

Reference-guided annotation of viral genomes

Eric Nawrocki

National Center for Biotechnology Information
National Institutes of Health



Most prevalent viral genomes in GenBank*

rank	species	#seqs	family	type	host	#CDS	#mature peptides
1	Influenza	503,115	Orthomyxoviridae	(-)ssRNA	humans+	11	-
2	Rotavirus A	58,405 [†]	Reoviridae	dsRNA	humans	12	-
3	Hepatitis B	9211	Hepadnaviridae	dsDNA-RT	humans	7	-
4	Dengue	4853	Flaviviridae	(+)ssRNA	humans	1	14
5	HIV-1	2597	Retroviridae	ssRNA-RT	humans	10	14
6	Hepatitis C	2185	Flaviviridae	(+)ssRNA	humans	2	10
7	Porcine circovirus	1905	Circoviridae	ssDNA	pigs	3	-
8	West Nile	1667	Flaviviridae	(+)ssRNA	humans	3	16
9	Ebola	1384	Flaviviridae	(+)ssRNA	humans	9	-
10	Enterovirus A	1222	Picornoviridae	(+)ssRNA	humans	1	11
11	RSV	1122	Orthopneumovirus	(-)ssRNA	humans	11	-
12	Norwalk virus	1009	Caliciviridae	(+)ssRNA	humans	3	6
13	Maize streak virus	884	Geminiviridae	ssDNA	plants	4	-
14	Rabies lyssavirus	826	Rhabdoviridae	(-)ssRNA	humans+	5	-
15	Enterovirus C	765	Picornoviridae	(+)ssRNA	humans	1	13

† sum of 11 segments

The Virus Variation Resource is a powerful tool for viral research

- value added database that includes annotations not in GenBank
- allows users to:
 - select subsets of data based on desired criteria (host, country, gene, etc.)
 - download alignments
 - compute trees
 - more...

The screenshot shows the 'Virus Variation' page on the NCBI website. At the top, there's a banner with a flask icon and text encouraging users to try a new experimental search interface and provide feedback via NCBI Insights. A 'NEW' badge is visible in the top right corner. Below the banner, the title 'Virus Variation Resource' is centered. A grid of virus names is displayed in three rows: Influenza virus, Dengue virus, Zika virus; Rotavirus, West Nile virus, MERS coronavirus; and Ebolavirus. At the bottom, there are three boxes: 'Help center' with a question mark icon, 'How to cite us' with a double quote icon, and 'Related links' with an information icon. An 'Announcements' section at the very bottom contains a message about the experimental search interface and a link to NCBI Insights.

Try our new, experimental Virus Sequence Search Interface and send us your feedback! Learn more about this tool at [NCBI Insights](#).

Virus Variation Resource

Influenza virus Dengue virus Zika virus

Rotavirus West Nile virus MERS coronavirus

Ebolavirus

Help center How to cite us Related links

Announcements

★ Try our new, experimental Virus Sequence Search Interface and send us your feedback! Learn more about this tool at [NCBI Insights](#).

The Virus Variation Resource is a powerful tool for viral research

NCBI Resources How To Sign in to NCBI

Virus Variation Dengue virus database

How to cite Contact us Help

Virus Variation home Virus resources ▾

Select sequence type

Protein Nucleotide Full-length sequences only

Define search set

Structural Non-structural

C	M	E	NS1	NS2A	NS2B	NS3	NS4A	2K	NS4B	NS5
---	---	---	-----	------	------	-----	------	----	------	-----

Type Disease Host Region/Country Genome region

any	any	any	any	any	any
1	known	Human	regions	C	
2	DF	Mammal	Africa	M	
3	DHF	Mosquito	Asia	E	
4	DSS	Primate	Europe	NS1	

Collection date: to Year Month Day Year Month Day

Release date: to Year Month Day

Additional filters ▾

Keyword Search in

Get sequences from:

Include Laboratory isolates
 Include Vaccine strains
 Include Environmental isolates

Get sequences by accession ▾

Enter a comma or space separated list of sequence accessions or upload text file with this list.

Upload No file chosen Accessions

Our initial goal was to develop an annotation pipeline to benefit Virus Variation

- general method for annotating existing and new viral genome sequences for a species using trusted annotation for that species (e.g. RefSeq)
- identify interesting characteristics that Virus Variation can allow users to sort/select based on:
 - identification of high quality sequences that meet specific expectations
 - identification of sequences that deviate from expectations in various ways
 - * early stop codon
 - * above or below a specific fractional identity to reference
 - * more...

Our initial goal was to develop an annotation pipeline to benefit Virus Variation

- general method for annotating existing and new viral genome sequences for a species using trusted annotation for that species (e.g. RefSeq)
- identify interesting characteristics that Virus Variation can allow users to sort/select based on:
 - identification of high quality sequences that meet specific expectations
 - identification of sequences that deviate from expectations in various ways
 - * early stop codon
 - * above or below a specific fractional identity to reference
 - * more...

A new goal is to use this annotation pipeline to validate/annotate incoming submissions of sequences from commonly deposited viruses

- use unexpected features identified by pipeline to flag issues for indexers and submitters to deal with
- currently testing on Norovirus, and soon to be testing on Ebolavirus

Four pilot species were chosen from the 15 most prevalent

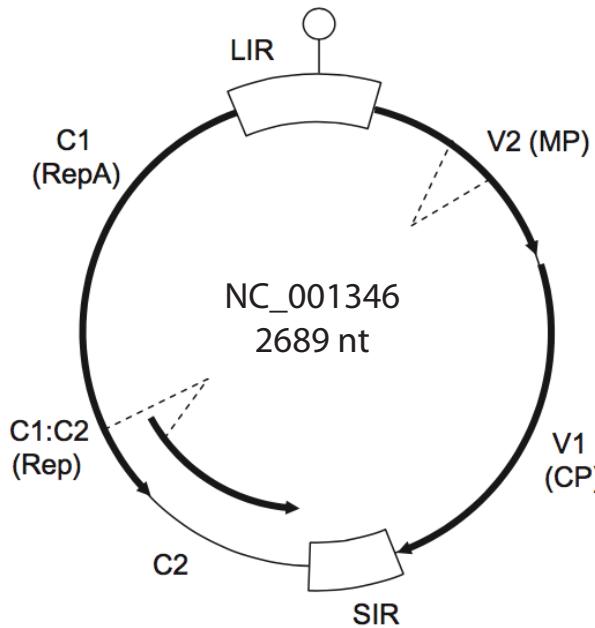
rank	species	#seqs	family	type	host	#CDS	#mature peptides
1	Influenza	503,115	Orthomyxoviridae	(-)ssRNA	humans+	11	-
2	Rotavirus A	58,405*	Reoviridae	dsRNA	humans	12	-
3	Hepatitis B	9211	Hepadnaviridae	dsDNA-RT	humans	7	-
4	Dengue	4853	Flaviviridae	(+)ssRNA	humans	1	14
5	HIV-1	2597	Retroviridae	ssRNA-RT	humans	10	14
6	Hepatitis C	2185	Flaviviridae	(+)ssRNA	humans	2	10
7	Porcine circovirus	1905	Circoviridae	ssDNA	pigs	3	-
8	West Nile	1667	Flaviviridae	(+)ssRNA	humans	3	16
9	Ebola	1384	Flaviviridae	(+)ssRNA	humans	9	-
10	Enterovirus A	1222	Picornoviridae	(+)ssRNA	humans	1	11
11	RSV	1122	Orthopneumovirus	(-)ssRNA	humans	11	-
12	Norwalk virus	1009	Caliciviridae	(+)ssRNA	humans	3	6
13	Maize streak virus	884	Geminiviridae	ssDNA	plants	4	-
14	Rabies lyssavirus	826	Rhabdoviridae	(-)ssRNA	humans+	5	-
15	Enterovirus C	765	Picornoviridae	(+)ssRNA	humans	1	13

† sum of 11 segments

Overview of annotation pipeline for Maize Streak Virus

INPUT:

1. RefSeq annotation



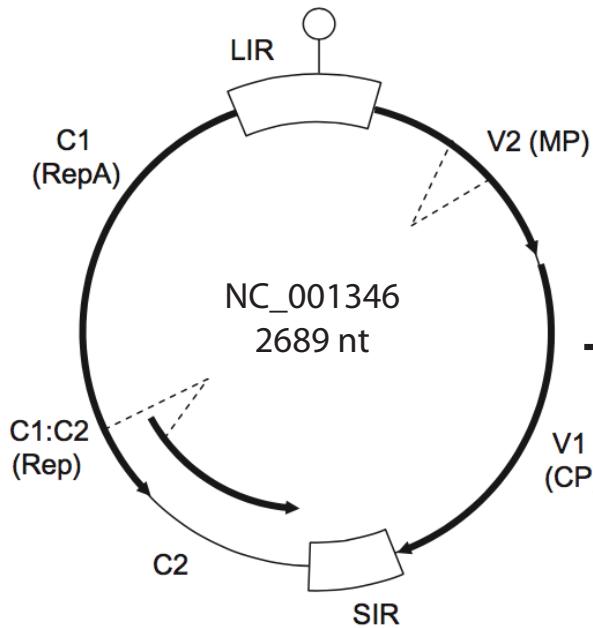
2. Target sequence(s) to annotate:

- HQ693295
 - EU628610
 - EU628635
 - FJ882144
 -
 -
 - HQ693434
- ⋮

Overview of annotation pipeline for Maize Streak Virus

INPUT:

1. RefSeq annotation

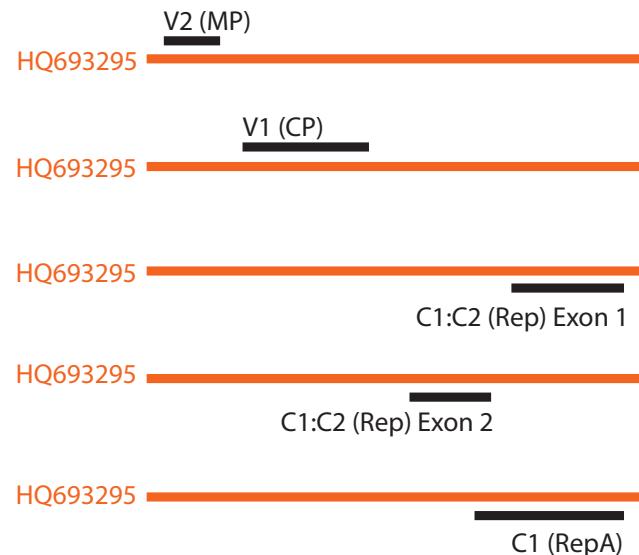


Build homology models and search targets:

V2 (MP)	306nt
V1 (CP)	735nt
C1:C2 (Rep) Exon 1	642nt
C1:C2 (Rep) Exon 2	441nt
C1 (RepA)	819nt

2. Target sequence(s) to annotate:

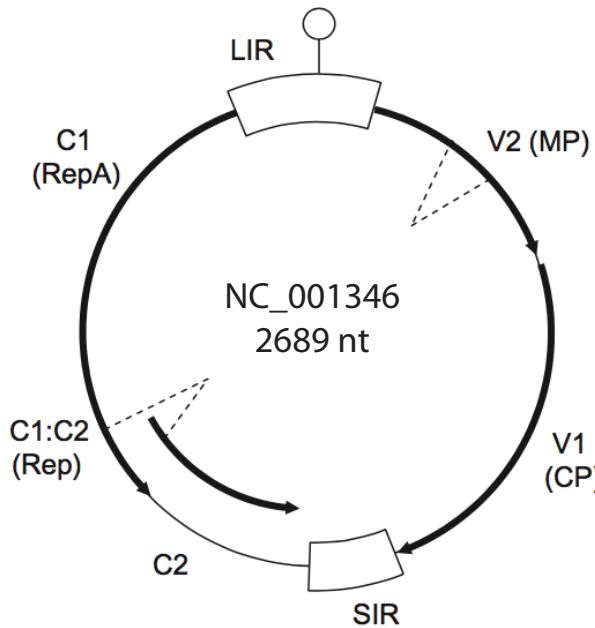
HQ693295
EU628610
EU628635
FJ882144
HQ693434



Overview of annotation pipeline for Maize Streak Virus

INPUT:

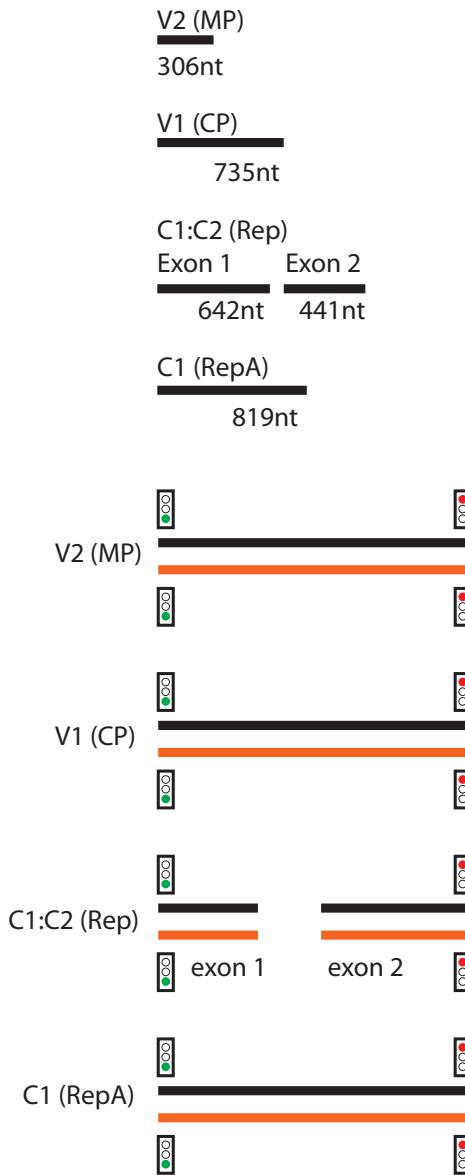
1. RefSeq annotation



2. Target sequence(s) to annotate:

HQ693295
EU628610
EU628635
FJ882144
HQ693434

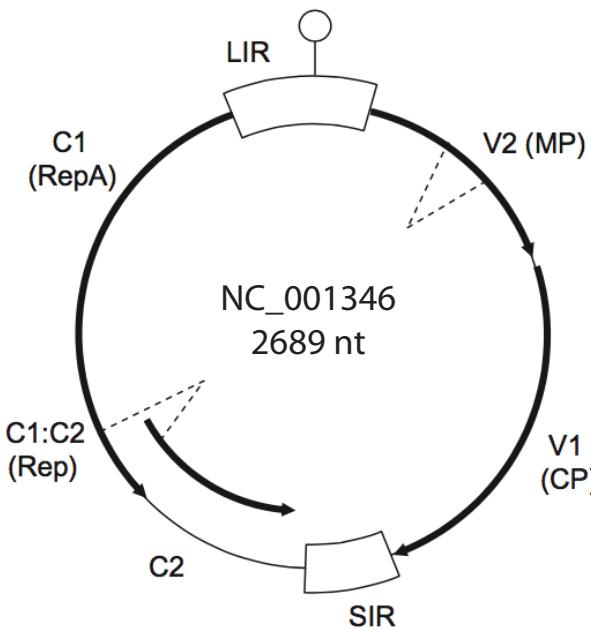
Combine models
into 'features', translate,
find stop codons, and
identify errors:



Overview of annotation pipeline for Maize Streak Virus

INPUT:

1. RefSeq annotation



2. Target sequence(s) to annotate:

HQ693295

EU628610

EU628635

FJ882144

HQ693434

⋮

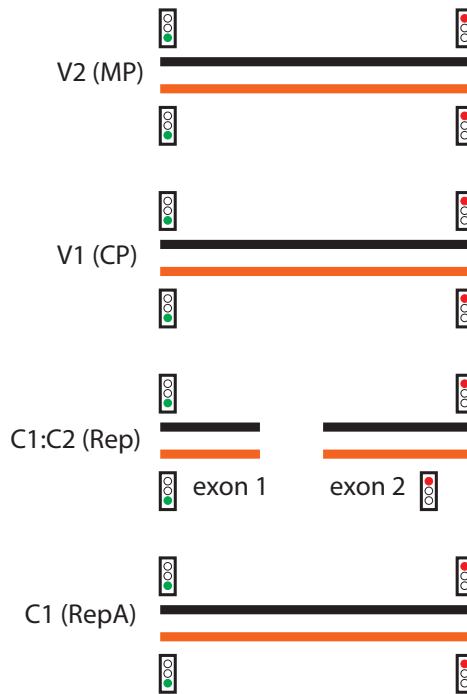
Combine models
into 'features', translate,
find stop codons, and
identify errors:

V2 (MP)
306nt

V1 (CP)
735nt

C1:C2 (Rep)
Exon 1 Exon 2
642nt 441nt

C1 (RepA)
819nt



OUTPUT:

1. Tabular annotations of all features:

EU628610:MP:157-462
EU628610:CP:473-1207
EU628610:Rep:2530-1889,1796-1371
EU628610:RepA:2530-1712

2. List of all 'error codes':

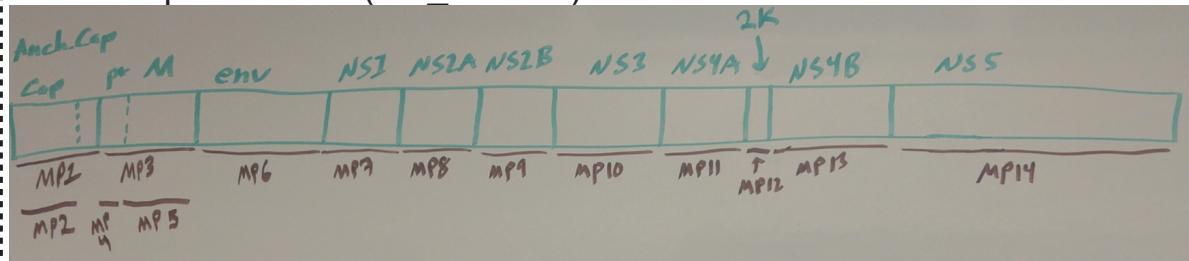
EU628610:Rep:trc

3. Nucleotide and protein multiple alignments (optional)

Overview of annotation pipeline for Dengue Virus

INPUT:

1. RefSeq annotation (NC_001477)



2. Target sequence(s) to annotate:

KC762654

EU179860

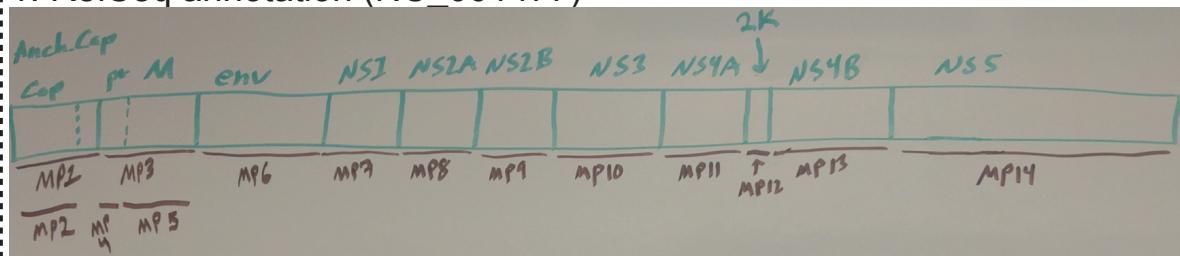
A75711

DQ193572

Overview of annotation pipeline for Dengue Virus

INPUT:

1. RefSeq annotation (NC_001477)



2. Target sequence(s) to annotate:

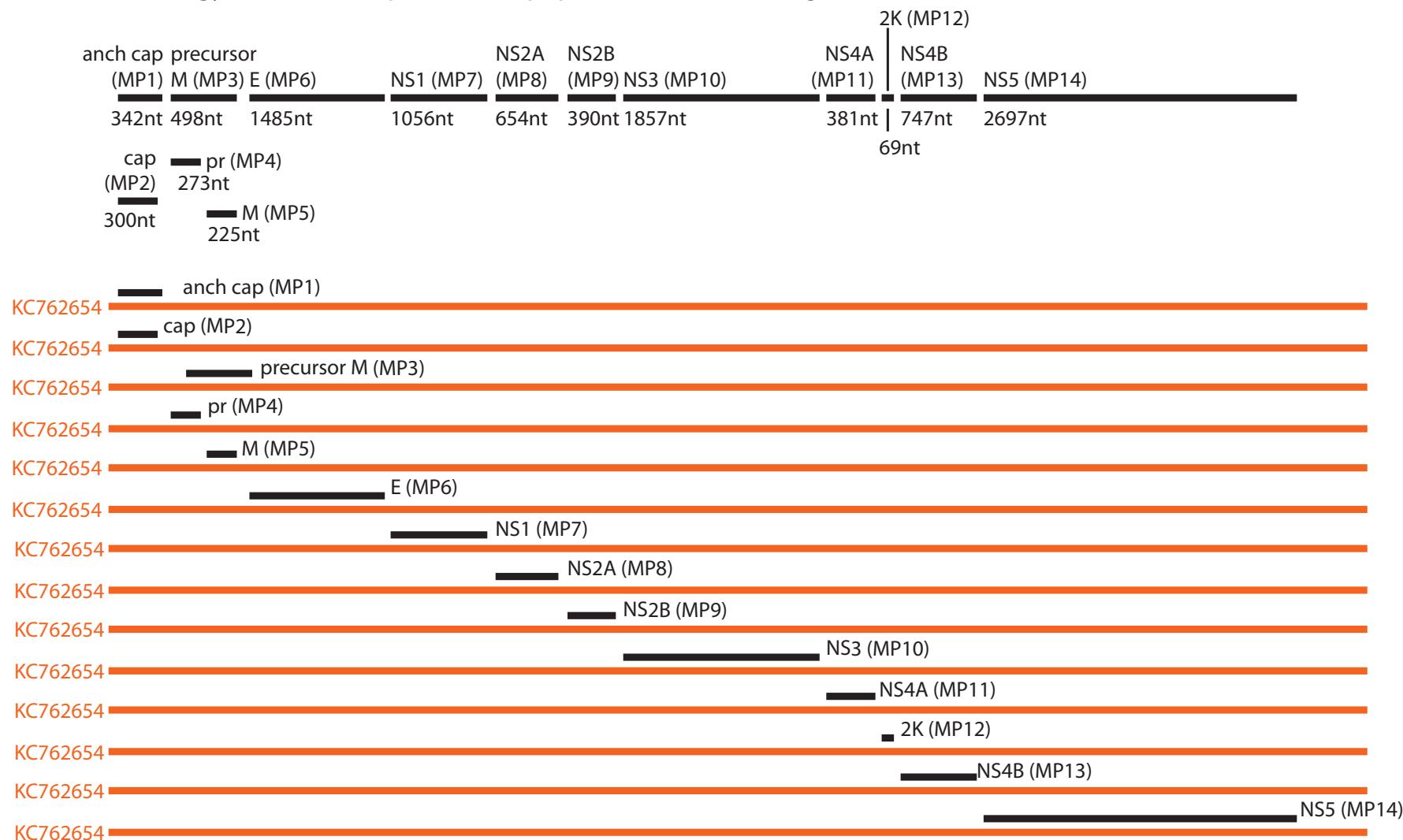
KC762654

EU179860

A75711

DQ193572

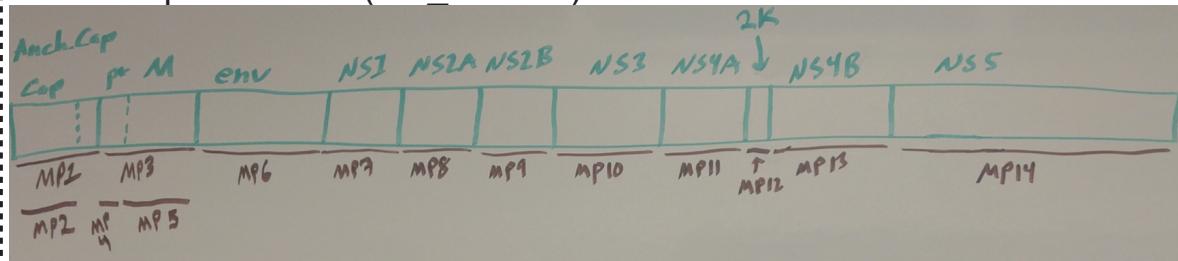
Build 14 homology models (one per mature peptide) and search targets:



Overview of annotation pipeline for Dengue Virus

INPUT:

1. RefSeq annotation (NC_001477)



2. Target sequence(s) to annotate:

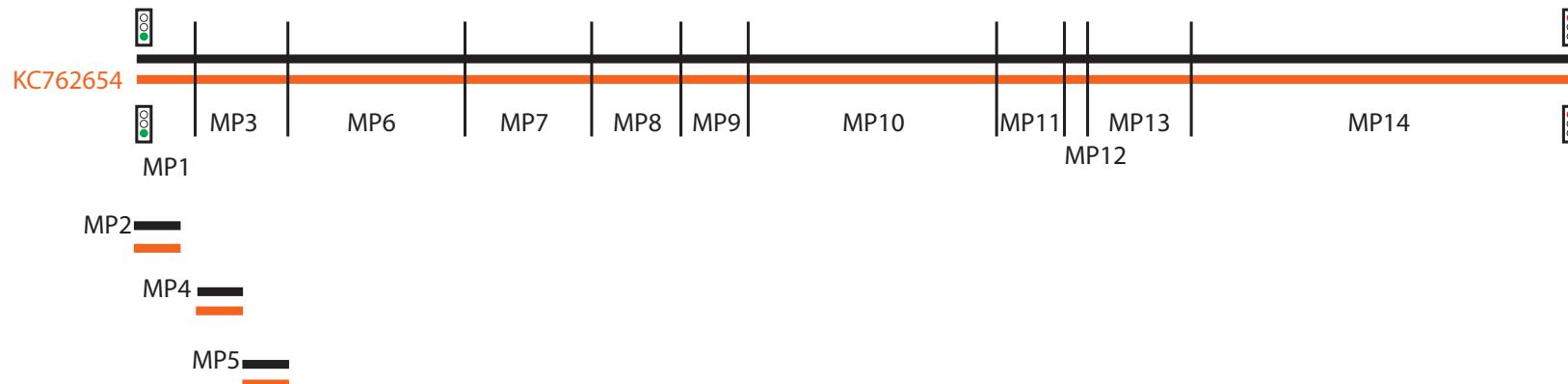
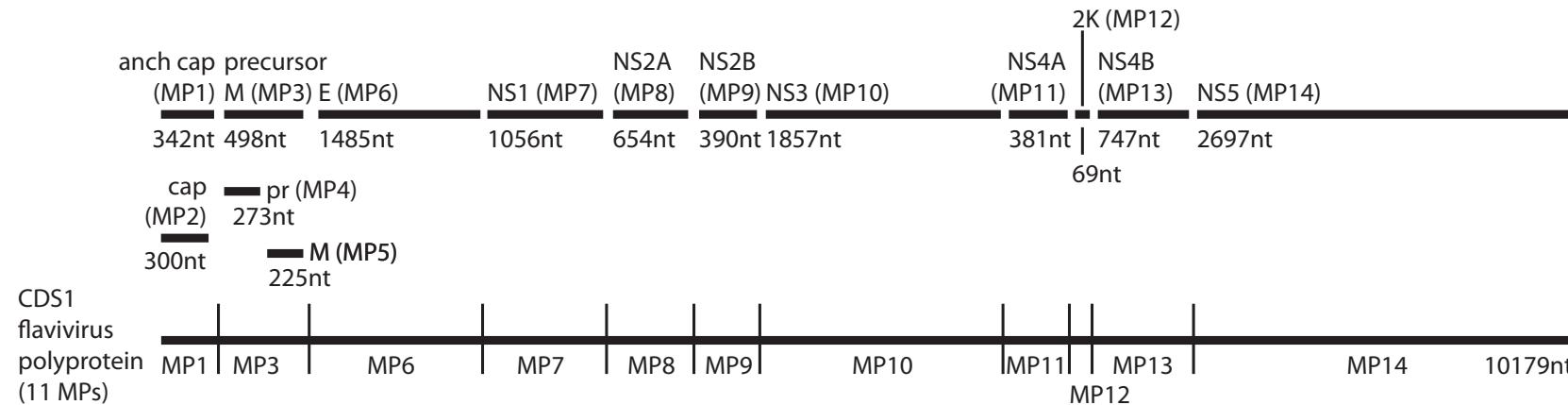
KC762654

EU179860

A75711

DQ193572

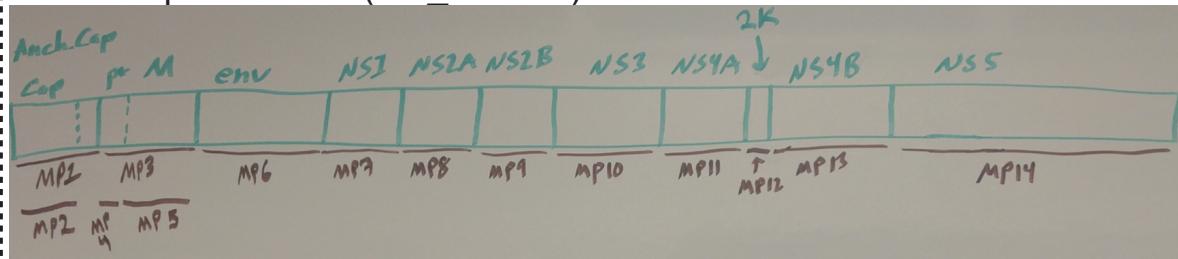
Combine models into 'features' (14 MPs and 1 CDS), translate, find stop codons, and identify errors:



Overview of annotation pipeline for Dengue Virus

INPUT:

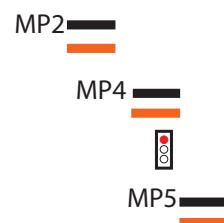
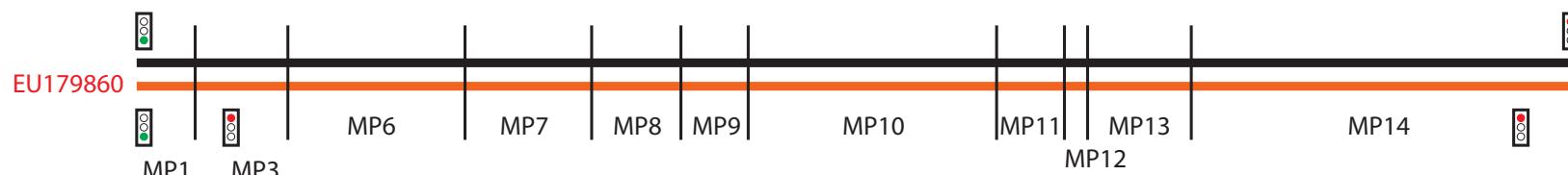
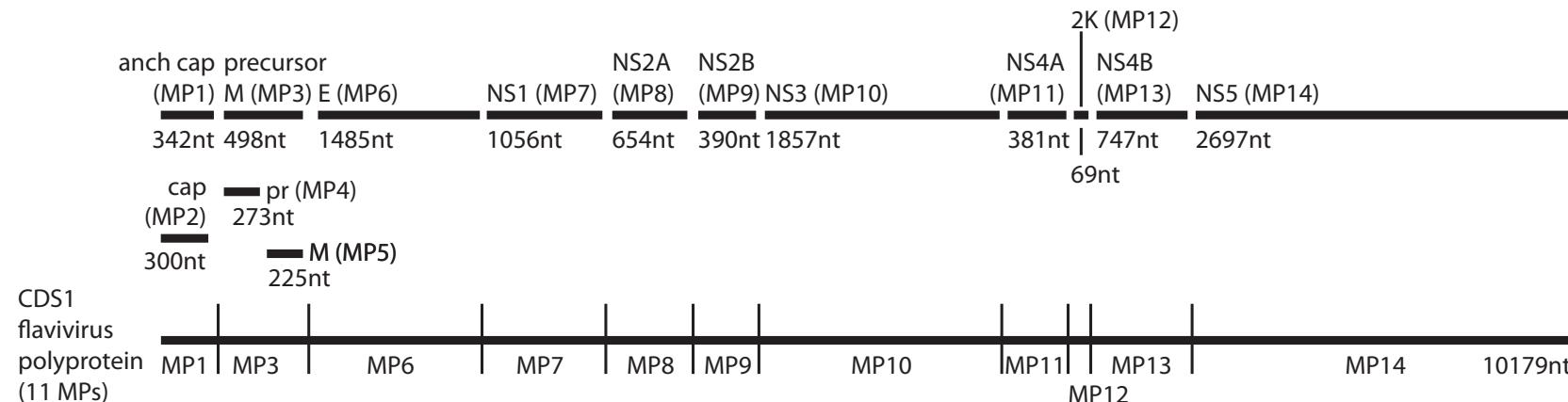
1. RefSeq annotation (NC_001477)



2. Target sequence(s) to annotate:

KC762654
EU179860
A75711
DQ193572

Combine models into 'features' (14 MPs and 1 CDS), translate, find stop codons, and identify errors:



OUTPUT:

1. Tabular annotations of all features
2. List of all 'error codes'
3. Nucleotide and protein multiple alignments (optional)

Error codes: 17 abnormal situations

- Per-feature (e.g. CDS, mature peptide) errors:
 - Unexpected stop codon errors (trc, ext, nst, ntr)
 - Missing expected features (str, stp, nm3)
 - Problem with homology search prediction (bd5, bd3, nop)
 - Unexpected relationship to other features (olp, aja, ajb)
 - Problem annotating CDS due to mature peptide errors (aji, int, inp)
- Per-sequence errors:
 - Lack of exactly one origin sequence (ori)

trc error code reports a truncation due to an early stop codon

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors			 MP#(i-1) MP#i MP#(i+1)	
trc (truncation) in-frame stop codon exists 5' of predicted stop			 MP#(i-1) MP#i MP#(i+1)	 MP1 2 3 4 5 6 ?

Example output from annotation script:

```
EU628610      3  CDS#3      trc  in-frame stop codon exists 5' of stop position predicted by homology to reference
                           [homology search predicted 1796..1356 exon 2 of 2 revised to 1796..1371 (stop shifted 15 nt)]
```

ext: extended feature due to a missing stop codon

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors	  	  	MP#(i-1) MP#i MP#(i+1)	
ext (extension) first in-frame stop codon exists 3' of predicted stop	  	  	MP#(N-2) MP#(N-1) MP#N of N	

Example output from annotation script:

```
HM631854      14  MP#14      ext  first in-frame stop codon exists 3' of stop position predicted by homology to reference
                  [homology search predicted 7544..10230 revised to 7544..10288 (stop shifted 58 nt)]

HM631854      15  CDS(MP)#1    ext  first in-frame stop codon exists 3' of stop position predicted by homology to reference
                  [homology search predicted 74..10233 revised to 74..10291 (stop shifted 58 nt)]
```

olp: lack of an expected overlap with another feature

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors				
olp (overlap) feature does not overlap with same set of features as in reference				

Example output from annotation script:

```
FJ562227      6  CDS#6      olp  feature does not overlap with same set of features as in reference [-(6.1,1.1),-(6.1,7.1)]
```

For screening submissions: mapping error codes to errors and warnings

- Error codes are too complex to return to submitters
- Sometimes error codes are legitimate (true early stop)
- Sometimes error codes reflect artefacts in the sequence (assembly error)
- Example: trc error code leads to misc_feature and note

```
>Feature NC_029646-1ntgap-outframe
5      5103  gene
                  gene ORF1
5      5103  misc_feature
                  note similar to nonstructural polyprotein; contains premature stop codon
                  gene ORF1
5      994   misc_feature
                  note similar to p48; polyprotein may not be translated
995    2091   misc_feature
                  note similar to NTPase; polyprotein may not be translated
2092    2628   misc_feature
                  note similar to p22; polyprotein may not be translated
2629    3027   misc_feature
                  note similar to VPg; polyprotein may not be translated
3028    3570   misc_feature
                  note similar to Pro; polyprotein may not be translated
3571    5100   misc_feature
                  note similar to RdRp; polyprotein may not be translated
5084    6691   gene
                  gene ORF2
5084    6691   CDS
                  product VP1
                  gene ORF2
6691    7470   gene
                  gene ORF3
6691    7470   CDS
                  product VP2
                  gene ORF3

Additional note(s) to submitter:
ERROR: Deletion of Nucleotides: Contains internal deletion of 1 nucleotide in nonstructural polyprotein/NTPase
ERROR: CDS Has Frameshift: Contains internal deletion of 1 nucleotide(s) in nonstructural polyprotein/NTPase
```

Future directions (long term)

- Protein homology searches to supplement or replace nucleotide searches
- Multiple alignment (profile) based homology searches
- Annotate structural RNA features (Infernal is well-suited for this)

Acknowledgements

Alejandro Schäffer

David Landsman

David Lipman

J. Rodney Brister

Ilene Mizrachi

Eneida Hatcher

Linda Yankie

Olga Blinkova

Anatoly Mnev

Sergey Zhdanov

Yiming Bao