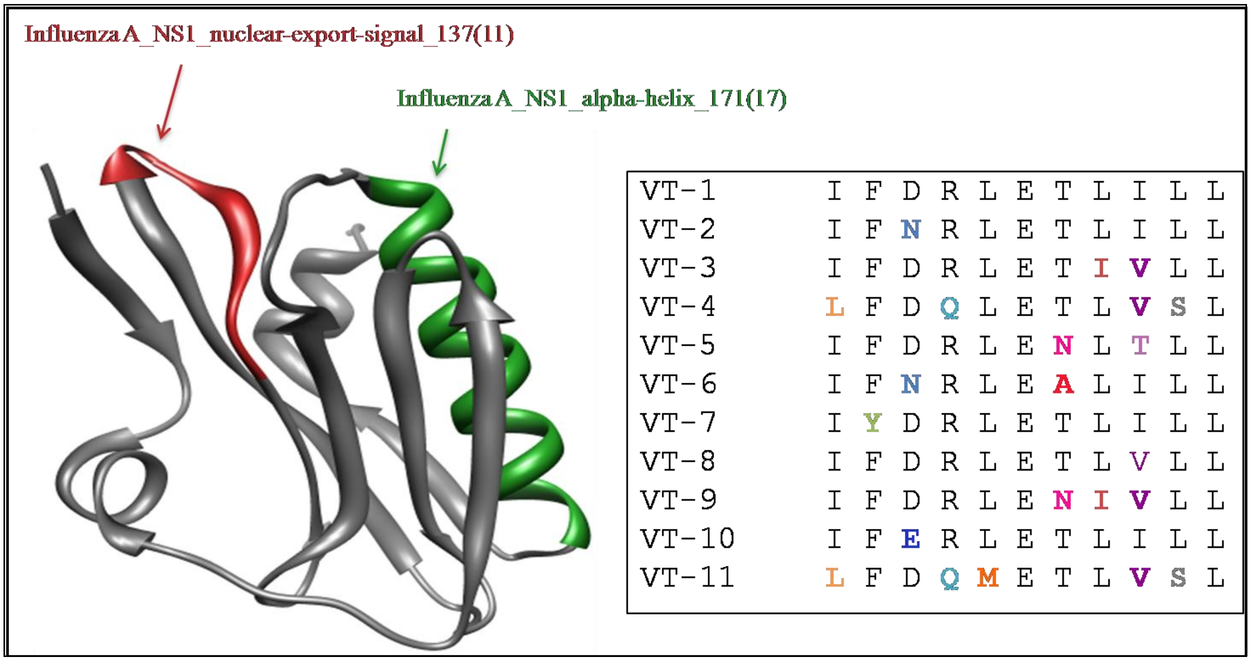
SEQUENCE FEATURE VARIANT TYPES

DEFINITION OF SFVT:

The “Sequence Feature Variant Type” (SFVT) component in IRD (http://www.fludb.org) is a relatively novel approach that delineates specific regions, called “Sequence Features” (SF), within influenza virus segments/proteins (or DNA or RNA molecules) based on their functional properties (e.g. protein-protein interaction site, enzyme active site, etc.), structural properties (e.g. beta-strand), sequence alteration effect (point mutations that can lead to a phenotypic change) or immune epitope locations. The extent of sequence variation for each SF is described as a collection of “Variant Types” (VT), computed by multiple sequence alignments of all relevant influenza virus genomes available in IRD. Therefore a set of unique sequence substitution(s) existing within a characterized region of the genome is termed a ‘*Sequence Feature Variant Type*’. (Example ***Fig 1***).

SFVT Citation:

Noronha JM, Liu M, Squires RB, Pickett BE, Hale BG, Air GM, Galloway SE, Takimoto T, Schmolke M, Hunt V, Klem E, García-Sastre A, McGee M, Scheuermann RH. Mar 2012. *Influenza Sequence Feature Variant Type (Flu-SFVT) analysis: evidence for a role of NS1 in influenza host range restriction.* J Virol. doi: 10.1128/JVI.06901-11
PMID: 22398283


Influenza A_NS1_nuclear-export-signal_137(11)
Influenza A_NS1_alpha-helix_171(17)

VT-1	I	F	D	R	L	E	T	L	I	L	L
VT-2	I	F	N	R	L	E	T	L	I	L	L
VT-3	I	F	D	R	L	E	T	I	V	L	L
VT-4	L	F	D	Q	L	E	T	L	V	S	L
VT-5	I	F	D	R	L	E	N	L	T	L	L
VT-6	I	F	N	R	L	E	A	L	I	L	L
VT-7	I	Y	D	R	L	E	T	L	I	L	L
VT-8	I	F	D	R	L	E	T	L	V	L	L
VT-9	I	F	D	R	L	E	N	I	V	L	L
VT-10	I	F	E	R	L	E	T	L	I	L	L
VT-11	L	F	D	Q	M	E	T	L	V	S	L

Note: This figure is created by IRD team and permission is required to reuse it in any form.

Fig 1: The crystal structure of Influenza A NS1 protein (PDB 2GX9) with the Nuclear Export Signal (NES) "Sequence Feature" (SF) highlighted in red and an alpha helix SF highlighted in green. A protein alignment with colors representing the variation that exists in the NES region is also shown. Each sequence that has one or more substitutions comprises a unique fingerprint or "Variant Type" (VT). Hence a set of unique sequence substitutions existing within any characterized region is termed a "sequence feature variant type" (SFVT). SFVTs can be used in statistical analyses to determine the relationship between sequence variations and metadata associated with these sequences.

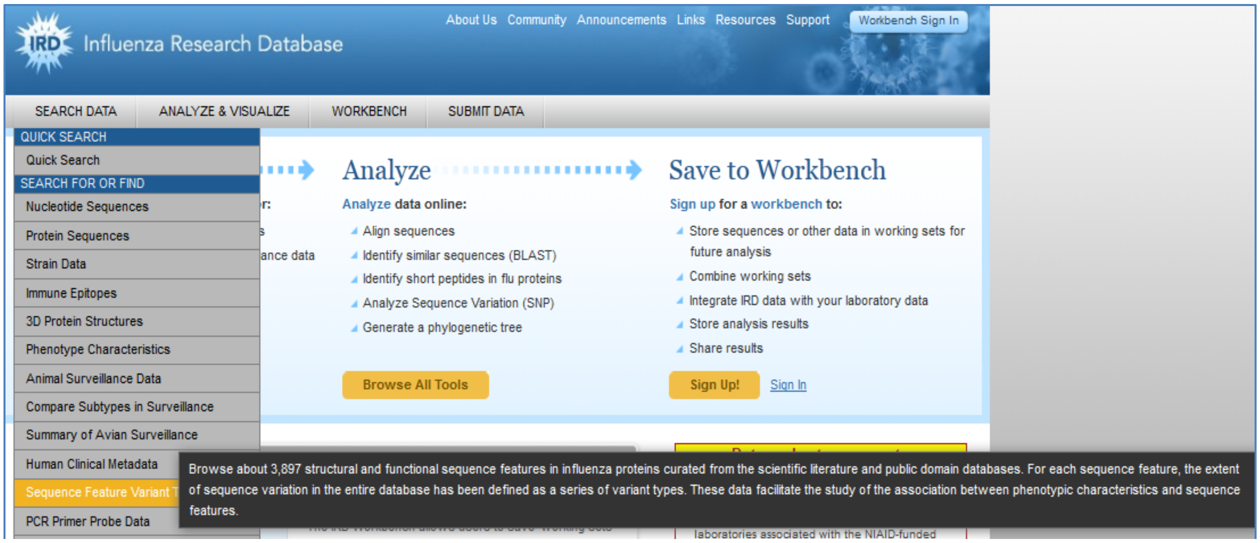
The SFVT approach is adapted from the MHC- SFVT approach developed by the DAIT- Data Interoperability Science Committee/HLA working group members for human MHC region sequence data (see Karp, D. et al. 2009 Human Molecular Genetics, PMID: 19933168).

Influenza virus segments are known to encode proteins with frequent variations in amino acid sequences between independent isolates. So defining SFVTs will aid in identifying sequence polymorphisms that correlate with important phenotypic characteristics, such as drug sensitivity/resistance, measures of virulence, host susceptibility, virus transmissibility, fatality in animal models, etc.

DETAILED DESCRIPTION:

How to access the SFVT component in IRD:

- SFVTs can be accessed from the IRD Homepage via the 'Search Data' tab (***Fig 2***).



Analyze data online:

- Align sequences
- Identify similar sequences (BLAST)
- Identify short peptides in flu proteins
- Analyze Sequence Variation (SNP)
- Generate a phylogenetic tree

Sign up for a workbench to:

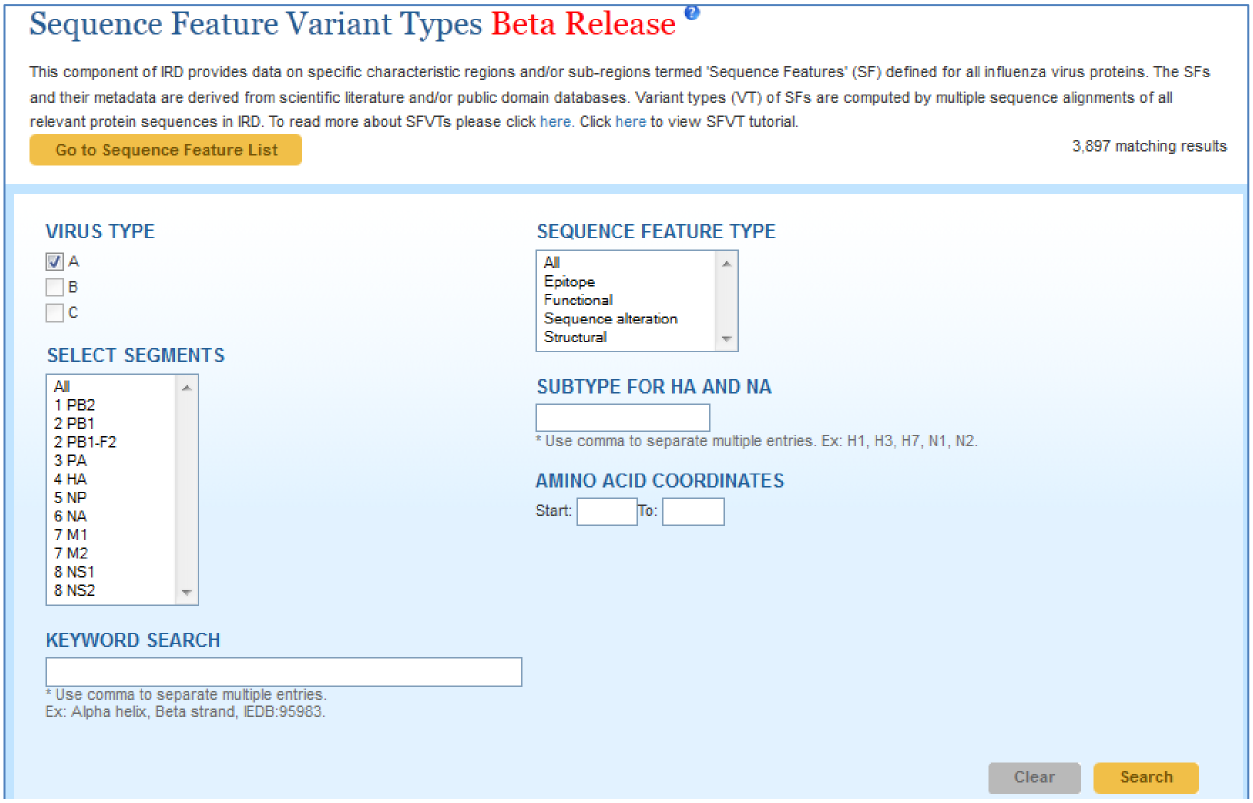
- Store sequences or other data in working sets for future analysis
- Combine working sets
- Integrate IRD data with your laboratory data
- Store analysis results
- Share results

Browse about 3,897 structural and functional sequence features in influenza proteins curated from the scientific literature and public domain databases. For each sequence feature, the extent of sequence variation in the entire database has been defined as a series of variant types. These data facilitate the study of the association between phenotypic characteristics and sequence features.

Fig 2: Screenshot of the IRD Homepage showing the 'Search Data' tab with SFVT search option highlighted.

- The entire list of SFVTs can be browsed using the ‘Go to Sequence Feature List’ tab or,

- Specific SFs can be searched using the ‘SFVT search interface’ (***Fig 3***). Searches can be based on criteria such as influenza virus type, segment and subtype, SF type, amino acid co-ordinates and keywords.



Sequence Feature Variant Types Beta Release

This component of IRD provides data on specific characteristic regions and/or sub-regions termed 'Sequence Features' (SF) defined for all influenza virus proteins. The SFs and their metadata are derived from scientific literature and/or public domain databases. Variant types (VT) of SFs are computed by multiple sequence alignments of all relevant protein sequences in IRD. To read more about SFVTs please click here. Click here to view SFVT tutorial.

Go to Sequence Feature List

3,897 matching results

VIRUS TYPE

☑ A
□ B
□ C

SELECT SEGMENTS

All
1 PB2
2 PB1
2 PB1-F2
3 PA
4 HA
5 NP
6 NA
7 M1
7 M2
8 NS1
8 NS2

KEYWORD SEARCH

* Use comma to separate multiple entries.
Ex: Alpha helix, Beta strand, IEDB:95983.

SEQUENCE FEATURE TYPE

All
Epitope
Functional
Sequence alteration
Structural

SUBTYPE FOR HA AND NA

* Use comma to separate multiple entries. Ex: H1, H3, H7, N1, N2.

AMINO ACID COORDINATES

Start: To:

Clear Search

<u>**Fig 3:** Screenshot displaying the SFVT browse and search interface in IRD.</u>

Naming guidelines for sequence features and variant types:

- The SF name is assigned using the following approach and each character in the name is separated by an underscore:

  Influenza virus type _protein symbol_ sequence feature type_ start position of the sequence feature (Total length/the total no. of amino acids that make up the SF)

  Example: Influenza A_H1_cytoplasmic-domain_550(16)

  In the above example, the virus in which the SF is described belongs to the Type A Influenza virus group and the SF is present in the H1 subtype of the Hemagglutinin (HA) protein; the SF type describes a cytoplasmic domain which starts at position 550 and the total length of this SF is 16 amino acids which is shown in parenthesis following the start position.

- There is no restriction on the size of the SF and they can be present on virtually any sub-region of the genomic sequence. SFs can be overlapping and continuous or discontinuous in the linear sequence.

  Example: *Influenza A_H1_sialic-acid-binding-site_98(17)* constitutes a string of discontinuous amino acids (residues: 98, 134-138, 154, 156, 184, 191, 195, 196, 225-229) that are involved in binding to sialic acid receptors on host cells, while the *Influenza A_NS1_nuclear-export-signal_137(11)* is an example for a continuous SF that is 11 amino acid long and ranges from position 137 to 147.

- SFs are curated from published literature (Pubmed, www.ncbi.nlm.nih.gov/pubmed) and from public domain databases such as UniProt www.uniprot.org) and Immune Epitope Database (IEDB, www.immuneepitope.org)

- In the case of surface protein such as hemagglutinin, SFs are currently defined for a few commonly studied subtypes namely, H1, H2, H3, H5 and H7. We plan to extend this to the remaining subtypes in subsequent versions. Similarly, in case of neuraminidase that has 9 known subtypes, SFs are currently defined for N1 and N2 subtypes.

- The current version of the SF list is validated for accuracy by domain experts in the field of Influenza research.

- For ease of data retrieval and filtering, SFs are classified into four broad categories: structural SFs, functional SFs, sequence alterations and immune epitopes based on the role they play in the genomic sequence. SFs can overlap between two or more categories.

- Evidence codes are assigned to each SF indicating whether it is curated from literature (EXP) or inferred from electronic annotations (IEA).

- All influenza virus sequences in IRD for a given segment are aligned to the reference strain chosen for that segment. Any sequence variation in the amino acid residue(s) that constitute the SF is defined as a variant type (VT) of the SF.

- So for the purposes of computing SFVTs, a standard reference strain (also called VT-1 strain) has been selected for each influenza virus segment based on criteria regarding various aspects of strain characterization including the presence of experimentally determined 3-D structures for that sequence and its importance in flu research. (*Table 1*)

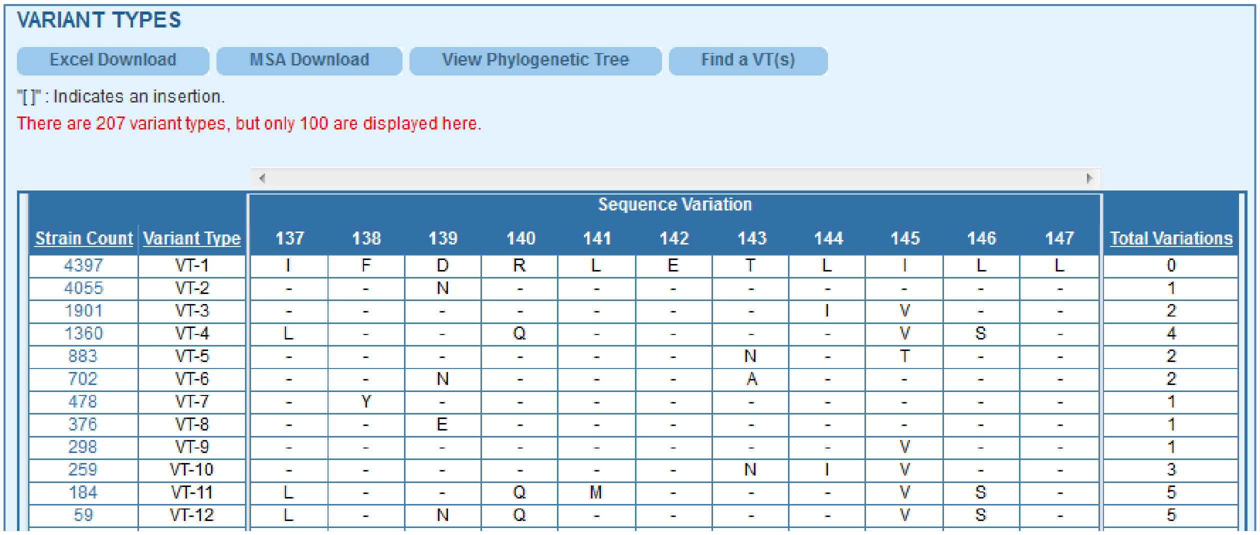
- VT-1 always represents the reference strain chosen for the protein and all other subsequent VTs are serially numbered from VT-2 onwards. VT-unknown indicates that the sequence is unknown or is truncated in that region of SF.

Table 1: List of reference strains for each Influenza virus protein

Segment	Protein	Serotype	Reference Strain	GenBank Accession No.
1	PB2	-	A/VietNam/1203/2004	EF467805
2	PB1	-	A/Hong Kong/156/1997	AF036362
2	PB1-F2		A/WSN/1933	CY034138
3	PA	-	A/WSN/1933	CY034137
4	HA	H1	A/California/04/2009	FJ966082
		H2	A/Japan/305/1957	J02127
		H3	A/Aichi/2/1968	AB284320
		H5	A/VietNam/1203/2004	AY818135
		H7	A/turkey/Italy/220158/2002	AY586409
5	NP	-	A/WSN/1933	CY034135
6	NA	N1	A/California/07/2009	FJ984386
		N2	A/Tokyo/3/1967	U38242
7	M	-	A/Udorn/1972	J02167
8	NS	-	A/Udorn/1972	V01102

- VTs from VT-2 onwards are numbered based on their frequency of occurrence in the IRD database as of March 2012. VTs can be sorted in ascending or descending order (Table 2). A dash indicates that the amino acid residue is identical to the residue at the corresponding position on the reference strain for that SF. Strain count links to the total number of Influenza strains in IRD that contain the specific VT

Table 2: List of VTs occurring in the NS1 nuclear export signal region..



VARIANT TYPES

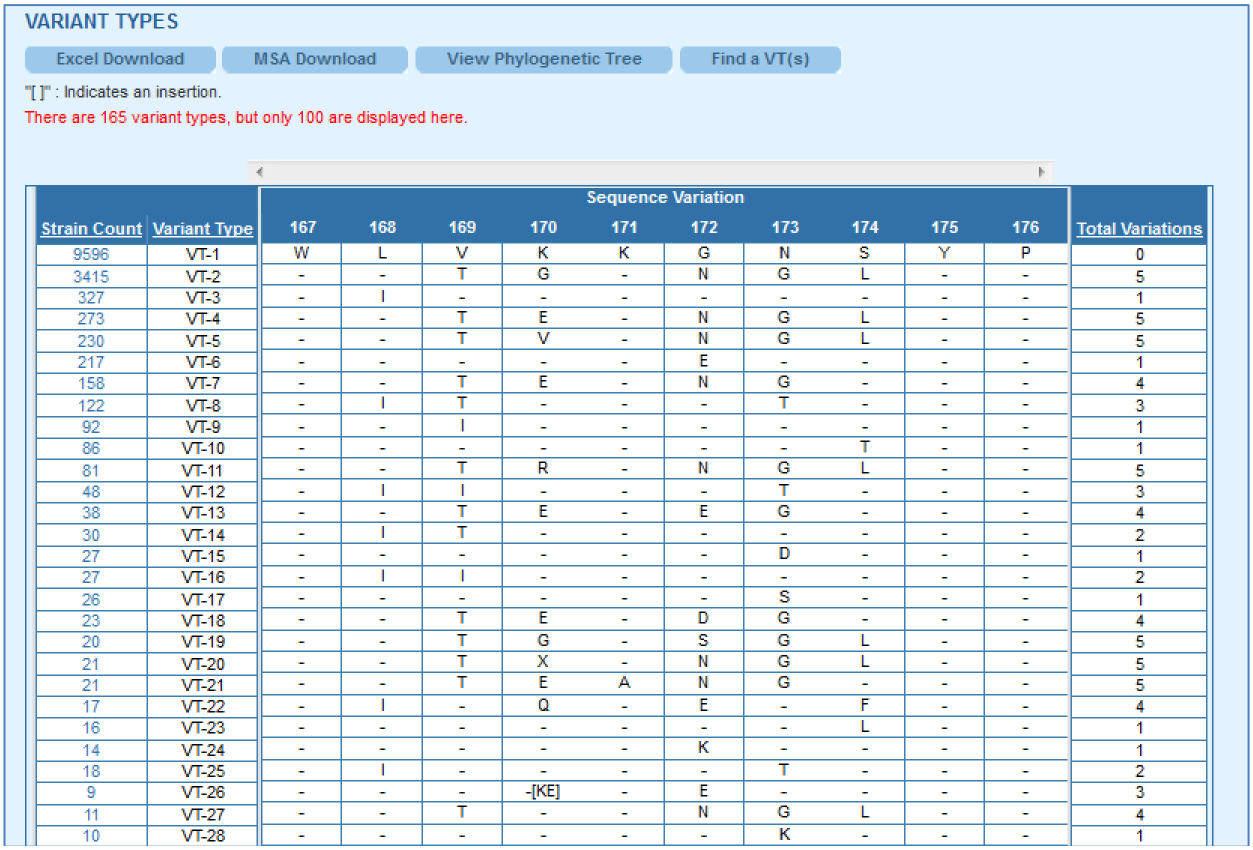
"[ ]" : Indicates an insertion.

There are 207 variant types, but only 100 are displayed here.

		Sequence Variation											
Strain Count	Variant Type	137	138	139	140	141	142	143	144	145	146	147	Total Variations
4397	VT-1	I	F	D	R	L	E	T	L	I	L	L	0
4055	VT-2	-	-	N	-	-	-	-	-	-	-	-	1
1901	VT-3	-	-	-	-	-	-	-	I	V	-	-	2
1360	VT-4	L	-	-	Q	-	-	-	-	V	S	-	4
883	VT-5	-	-	-	-	-	-	N	-	T	-	-	2
702	VT-6	-	-	N	-	-	-	A	-	-	-	-	2
478	VT-7	-	Y	-	-	-	-	-	-	-	-	-	1
376	VT-8	-	-	E	-	-	-	-	-	-	-	-	1
298	VT-9	-	-	-	-	-	-	-	-	V	-	-	1
259	VT-10	-	-	-	-	-	-	N	I	V	-	-	3
184	VT-11	L	-	-	Q	M	-	-	-	V	S	-	5
59	VT-12	L	-	N	Q	-	-	-	-	V	S	-	5

Display of insertions within the SF region:

- Any insertion(s) in the amino acid sequence of a VT with respect to the position on the reference strain is denoted in brackets [ ] with a dash prefixed to it as shown in the figure below (**Fig 4**)



"[ ]" : Indicates an insertion.

There are 165 variant types, but only 100 are displayed here.

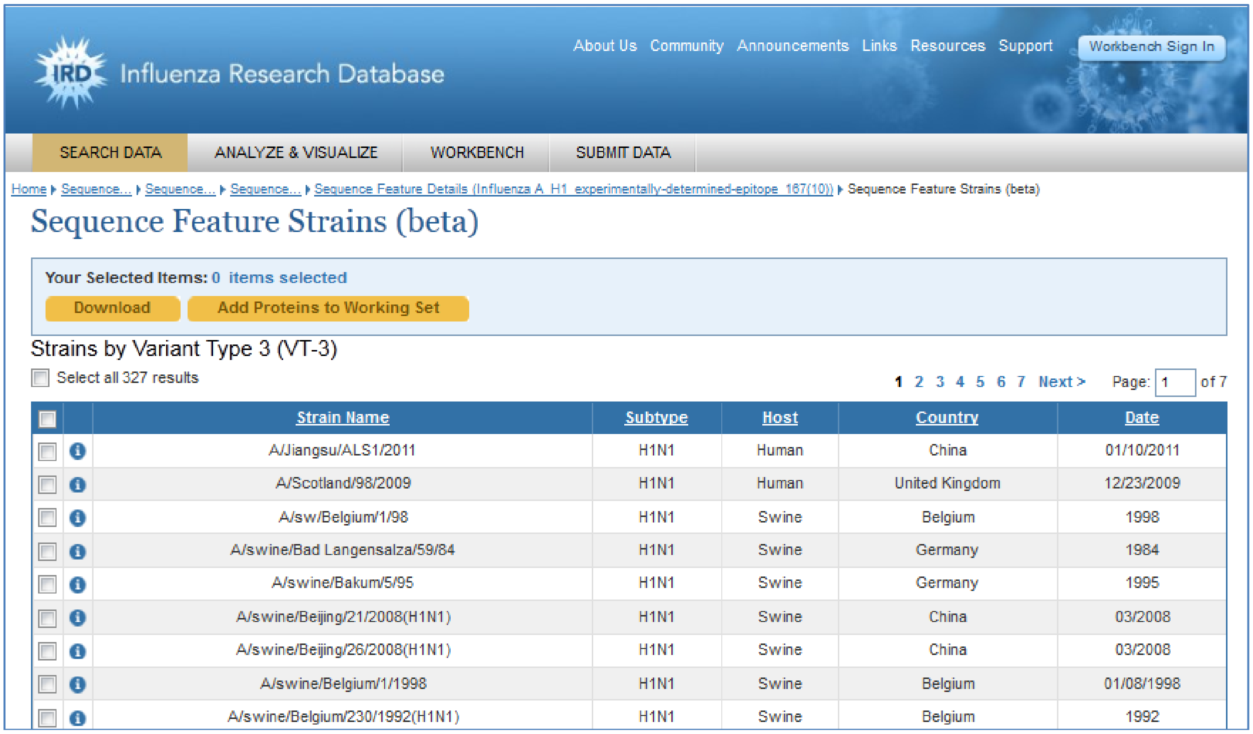
Strain Count	Variant Type	Sequence Variation										Total Variations
		167	168	169	170	171	172	173	174	175	176	
9596	VT-1	W	L	V	K	K	G	N	S	Y	P	0
3415	VT-2	-	-	T	G	-	N	G	L	-	-	5
327	VT-3	-	I	-	-	-	-	-	-	-	-	1
273	VT-4	-	-	T	E	-	N	G	L	-	-	5
230	VT-5	-	-	T	V	-	N	G	L	-	-	5
217	VT-6	-	-	-	-	-	E	-	-	-	-	1
158	VT-7	-	-	T	E	-	N	G	-	-	-	4
122	VT-8	-	I	T	-	-	-	T	-	-	-	3
92	VT-9	-	-	I	-	-	-	-	-	-	-	1
86	VT-10	-	-	-	-	-	-	-	T	-	-	1
81	VT-11	-	-	T	R	-	N	G	L	-	-	5
48	VT-12	-	I	I	-	-	-	T	-	-	-	3
38	VT-13	-	-	T	E	-	E	G	-	-	-	4
30	VT-14	-	I	T	-	-	-	-	-	-	-	2
27	VT-15	-	-	-	-	-	-	D	-	-	-	1
27	VT-16	-	I	I	-	-	-	-	-	-	-	2
26	VT-17	-	-	-	-	-	-	S	-	-	-	1
23	VT-18	-	-	T	E	-	D	G	-	-	-	4
20	VT-19	-	-	T	G	-	S	G	L	-	-	5
21	VT-20	-	-	T	X	-	N	G	L	-	-	5
21	VT-21	-	-	T	E	A	N	G	-	-	-	5
17	VT-22	-	I	-	Q	-	E	-	F	-	-	4
16	VT-23	-	-	-	-	-	-	-	L	-	-	1
14	VT-24	-	-	-	-	-	K	-	-	-	-	1
18	VT-25	-	I	-	-	-	-	T	-	-	-	2
9	VT-26	-	-	-	-[KE]	-	E	-	-	-	-	3
11	VT-27	-	-	T	-	-	N	G	L	-	-	4
10	VT-28	-	-	-	-	-	-	K	-	-	-	1

Fig 4: Screenshot of a VT table for the SF, *Influenza A_H1_experimentally-determined-epitope_167(10)* showing two amino acid insertions at position 170 in the VT-26.

- Strains that belong to each VT group have associated metadata in IRD such as the country of isolation, host, year and subtype of the virus. (**Fig 5**) All the metadata columns are sortable for easy browsing of the data.

- This strain metadata can be used in statistical and computational analysis to study interesting genotype-phenotype associations.

- Strain list can be downloaded in excel format or can be added to the personal workbench as a working set for further analysis.



IRD
Influenza Research Database
About Us
Community
Announcements
Links
Resources
Support
Workbench Sign In
SEARCH DATA
ANALYZE & VISUALIZE
WORKBENCH
SUBMIT DATA
Home ▸ Sequence... ▸ Sequence... ▸ Sequence... ▸ Sequence Feature Details (Influenza A_H1_experimentally-determined-epitope_167(10)) ▸ Sequence Feature Strains (beta)

Sequence Feature Strains (beta)

Your Selected Items: 0 items selected

Download
Add Proteins to Working Set

Strains by Variant Type 3 (VT-3)

□ Select all 327 results

1 2 3 4 5 6 7 Next >
Page: 1 of 7

□		Strain Name	Subtype	Host	Country	Date
□	ℹ	A/Jiangsu/ALS1/2011	H1N1	Human	China	01/10/2011
□	ℹ	A/Scotland/98/2009	H1N1	Human	United Kingdom	12/23/2009
□	ℹ	A/sw/Belgium/1/98	H1N1	Swine	Belgium	1998
□	ℹ	A/swine/Bad Langensalza/59/84	H1N1	Swine	Germany	1984
□	ℹ	A/swine/Bakum/5/95	H1N1	Swine	Germany	1995
□	ℹ	A/swine/Beijing/21/2008(H1N1)	H1N1	Swine	China	03/2008
□	ℹ	A/swine/Beijing/26/2008(H1N1)	H1N1	Swine	China	03/2008
□	ℹ	A/swine/Belgium/1/1998	H1N1	Swine	Belgium	01/08/1998
□	ℹ	A/swine/Belgium/230/1992(H1N1)	H1N1	Swine	Belgium	1992

Fig 5: Screenshot of the SF strain list page. Table displays the influenza virus strains containing the VT-3 sequence for an immune epitope SF named *Influenza A_H1_experimentally-determined-epitope_167(10)*

Finding specific VT(s)

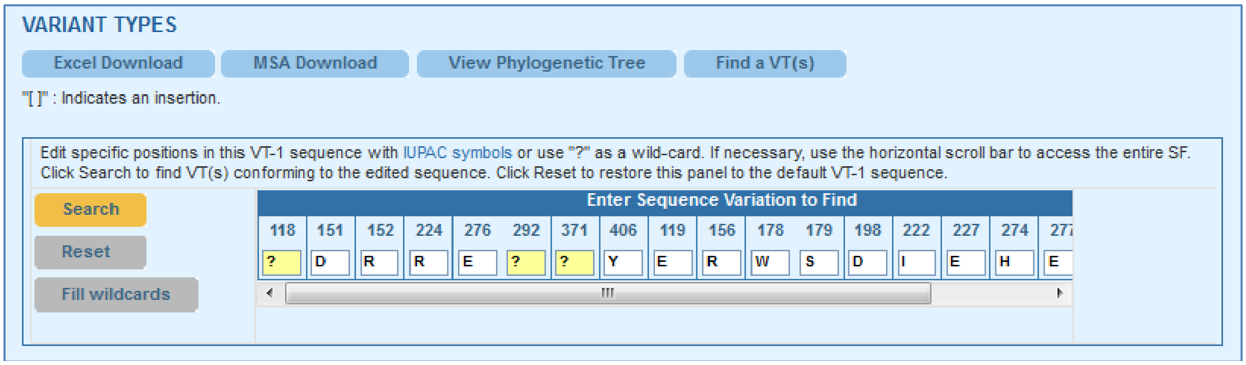
We have made it easy to find specific VT(s) and the strains that contain them in IRD. (***Fig 6***). In an example workflow, PDB structure (e.g. 1IVC) indicates specific amino acid residues (118, 292, and 371) that make contact with an inhibitor. A user might want to find strains in which those residues are altered.

- On the .PDB file find SFs that contain the residues contacting the inhibitor. In the file cited above, SFs: Influenza A_N2_SF60 and Influenza A_N2_SF61 contain all three candidate amino acids.

- From the SF Landing Page, go to the Sequence Feature list and click the link to the SFs for protein N2. Find and click on the 'Details' link of either SF (for this example use Influenza A_N2_SF60).

- Just above the VT table, click the menu button "Find a VT(s)".

- A panel shows the VT-1 sequence. Make these edits: 118 = ?, 292 = ?, and 371 = ?. Click Search.



VARIANT TYPES

"[ ]" : Indicates an insertion.

Edit specific positions in this VT-1 sequence with IUPAC symbols or use "?" as a wild-card. If necessary, use the horizontal scroll bar to access the entire SF. Click Search to find VT(s) conforming to the edited sequence. Click Reset to restore this panel to the default VT-1 sequence.

Enter Sequence Variation to Find

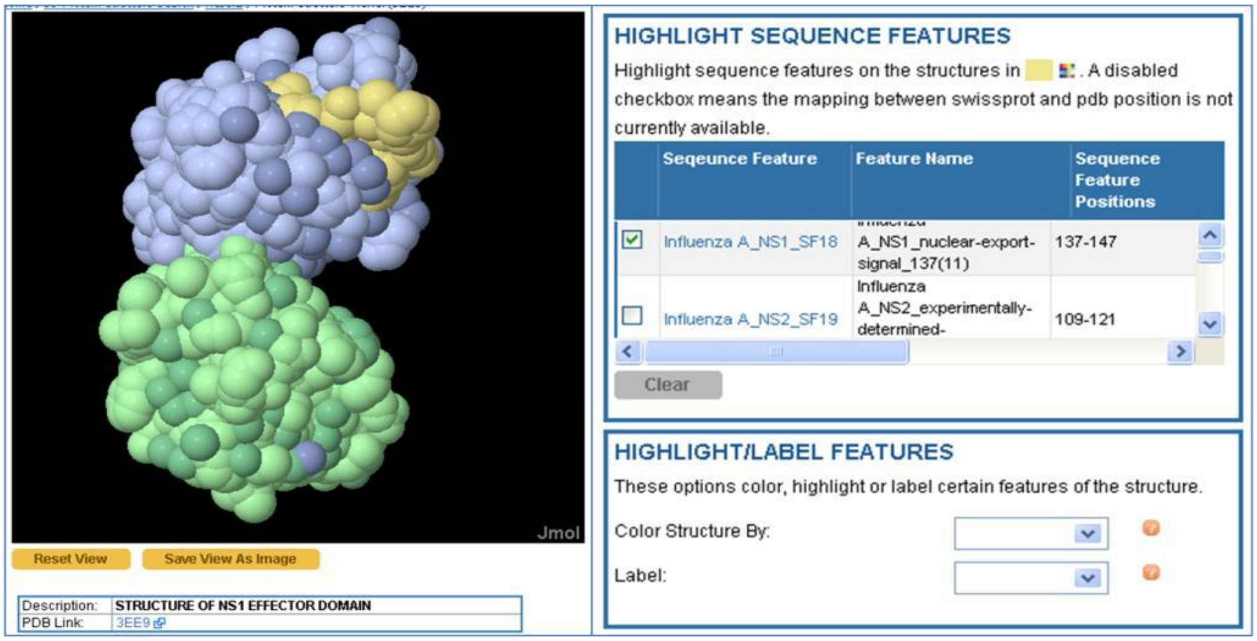
118	151	152	224	276	292	371	406	119	156	178	179	198	222	227	274	277
?	D	R	R	E	?	?	Y	E	R	W	S	D	I	E	H	E

Fig 6: Screenshot of the 'Find a VT' section on the SFVT details page

- In addition to the VT-1 sequence for comparison, the search returns VT-17, VT-36, VT-37 and VT-39. Two of these are substituted at 292 and two are substituted at 371. For each search result, click the Strain Count link to find strain(s) harboring these N2 substitutions.

Representation of SFVTs in IRD:

SFVTs can be highlighted on protein 3D structures, viewed as multiple sequence alignments in Jalview and downloaded in excel format. Pre-computed phylogenetic trees of VTs are also available for each SFVT group. (See screenshots of all these features below- Figures 7, 8 and 9):



Description:	STRUCTURE OF NS1 EFFECTOR DOMAIN
PDB Link:	3EE9

HIGHLIGHT SEQUENCE FEATURES

Highlight sequence features on the structures in . A disabled checkbox means the mapping between swissprot and pdb position is not currently available.

	Seqeunce Feature	Feature Name	Sequence Feature Positions
☑	Influenza A_NS1_SF18	A_NS1_nuclear-export-signal_137(11)	137-147
☐	Influenza A_NS2_SF19	Influenza A_NS2_experimentally-determined-	109-121

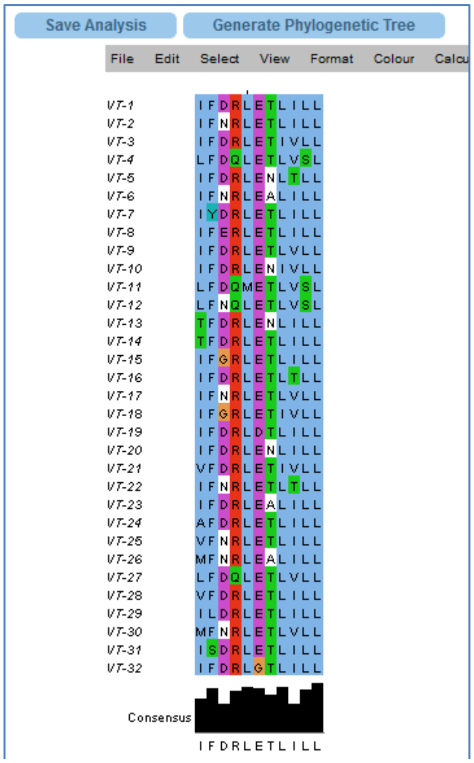
HIGHLIGHT/LABEL FEATURES

These options color, highlight or label certain features of the structure.

Color Structure By:

Label:

Fig 7. NS1 protein effector domain (PDB: 3EE9). The residues highlighted in brown between amino acids 137-147 represents the *Influenza A_NS1_nuclear-export-signal_137(11)* sequence feature.



```
VT-1      IFDRLETLILL
VT-2      IFNRLETLILL
VT-3      IFDRLETIVLL
VT-4      LFDQLETLVSL
VT-5      IFDRLENLTLL
VT-6      IFNRLEALILL
VT-7      IYDRLETLILL
VT-8      IFERLETLILL
VT-9      IFDRLETLVLL
VT-10     IFDRLENIVLL
VT-11     LFDQMETLVSL
VT-12     LFNQLETLVSL
VT-13     TFDRLENLILL
VT-14     TFDRLETLILL
VT-15     IFGRLETLILL
VT-16     IFDRLETLTLL
VT-17     IFNRLETLVLL
VT-18     IFGRLETIVLL
VT-19     IFDRLDTLILL
VT-20     IFDRLENLILL
VT-21     VFDRLETIVLL
VT-22     IFNRLETLTLL
VT-23     IFDRLEALILL
VT-24     AFDRLETLILL
VT-25     VFNRLETLILL
VT-26     MFNRLEALILL
VT-27     LFDQLETLVLL
VT-28     VFDRLETLILL
VT-29     ILDRLETLILL
VT-30     MFNRLETLVLL
VT-31     ISDRLETLILL
VT-32     IFDRLGTLILL

Consensus
          IFDRLETLILL
```

Fig 8: MSA of the SVTs of NS1 nuclear export protein

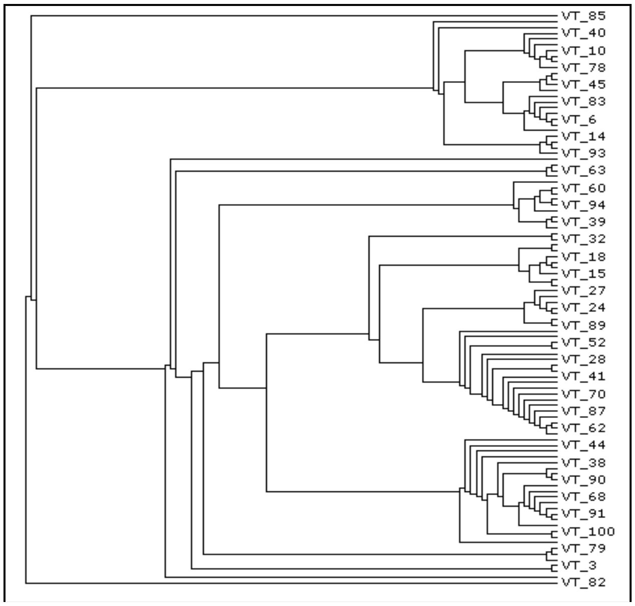
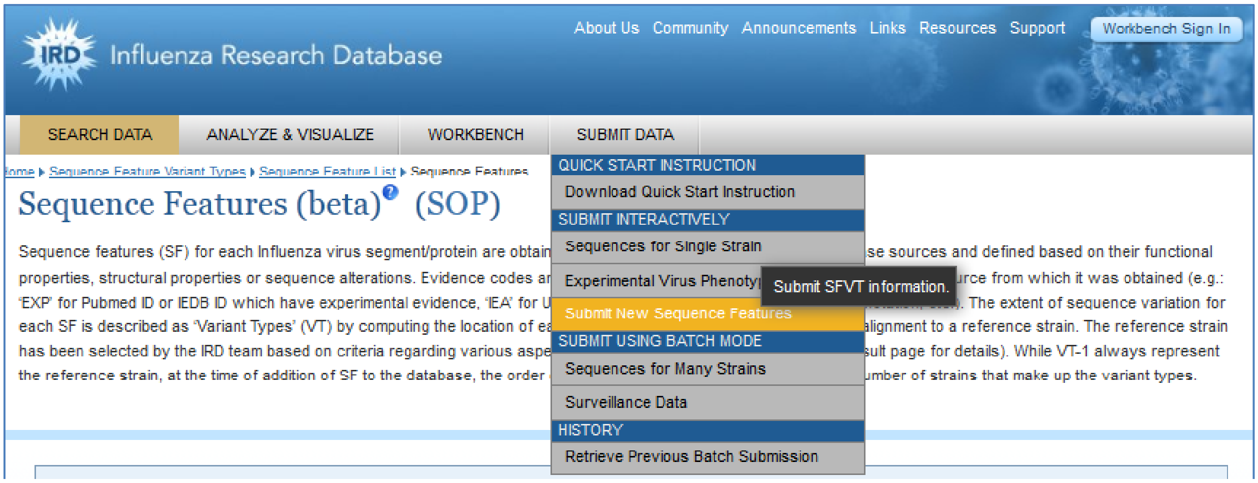


Fig 9: A pre-computed phylogenetic tree of the SVTs of NS1 nuclear export protein SF.

SUBMIT NEW SEQUENCE FEATURES:

- In an effort to keep the SF list up-to-date and to facilitate greater community involvement in annotation efforts, we have created a novel SF submission interface in IRD. (**Fig 10**)


IRD
Influenza Research Database
About Us
Community
Announcements
Links
Resources
Support
Workbench Sign In
SEARCH DATA
ANALYZE & VISUALIZE
WORKBENCH
SUBMIT DATA
QUICK START INSTRUCTION
Download Quick Start Instruction
SUBMIT INTERACTIVELY
Sequences for Single Strain
Experimental Virus Phenoty
Submit SFVT information.
Submit New Sequence Features
SUBMIT USING BATCH MODE
Sequences for Many Strains
Surveillance Data
HISTORY
Retrieve Previous Batch Submission
Home ▸ Sequence Feature Variant Types ▸ Sequence Feature List ▸ Sequence Features

Sequence Features (beta) (SOP)

Sequence features (SF) for each Influenza virus segment/protein are obtain se sources and defined based on their functional
properties, structural properties or sequence alterations. Evidence codes ar rce from which it was obtained (e.g.:
'EXP' for Pubmed ID or IEDB ID which have experimental evidence, 'IEA' for U . The extent of sequence variation for
each SF is described as 'Variant Types' (VT) by computing the location of ea alignment to a reference strain. The reference strain
has been selected by the IRD team based on criteria regarding various aspe sult page for details). While VT-1 always represent
the reference strain, at the time of addition of SF to the database, the order umber of strains that make up the variant types.

Fig 10: A screenshot of the SF submission page in IRD.

- The SF submission page is accessible from the 'Submit Data' tab on the IRD homepage.

- This tool provides a user-friendly interface for online submission of new SFs. Drop-down menus are made available for fields such as Virus type, protein, SF category, etc. to maximize the efficiency.

- Newly submitted SFs are first manually inspected and then validated using appropriate QC measures to ensure accuracy and authenticity, and to avoid duplicates.

- Following QC process, the new SFs are either accepted or rejected and the contributors are notified of the decisions. Submitters of accepted SFs are acknowledged on the IRD site.