

The Immune Epitope Database Analysis Resource

**Epitope Analysis Tools:
Adding value to epitope datasets**

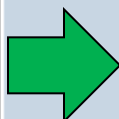
**Bjoern Peters
IEDB User Workshop
October 29, 2015**

Epitope Analysis Tools: Add value to epitope datasets

Conservation of swine flu (S-OIV)
epitopes as an example

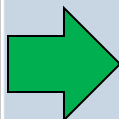
Analysis Tools

The tools below are intended for the detailed analysis of a known epitope sequence or group of sequences.



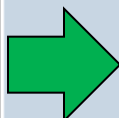
[Population Coverage](#)

This tool calculates the fraction of individuals predicted to respond to a given set of epitopes with calculation is made on the basis of HLA genotypic frequencies assuming non-linkage disequilibrium.



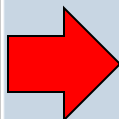
[Epitope Conservancy Analysis](#)

This tool calculates the degree of conservancy of an epitope within a given protein sequence set identity. The degree of conservation is defined as the fraction of protein sequences containing the epitope.



[Epitope Cluster Analysis](#)

This tool groups epitopes into clusters based on sequence identity. A cluster is defined as a group of epitopes with sequence similarity greater than the minimum sequence identity threshold specified.



[Homology Mapping](#)

This tool maps linear and conformational epitopes to 3D structures of proteins. This is done by comparing the protein sequence with that of proteins with known 3D structures in the PDB. The tool generates a list of the epitope source sequence and a homologous sequence from the PDB, and allows visualization in EpitopeViewer.

Population Coverage

- Calculates the fraction of individuals predicted to respond to a given set of epitopes with known MHC restrictions
- Based on HLA genotypic frequencies assuming non-linkage disequilibrium between HLA loci
- HLA genotypic frequencies have been updated with more comprehensive dataset (<http://allelefrequencies.net> rather than *dbMHC*)

IEDB Analysis Resource

Population Coverage Calculation

Population / Area Information

Example Data Sets

Help

Reference

Download

Contact

Population Coverage Calculation

Number of epitope(s):

Set

Query by:

Area, Country and Ethnicity *
▼ *

Select area(s) and/or population(s):

World

- East Asia

- Japan

- Oriental

- Korea; South

- Oriental

- Mongolia

- Oriental

Select calculation option(s):

Class I separate

Class II separate

Class I and II combined

Add user population(s) ?

Compute

Reset

* Population datasets generously provided by Derek Middleton at [The Allele Frequency Net Database](#)

Click here to load epitope / MHC restriction data from file ?

or enter epitope / MHC restriction data in the form below:

No.	Epitope	MHC Restricted Allele(s)
1	<input type="text"/>	<input type="text"/> <input type="button" value="Browse..."/>
2	<input type="text"/>	<input type="text"/> <input type="button" value="Browse..."/>
3	<input type="text"/>	<input type="text"/> <input type="button" value="Browse..."/>

Query: For a set of 11 MHC class II restricted epitopes with promiscuous HLA binding, what is the population coverage in different North African populations?

Population Coverage Calculation

Number of epitope(s):

11

Set

Query by:

Area, Country and Ethnicity ▾ *

Select area(s) and/or population(s):	Select calculation option(s):
<ul style="list-style-type: none"> - North Africa - Algeria - Arab - Ethiopia - Black - Mali - Black - Morocco - Arab 	<input type="checkbox"/> Class I separate <input checked="" type="checkbox"/> Class II separate <input type="checkbox"/> Class I and II combined
<input type="button" value="Add user population(s)"/> ?	<input type="button" value="Compute"/> <input type="button" value="Reset"/>

* Population datasets generously provided by Derek Middleton at [The Allele Frequency Net Database](#)

 ?

or enter epitope / MHC restriction data in the form below:

No.	Epitope	MHC Restricted Allele(s)
1	Gag 171	HLA-DRB1*01:01, HLA-DRB1*15:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*1 <input style="float: right;" type="button" value="Browse..."/>
2	Gag 294	HLA-DRB1*01:01, HLA-DRB1*15:01, HLA-DRB1*04:05, HLA-DRB1*11:01, HLA-DRB1*1 <input style="float: right;" type="button" value="Browse..."/>
3	Gag 298	HLA-DRB1*01:01, HLA-DRB1*15:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*0 <input style="float: right;" type="button" value="Browse..."/>
4	Pol 303	HLA-DRB1*01:01, HLA-DRB1*15:01, HLA-DRB1*03:01, HLA-DRB1*04:05, HLA-DRB1*0 <input style="float: right;" type="button" value="Browse..."/>
5	Pol 335	HLA-DRB1*01:01, HLA-DRB1*15:01, HLA-DRB1*04:05, HLA-DRB1*13:02, HLA-DRB1*0 <input style="float: right;" type="button" value="Browse..."/>
6	Pol 596	HLA-DRB1*01:01, HLA-DRB1*15:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*1 <input style="float: right;" type="button" value="Browse..."/>

Population Coverage Calculation Result

Population / Area	Class II		
	Coverage ^a	Average hit ^b	PC90 ^c
Algeria	77.15%	7.25	0.88
Ethiopia	83.00%	8.71	1.18
Mali	0.00%	0.00	?
Morocco	83.44%	8.14	1.21
Sudan	60.56%	4.52	0.51
Tunisia	74.26%	6.82	0.78
Average (Standard deviation)	63.07% (29.21%)	5.91 (2.95)	? (?)

Summary of Results
in Table; breakdown
in individual graphs

^a projected population coverage

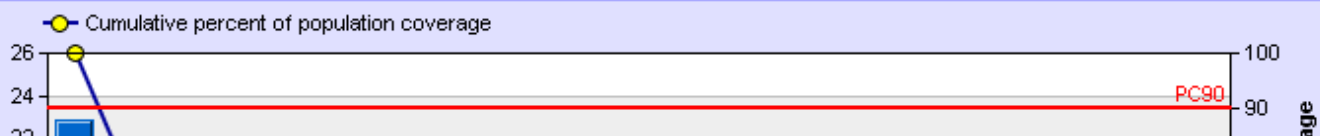
^b average number of epitope hits / HLA combinations recognized by the population

^c minimum number of epitope hits / HLA combinations recognized by 90% of the population

Algeria

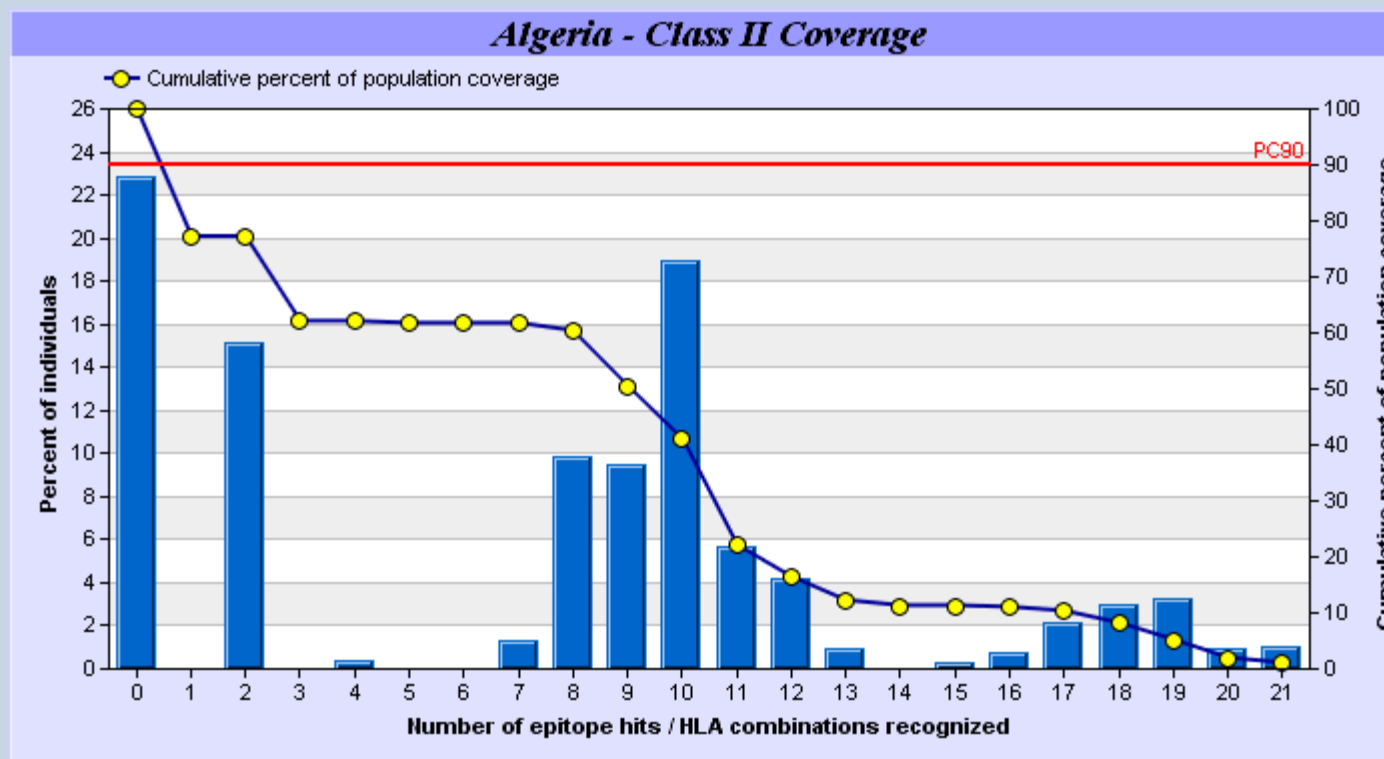
MHC Class	Coverage ^a	Average hit ^b	PC90 ^c
Class II	77.15%	7.25	0.88

Algeria - Class II Coverage



Algeria

MHC Class	Coverage ^a	Average hit ^b	PC90 ^c
Class II	77.15%	7.25	0.88



[View chart data in table format](#)

- ^a projected population coverage
- ^b average number of epitope hits / HLA combinations recognized by the population
- ^c minimum number of epitope hits / HLA combinations recognized by 90% of the population

[View coverage of individual epitope in Algeria](#)

[Back to top](#)

Epitope Conservancy Analysis

- Calculates the degree of conservancy of an epitope within a given protein sequence set at different degrees of sequence identity
- Degree of conservation is defined as the fraction of protein sequences containing the epitope at a given identity level.
- Example: Swine flu project (done in introduction)

Epitope Conservancy Analysis

Step 1. Epitope Sequence(s)

Enter epitope sequence(s) in PLAIN or FASTA format: ?

```
EVETPIRN
GLFGAIAGF
GLFGAIAGFIE
GTLVKTITDDQIEV
HHPSTNQEQTSLYVQAS
MSLLTEVETLTRNGWECKCRDSSD
MSLLTEVETLTRNGWGCRCSDSSD
MSLLTEVETPIRNEWGCRCNDSSD
NVPEKQTRGIFGAIAGFIE
```

Or upload epitope sequence(s) from a file: ?

Click here to upload

Browse...

Step 2. Protein Sequence(s)

Enter protein sequence(s) in PLAIN or FASTA format: ?

```
>gi|229609555|gb|ACQ83397| /Human/HA/H1N1/Sweden/2009/04/30/
hemagglutinin[Influenza A virus (A/Stockholm/28/2009(H1N1))]
MKAILVLLYTFATANADTLICIGYHANNSTDTVDTVLEKNVTVTHSVNLLLEDKHNGKLCCLRGVAPLHLG
KCNIAGWILGNPECESLSTASSWSYIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKTSSW
PNHDXNKGVTAACPHAGAKSFYKNLIWLVKKGNYSYKLSKSYINDKGKEVLVLWGIHHPSTSADQQSLYQ
NADAYVFGSSRYSKFKPEIAIRPKVRDQEGRMNYYWTLVEPGDKITFEATGNLVVPRYAFAMERNAGS
GIIISDTPVHDCNTTTCQTPKGAINSLPFQNIHPITIGKCPKYVKSTKLRLATGLRNVPSIQSRGLFGAI
AGFIEGGWTGMVDGWYGYHHQNEQGSYAA
```

Or upload protein sequence(s) from a file: ?

Click here to upload

Browse...

[Click here](#) to browse for sequences in NCBI ?

Step 3. Calculation option(s)

Epitope No. ▲ ▼	Epitope name	Epitope sequence	Epitope length	Percent of protein sequence matches at identity $\geq 100\%$ ▲ ▼	Minimum identity	Maximum identity ▲ ▼	View details
1	1	EVETPIRN	8	0.00% (0/439)	12.50%	75.00%	Go
2	2	GLFGAIAGF	9	12.53% (55/439)	22.22%	100.00%	Go
3	3	GLFGAIAGFIE	11	12.53% (55/439)	18.18%	100.00%	Go
4	4	GTLVKTITDDQIEV	14	0.00% (0/439)	21.43%	35.71%	Go
5	5	HHPSTNQEQTSLYVQAS	17	0.00% (0/439)	17.65%	58.82%	Go
6	6	MSLLTEVETLTRNGWECKCRDSSD	24	0.00% (0/439)	12.50%	79.17%	Go
7	7	MSLLTEVETLTRNGWGCRCDSSD	24	0.00% (0/439)	12.50%	83.33%	Go
8	8	MSLLTEVETPIRNEWGCRCDSSD	24	0.00% (0/439)	12.50%	83.33%	Go
9	9	NVPEKQTRGIFGAIAGFIE	19	0.00% (0/439)	21.05%	78.95%	Go
10	10	SKAFSNCYPYDVPDYASL	18	0.00% (0/439)	16.67%	44.44%	Go
11	11	SLLTEVETPIRNEWGCRCDSS	22	0.00% (0/439)	13.64%	81.82%	Go
12	12	SLLTEVETPIRNEWGCRCDSSD	23	0.00% (0/439)	13.04%	82.61%	Go
13	13	WTGVTQN	7	0.00% (0/439)	0.00%	57.14%	Go
14	14	LTEVETPIRN	10	0.00% (0/439)	20.00%	80.00%	Go
15	15	MSLLTEVETPTKNEWECRCDSSD	24	0.00% (0/439)	12.50%	87.50%	Go
16	16	SLLTEVET	8	19.59% (86/439)	25.00%	100.00%	Go
17	17	VETPIRN	7	0.00% (0/439)	14.29%	71.43%	Go



Conservancy analysis for epitope #2

Epitope name	Epitope sequence	Epitope length	Percent of protein sequence matches at identity \geq 100%
2	GLFGAIAGF	9	12.53% (55/439)

Show records with identity

\geq

100%

Show records

Download data to file

Protein No. ▲ ▼	Protein name	Positions	Protein sub-sequence(s)	Identity ▼
1	gi 228860929 gb ACQ45338 /Human/HA/H1N1/Netherlands/2009/04/29/hemagglutinin[Influenza A virus (A/Netherlands/602/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
3	gi 227809925 gb ACP41145 /Human/HA/H1N1/China/2009/03/11/hemagglutinin[Influenza A virus (A/Yuelu/314/2009(H1N1))]	324-332	GLFGAIAGF	100.00%
7	gi 227809830 gb ACP41105 /Human/HA/H1N1/USA/2009/04/01/hemagglutinin[Influenza A virus (A/California/04/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
14	gi 227831759 gb ACP41926 /Human/HA/H1N1/USA/2009/03/30/hemagglutinin[Influenza A virus (A/California/05/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
22	gi 227831774 gb ACP41934 /Human/HA/H1N1/USA/2009/04/15/hemagglutinin[Influenza A virus (A/Texas/05/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
23	gi 227831776 gb ACP41935 /Human/HA/H1N1/USA/2009/04/16/hemagglutinin[Influenza A virus (A/California/06/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
41	gi 227831808 gb ACP41953 /Human/HA/H1N1/USA/2009/04/09/hemagglutinin[Influenza A virus (A/California/07/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
51	gi 227831826 gb ACP41963 /Human/HA/H1N1/USA/2009/04/14/hemagglutinin[Influenza A virus (A/Texas/04/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
54	gi 227977099 gb ACP44147 /Human/HA/H1N1/USA/2009/04/25/hemagglutinin[Influenza A virus (A/New York/19/2009(H1N1))]	345-353	GLFGAIAGF	100.00%
57	gi 227977104 gb ACP44150 /Human/HA/H1N1/USA/2009/04/08/	345-353	GLFGAIAGF	100.00%

Epitope Cluster Analysis

- Groups epitopes into clusters based on sequence identity.
- Clusters are defined as a group of sequences which have a sequence similarity greater than the minimum sequence identity threshold specified.
- Slightly different implementation than the clustering tool integrated on the IEDB site (which relies on pre-computing)

Epitope Cluster Analysis

Epitope sequences

Enter epitope sequences in PLAIN or FASTA format: [?](#)

Or upload epitope sequence(s) from a file: [?](#)

No file chosen

Sequence identity threshold

Select minimum sequence identity threshold: [?](#)

80%

Epitope Cluster Analysis

Epitope sequences

Enter epitope sequences in PLAIN or FASTA format: [?](#)

```
AIYHTENAYVSVVSSHYNR  
AMEQMAGSSEQAAEAMEVASQARQMVQAMRTIGTHPSSS  
CKRGPDSGFFSRLNWLY  
CKRGPDSGFFSRLNWLYKSGSTYPVQNVTMPNNDNS  
CLGHHAVPNGTLVKITITNDQIEVTNATELVQSSSTGKIC  
CNNPHRIL  
CNNPHRILDGINC  
CNNPHRILDGINCTLIDALLGDPHCDGFQNEKWDL
```

Or upload epitope sequence(s) from a file: [?](#)

No file chosen

Sequence identity threshold

Select minimum sequence identity threshold: [?](#)

80%

Epitope Cluster Analysis Results

Number of epitopes: 75

Number of clusters: 30

Identity threshold: 80%

[Download data to file](#)

Cluster No.	Number of epitopes in the cluster	Epitope No.	Epitope Name	Epitope Sequence
1	1	1	1	AIYHTENAYVSVVSSHYNR
2	1	1	2	AMEQMAGSSEQAAEAMEVASQARQMVQAMRTIGTHPSSS
3	2	1	3	CKRGPDSGFFSRLNWLY
		2	4	CKRGPDSGFFSRLNWLYKSGSTYPVQNVTMPNNDNS
4	1	1	5	CLGHHAVPNGTLVKITITNDQIEVTNATELVQSSSTGKIC
5	20	1	6	CNNPHRIL
		2	7	CNNPHRILDGINC
		3	8	CNNPHRILDGINCTLIDALLGDPHCDGFQNEKWDL
		4	22	GKICNNPHRILDGIDCTLID
		5	40	NATELVQSSSTGKICNNPHRILDGINC
		6	66	TNATELVQSSSTGKICNNPHRILDGIN
		7	20	FQNEKWDL
		8	28	HCDGFQNEKWDL
		9	29	HCDGFQNEKWDLFVE
		10	30	HCDGFQNEKWDLFVERSKAFSNCYPYDVPDYASLRS

Questions?