned or the epitope match didn't satisfy identity criteria, the needlman-wunsch algorithm for global alignment was used.

- o Epitope matching criteria:

  - Short epitopes of length <= 5 aa. Sequence identity = 100%, no gaps in the matching alignment.

  - Long epitope > 5 aa. Sequence identity >= 50%, maximum length fraction of gaps in the alignment (for each source and target) = 20%.

- Long epitopes (> 5 aa) that have multiple exact matches in the source protein. Sequence identity >= 70%, sequence similarity (based on BLOSUM62) = 100%, no gaps in the matching alignment.

** Response Frequency (RF) and Confidence Interval (CI) calculation:

N is the number of subjects tested, and **R is the number of subjects responded.**

If an assay lacks information on the number of subjects responded or tested, the following values of N and R are used:

R = 1, N = 1, if the assay outcome is positive;

R = 0, N = 1, if the assay outcome is negative.

Reporting CI instead of directly calculated RF *(RF = R/N)* allows correcting for the sample size. Otherwise, a peptide found to be positive in only one donor (*R=1; N=1*) would be considered more immunogenic than a peptide showing response in 99 out of 100 donors as the former has the RF score of 1.0 and the latter of 0.99. However CI will be very large in the former case, and small in the latter.

The 95% CI is a range of possible values that will contain the true RF value 95% of the times if the experiment (same epitopes & assays) is repeated an infinite number of times under the same conditions, assuming that the statistical distribution of outcomes of either a B- or T-cell assay can be approximated by the binomial distribution as, in general, the following conditions of experiments are satisfied: (1) an assay is repeated for a fixed number N of tested subjects; (2) each tested subject can show only two possible outcomes (response or no response); (3) the probability of response, or no response, is the same for each subject; and (4) the subjects are statistically independent (see http://en.wikipedia.org/wiki/Binomial_distribution).

CI (light pink area on the upper plot) ranges from 0 to 1.0.

Given the observed RF score, which is binomially distributed, it is possible to estimate the 95% CI for the true RF. There are several ways to compute a confidence interval for a binomial proportion (see http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval). One of the widely used formulas is the Wilson score interval:

$$\frac{1}{1+\frac{1}{n}z^2}\left[\hat{p}+\frac{1}{2n}z^2 \pm z\sqrt{\frac{1}{n}\hat{p}\left(1-\hat{p}\right)+\frac{1}{4n^2}z^2}\right]$$

where n is the number of tested subjects (N); p is the number of responded subjects divided by the number of tested (R/N); and z is the *(1-½ α)* percentile of a standard normal distribution; for a 95% confidence interval (an error level **α** = 0.05), z = 1.96. In this formula, the upper bound for the CI is calculated using the sign '+', and the lower bound, using the sign '-'. The Wilson score interval is asymmetric unless p=0.5.

It was shown that for small sample size n (or N, the number of tested subjects) the Wilson interval underestimates the error (http://www.ucl.ac.uk/english-

) and a continuity-corrected version of the Wilson interval is suggested. Therefore, **the Wilson score interval was used for N>=50**, and the lower bound of 95% CI was calculated as:

$(((R/N) + 1.96*1.96/(2*N) - 1.96 * sqrt(((R/N)*(1-(R/N))+1.96*1.96/(4*N))/N))/(1+1.96*1.96/N));$

while, the upper bound as:

$(((R/N) + 1.96*1.96/(2*N) + 1.96 * sqrt(((R/N)*(1-(R/N))+1.96*1.96/(4*N))/N))/(1+1.96*1.96/N))$

For N<50, we calculate the approximations for CI directly based on the binomial cumulative distribution function.