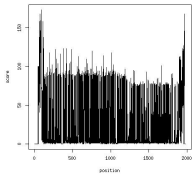




Sequence Conservation/Variation Analysis



- Analyze sequence polymorphism at the nucleotide or amino acid level.
- Calculate consensus sequence and polymorphism of ViPR sequences or your own

Sequence Variation Analysis Sample Report

Save SNP result to Workbench for future retrieval

Download consensus sequence in FASTA format

Download raw alignment of all sequences

Score ranges from 0 (no polymorphism) to 232 (highest polymorphism).

At each position, the consensus is the allele with frequency greater than 50%. If no allele exceeds 50%, N (for nucleotide) or Xaa (for amino acid) is used to indicate ambiguity.

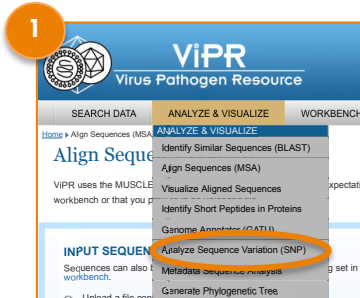
Sequence Variation Analysis Result									
Position	Score (SOP)	Consensus	A	T	G	C	Deletion	# Sequences	
1	0	G	0	0	1	0	0	1	1
2	0	G	0	0	1	0	0	1	1
3	0	A	1	0	0	0	0	1	1
4	0	C	0	0	0	1	0	1	1
5	0	C	0	0	0	1	0	1	1
6	0	G	0	0	1	0	0	1	1
7	0	A	27	0	0	0	0	27	1
8	0	C	0	0	0	27	0	27	1
9	0	A	27	0	0	0	0	27	1
9905	0	G	0	0	32	0	0	32	1
9906	100	N	0	16	0	16	0	32	1
9907	0	C	0	0	0	32	0	32	1
9908	20	A	31	0	1	0	0	32	1
9909	0	C	0	0	0	32	0	32	1
9910	0	A	32	0	0	0	0	32	1
9911	0	T	0	32	0	0	0	32	1

The analysis report page shows the polymorphism score, consensus, and counts for each different base/amino acid at each position.

Consensus sequence and raw alignment are available for download.

Count for different nucleotides at each position

Option 1: Calculate consensus sequence and sequence variation of your own sequences



On the ViPR homepage, choose a virus family or a Featured Virus to start.

- Mouse-over the "Analyze & Visualize" tab and click "Analyze Sequence Variation (SNP)".
- On the SNP landing page, use one of the three options to input sequences:
 - Upload a sequence file in FASTA format OR
 - Paste sequences in FASTA format OR
 - Use a working set from your Workbench.
- Then click "Run" to run the analysis.
- As soon as the analysis is finished, a report similar to the above sample report will be displayed on the screen.

1

2

SEARCH DATA ANALYZE & VISUALIZE WORKBENCH SUBMIT DATA VIRUS FAMILIES HOME Dengue

Analyze Sequence Variation

Analyze Sequence Variation (SNP)

For polymorphism calculation MUSCLE is used for multiple sequence alignment. A consensus sequence is created by "majority rule". At each position, the consensus is the allele with frequency greater than 50%, regardless of coverage. If no allele exceeds 50%, N (for nucleotide) or Xaa (for amino acid) indicates ambiguity. Sequences in the alignment are then compared to the consensus to identify polymorphisms.

To score polymorphism at each position, a formula modified from the one cited in Crooks et al. is used.

$S = -100 \cdot \sum (P_i \cdot \log P_i)$ where P_i is the frequency of the i th allele.

The score is the normalized entropy of the observed allele distribution. For nucleotides, scores can range from 0 (no polymorphism) to 232 (highest polymorphism).

INPUT SEQUENCES

Sequences can also be selected from search results or a working set in your workbench.

Upload a file containing my sequences in FASTA format.

File Path:

The minimum number of sequences is 2.

☐ Paste sequences in FASTA format.

☐ Use working sets

2.1

2.2

Paste sequence in FASTA or Phylip format. Define in your FASTA file will be used to label the display

```
>gb:FJ850072|Organism:Dengue virus
DENV-2/BR/BID-V2376
/2000|Subtype:2|Host:Human
ACAAGACAGATTCTTGAGGGAGCTAAGCTCAAG
TAGTTCTACAGCTTTTCTGATGAGAGCAGATC
TCTGATGAATAACCAAGCAAAAAGGCGAGAGTAC
GCCTTCAATATGCTGAAACGCGAGAGAACCCG
GTGTCACTGTCAACAGCTGACAAAGAGATTCTCA
```

2.3

Choose Working Set

Name	Type	Number of Sequences	Date
<input type="radio"/> Dengue2_genome_human-1999-2000	Genome	32	08/05/2011 3:37 PM
<input type="radio"/> DENV1-4_99-00_human_Genomes	Genome	82	06/24/2011 10:43 AM
<input type="radio"/> hepatitis c	Genome	1	03/29/2011 11:10 AM



Option 2. Calculate consensus sequence and sequence variation of ViPR sequences

1 Click **i** to view details of the record

2 Select sequences and add them to a working set for future analysis. You'll need to register for a Workbench account to use this feature.

3 Select display fields
Custom-sort records

4 Select Sequence Type

5 Analyze Sequence Variation (SNP)

6 Processing...

On the ViPR homepage, choose a virus family or a Featured Virus to start.

1. Search for nucleotide or protein sequences in ViPR by using the "Genomes" or "Genes & Proteins" search option available from the "Search Data" tab. For this example, we will use genome sequences.
2. Select search criteria on the Genome Search page and click the "Search" button to run your query.
3. On the search results page, select the desired sequences by clicking the checkboxes, mouse-over the yellow "Run Analysis" button and click "Analyze Sequence Variation (SNP)". If you want to include sequences that are not in this search result or to use the sequences to do further analysis, select the desired sequences and click "Add to Working Set". Then add other sequences to the same working set later by repeating the process. Click the "Workbench" tab and find the working set you saved. Click **i** next to it to view the details of the working set. Then mouse-over the yellow "Run Analysis" button and click "Analyze Sequence Variation (SNP)".
4. A "Select Sequence Type" lightbox will pop up. Select the appropriate sequence type and click "Continue".
5. On the next page, you will see a brief description of the SNP tool. Click "Run" to proceed.
6. If you have a large number of long sequences to analyze, it may take a few minutes to run. While the analysis is running, you can choose to save it (upon completion) to your Workbench by entering a name for the analysis and then clicking the "Save to Workbench" button. Then you can move to other parts of the ViPR site, and retrieve the SNP analysis result later from your Workbench.
7. As soon as the analysis is finished, a report similar to the sample report on the reverse page will be displayed on the screen.