

# Reference-guided annotation of viral genomes

Eric Nawrocki

Alejandro Schäffer's group

National Center for Biotechnology Information  
National Institutes of Health



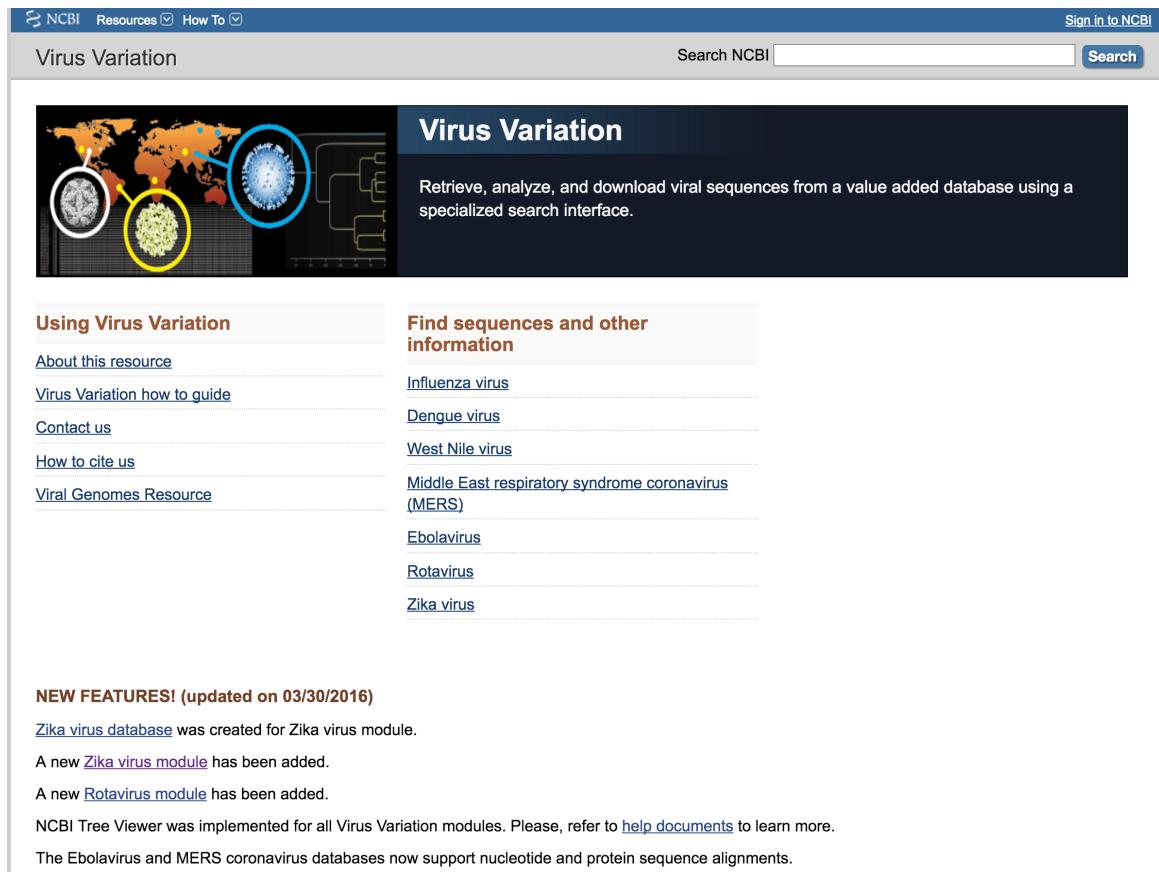
# Most prevalent viral genomes in GenBank\*

rank	species	#seqs	family	type	host	#cds	#mature peptides
1	Hepatitis B	7061	Hepadnaviridae	dsDNA-RT	humans	7	-
2	Dengue	3875	Flaviviridae	(+)ssRNA	humans	1	14
3	Rotavirus	22536 <sup>†</sup>	Reoviridae	dsRNA	humans	12	-
4	HIV-1	2132	Retroviridae	ssRNA-RT	humans	10	14
5	Porcine circovirus	1559	Circoviridae	ssDNA	pigs	3	-
6	Hepatitis C	1457	Flaviviridae	(+)ssRNA	humans	2	10
7	West Nile	1109	Flaviviridae	(+)ssRNA	humans	3	16
8	Ebola	1105	Flaviviridae	(+)ssRNA	humans	9	-
9	Enterovirus A	932	Picornoviridae	(+)ssRNA	humans	1	11
10	RSV	623	Paramyxoviridae	(-)ssRNA	humans	11	-
11	Enterovirus C (Polio)	618	Picornoviridae	(+)ssRNA	humans	1	13
12	JC polyomavirus	598	Polyomaviridae	dsDNA	humans	6	-
13	PRRSV	520	Arteriviridae	(+)ssRNA	pigs	11	13
14	Maize streak virus	508	Geminiviridae	ssDNA	plants	4	-
15	Norwalk virus	491	Caliciviridae	(+)ssRNA	humans	3	6

† 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 NCBI Virus Variation homepage. At the top, there's a navigation bar with links for NCBI, Resources, How To, Sign in to NCBI, and a search bar. The main header "Virus Variation" is centered above a banner. The banner features a world map with three highlighted regions: South America (yellow), Africa (yellow), and Asia (blue). Below the banner, there's a section titled "Using Virus Variation" with links to "About this resource", "Virus Variation how to guide", "Contact us", "How to cite us", and "Viral Genomes Resource". To the right, a section titled "Find sequences and other information" lists various virus types: Influenza virus, Dengue virus, West Nile virus, Middle East respiratory syndrome coronavirus (MERS), Ebolavirus, Rotavirus, and Zika virus. At the bottom of the page, there's a "NEW FEATURES!" section updated on 03/30/2016, which mentions the creation of a Zika virus database, the addition of a Zika virus module and Rotavirus module, the implementation of NCBI Tree Viewer, and support for nucleotide and protein sequence alignments in the Ebolavirus and MERS databases.

**Virus Variation**

Retrieve, analyze, and download viral sequences from a value added database using a specialized search interface.

**Using Virus Variation**

[About this resource](#)  
[Virus Variation how to guide](#)  
[Contact us](#)  
[How to cite us](#)  
[Viral Genomes Resource](#)

**Find sequences and other information**

[Influenza virus](#)  
[Dengue virus](#)  
[West Nile virus](#)  
[Middle East respiratory syndrome coronavirus \(MERS\)](#)  
[Ebolavirus](#)  
[Rotavirus](#)  
[Zika virus](#)

**NEW FEATURES! (updated on 03/30/2016)**

[Zika virus database](#) was created for Zika virus module.

A new [Zika virus module](#) has been added.

A new [Rotavirus module](#) has been added.

NCBI Tree Viewer was implemented for all Virus Variation modules. Please, refer to [help documents](#) to learn more.

The Ebolavirus and MERS coronavirus databases now support nucleotide and protein sequence alignments.

# 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

Add query Show results Clear form

#### 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

Add query Show results

## **Our goal is 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...

# Four pilot species were chosen from the 15 most prevalent

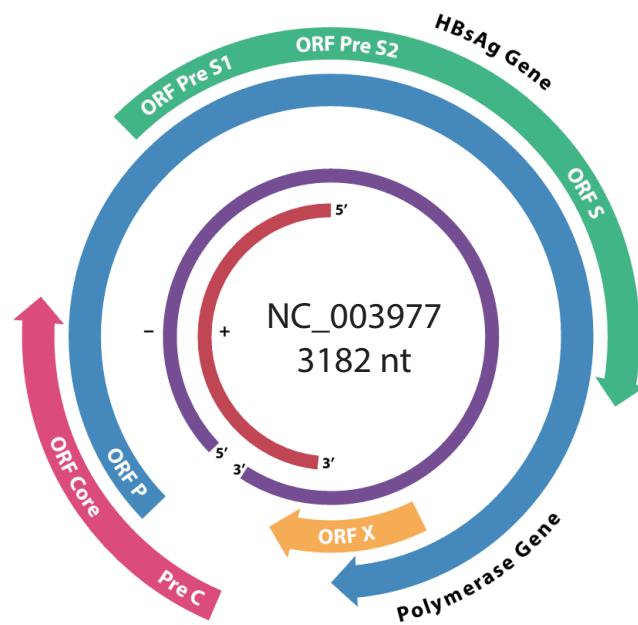
rank	species	#seqs	family	type	host	#cds	#mature peptides
1	Hepatitis B	7061	Hepadnaviridae	dsDNA-RT	humans	7	-
2	Dengue	3875	Flaviviridae	(+)ssRNA	humans	1	14
3	Rotavirus	22536*	Reoviridae	dsRNA	humans	12	-
4	HIV-1	2132	Retroviridae	ssRNA-RT	humans	10	14
5	Porcine circovirus	1559	Circoviridae	ssDNA	pigs	3	-
6	West Nile	1109	Flaviviridae	(+)ssRNA	humans	3	16
7	Hepatitis C	1457	Flaviviridae	(+)ssRNA	humans	2	10
8	Ebola	1105	Flaviviridae	(+)ssRNA	humans	9	-
9	Enterovirus A	932	Picornoviridae	(+)ssRNA	humans	1	11
10	RSV	623	Paramyxoviridae	(-)ssRNA	humans	11	-
11	Enterovirus C (Polio)	618	Picornoviridae	(+)ssRNA	humans	1	13
12	JC polyomavirus	598	Polyomaviridae	dsDNA	humans	6	-
13	PRRSV	520	Arteriviridae	(+)ssRNA	pigs	11	13
14	Maize streak virus	508	Geminiviridae	ssDNA	plants	4	-
15	Norwalk virus	491	Caliciviridae	(+)ssRNA	humans	3	6

\* sum of 11 segments

# Overview of annotation pipeline for Hepatitis B Virus

INPUT:

## 1. RefSeq annotation



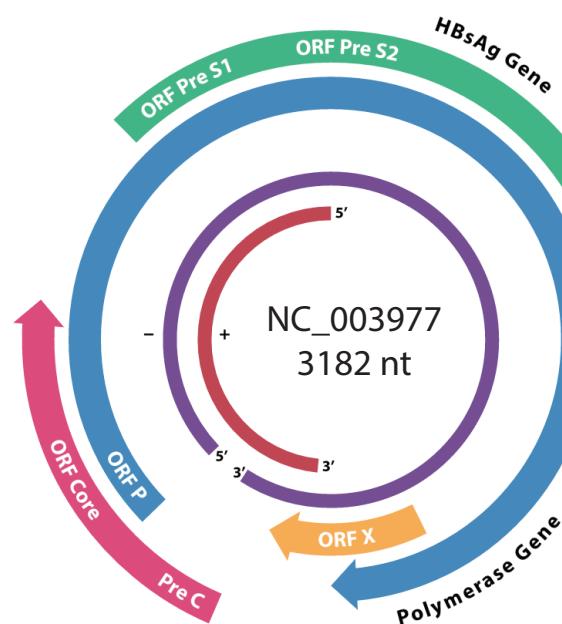
## 2. Target sequence(s) to annotate:

- KF922409
  - KJ410493
  - DQ448619
  - FJ349209
  - JQ027329
- ⋮

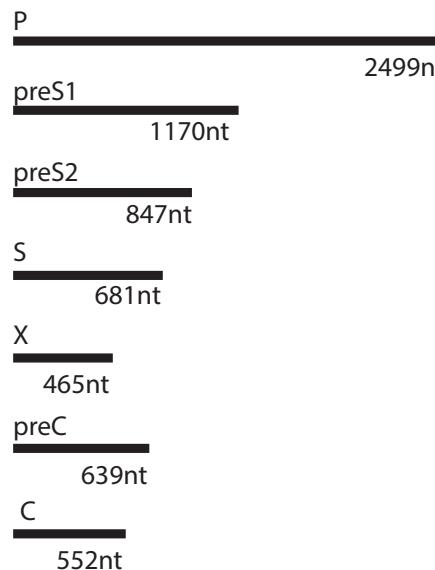
# Overview of annotation pipeline for Hepatitis B Virus

INPUT:

## 1. RefSeq annotation

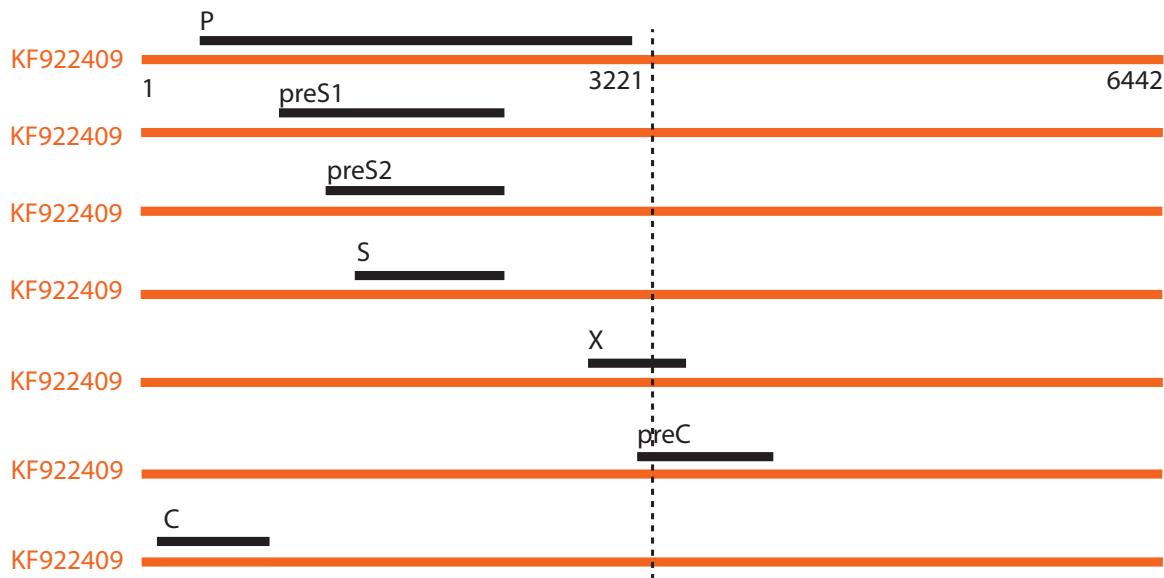


Build homology  
models and search  
duplicated genome targets:



## 2. Target sequence(s) to annotate:

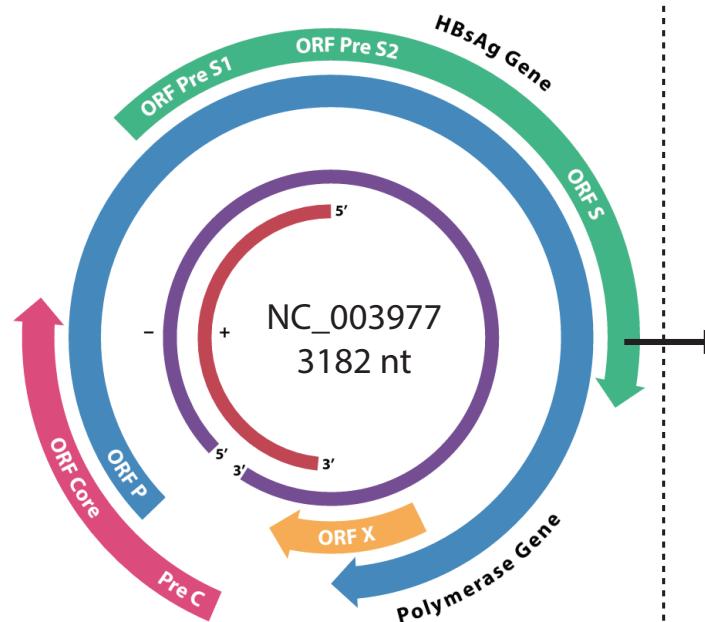
KF922409  
KJ410493  
DQ448619  
FJ349209  
  
JQ027329



# Overview of annotation pipeline for Hepatitis B Virus

INPUT:

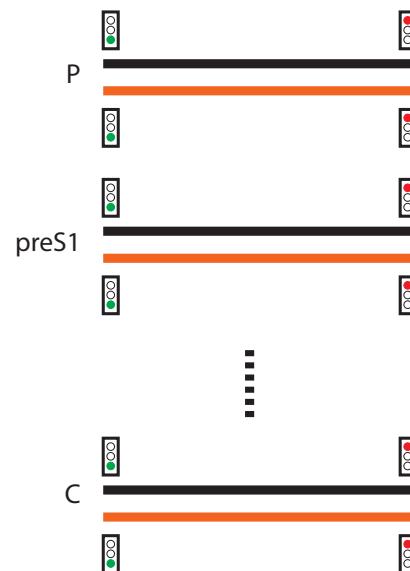
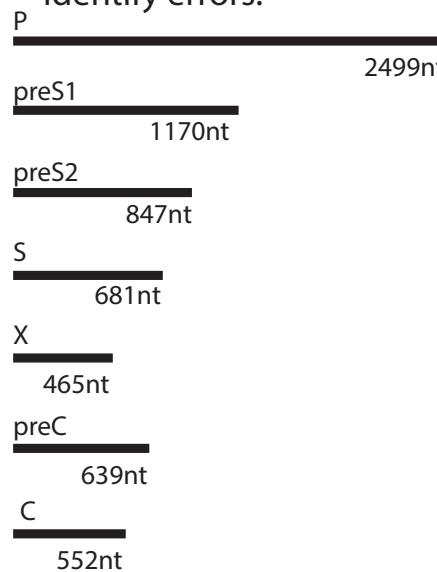
## 1. RefSeq annotation



## 2. Target sequence(s) to annotate:

KF922409  
KJ410493  
DQ448619  
FJ349209  
  
JQ027329

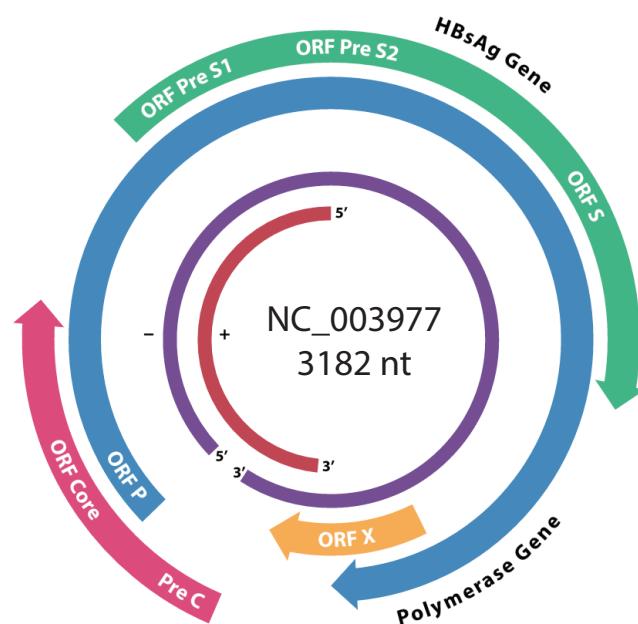
Combine models  
into 'features', translate,  
find stop codons, and  
identify errors:



# Overview of annotation pipeline for Hepatitis B Virus

## INPUT:

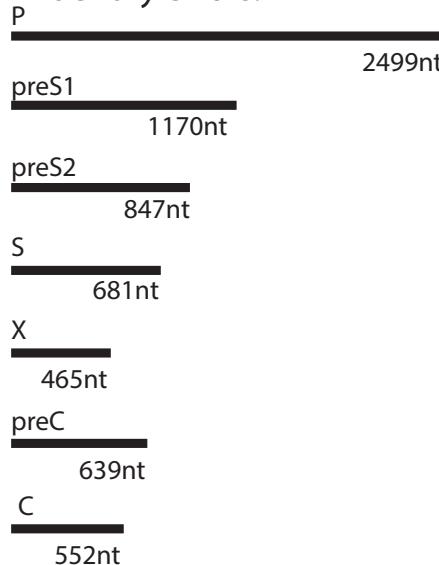
### 1. RefSeq annotation



### 2. Target sequence(s) to annotate:

KF922409  
**KJ410493**  
DQ448619  
FJ349209  
  
JQ027329

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



## OUTPUT:

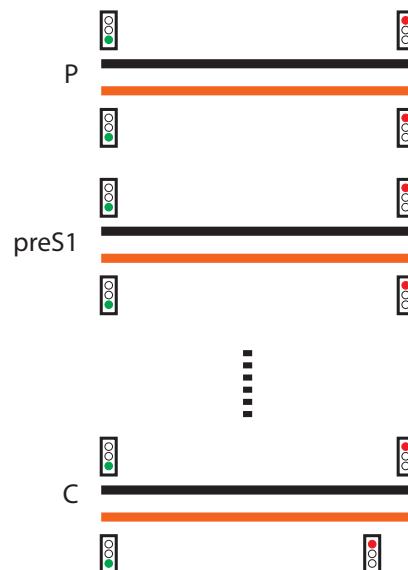
### 1. Tabular annotations of all features:

KJ410493:P:-909-1623  
KJ410493:preS1:-335-835  
KJ410493:preS2:-11-835  
KJ410493:S:155-835  
KJ410493:X:1374-1838  
KJ410493:preC:1814-2446  
KJ410493:C:1901-2446

### 2. List of all 'error codes':

EU628610:Rep:trc

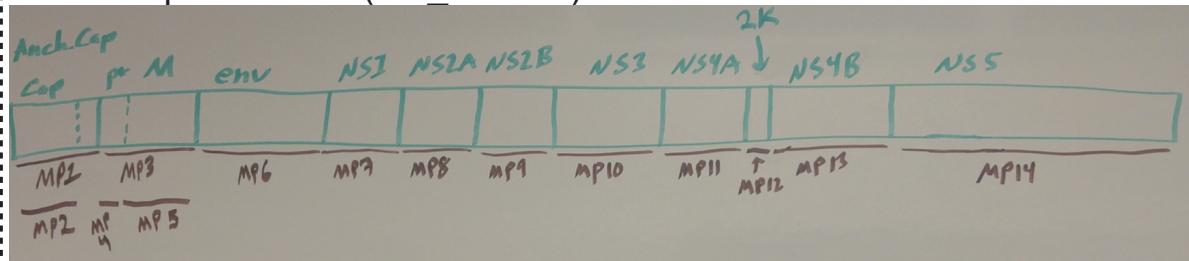
### 3. Nucleotide and protein multiple alignments (optional)



# Overview of annotation pipeline for Dengue Virus

## INPUT:

### 1. RefSeq annotation (NC\_001477)



### 2. Target sequence(s) to annotate:

KC762654

EU179860

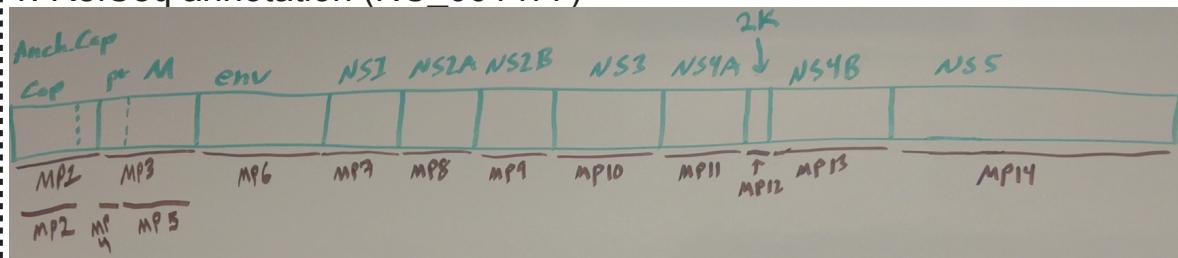
A75711

DQ193572

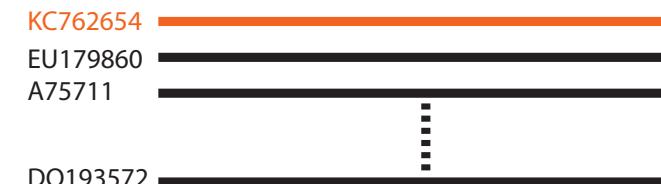
# Overview of annotation pipeline for Dengue Virus

INPUT:

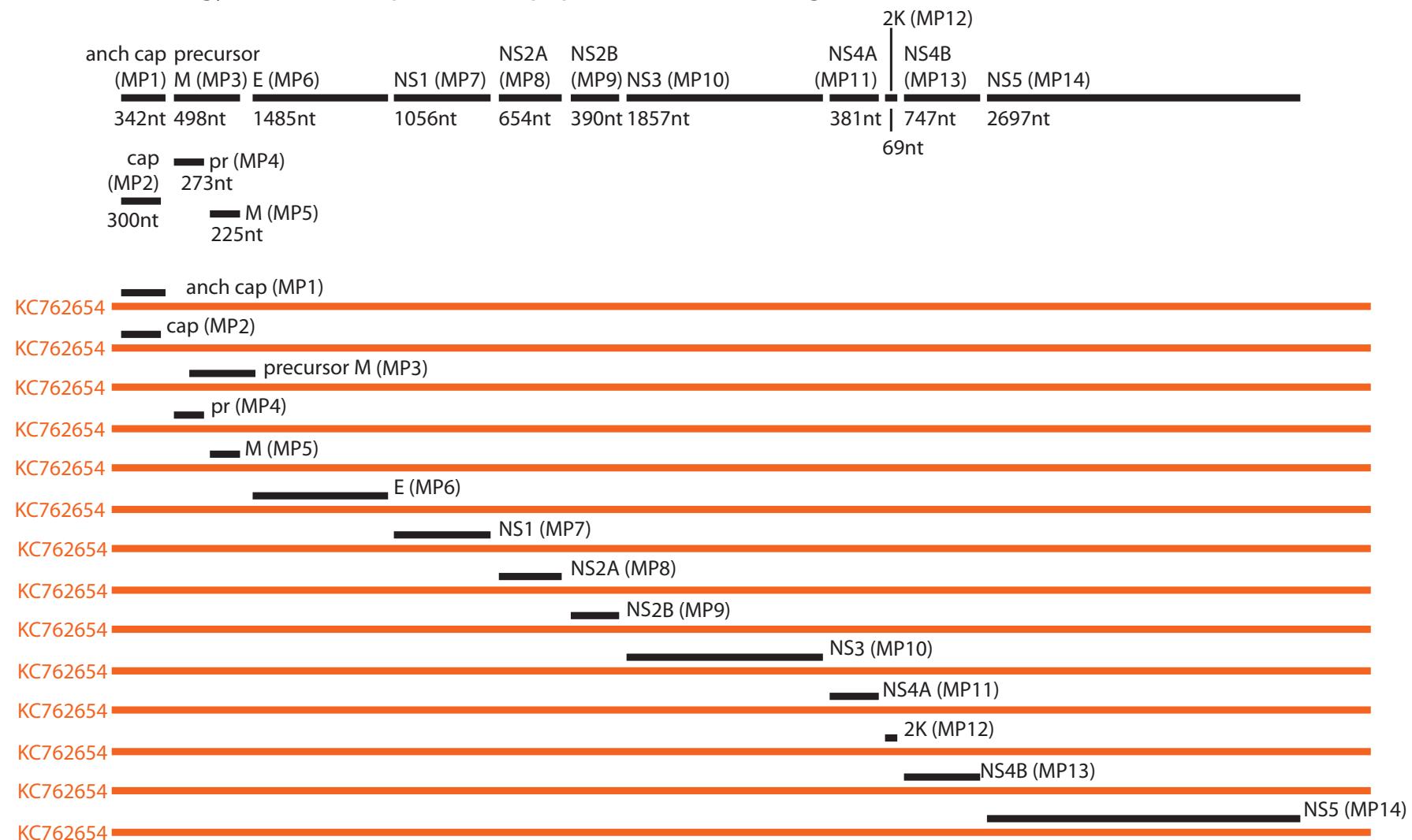
## 1. RefSeq annotation (NC\_001477)



## 2. Target sequence(s) to annotate:



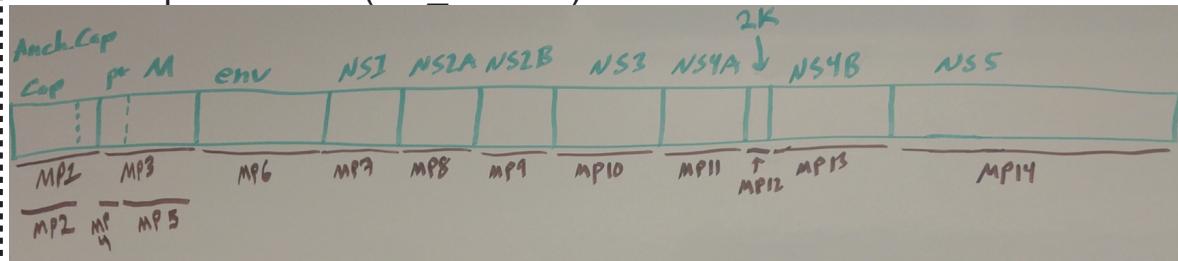
Build 14 homology models (one per mature peptide) and search targets:



# Overview of annotation pipeline for Dengue Virus

INPUT:

1. RefSeq annotation (NC\_001477)



2. Target sequence(s) to annotate:

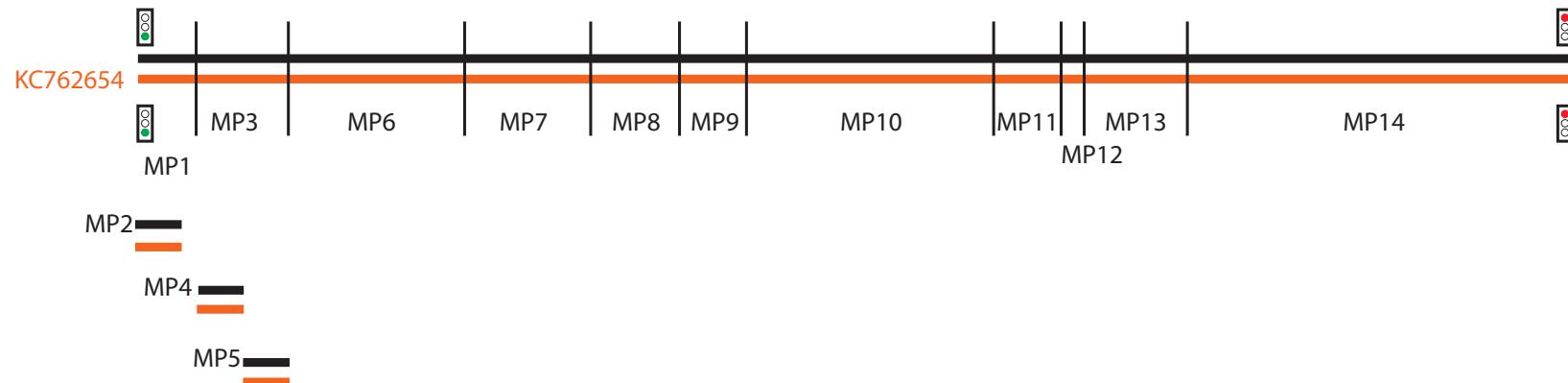
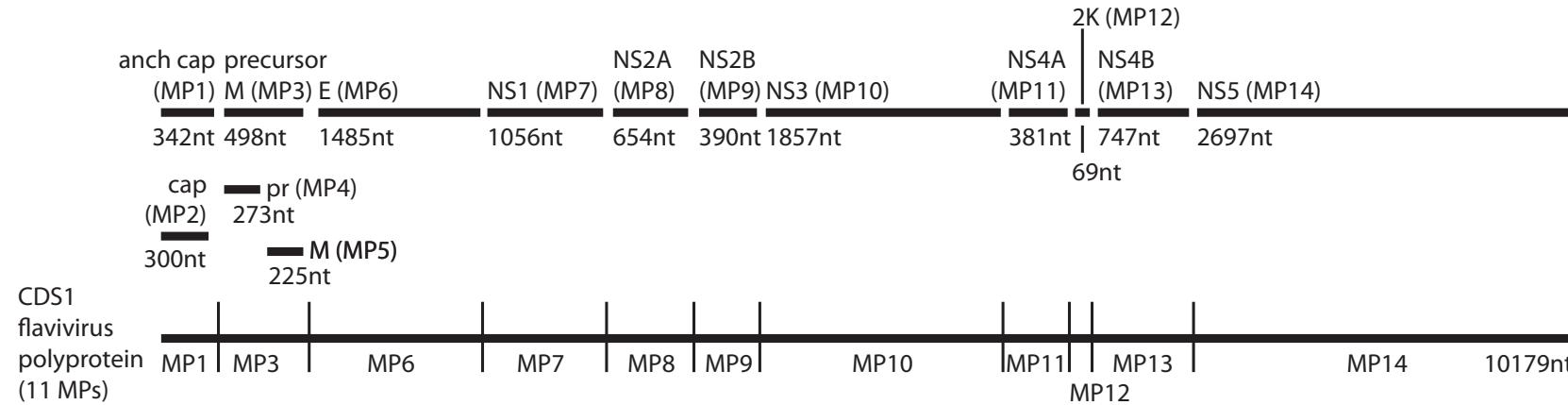
KC762654

EU179860

A75711

DQ193572

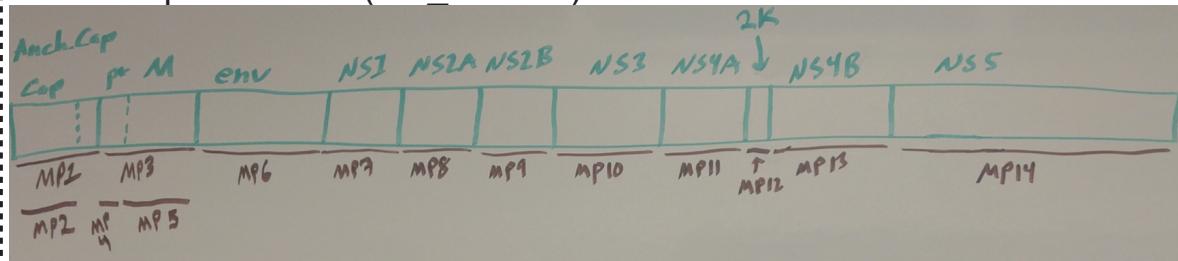
Combine models into 'features' (14 MPs and 1 CDS), translate, find stop codons, and identify errors:



# Overview of annotation pipeline for Dengue Virus

INPUT:

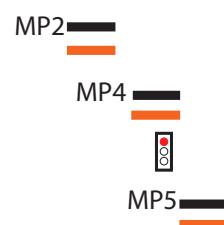
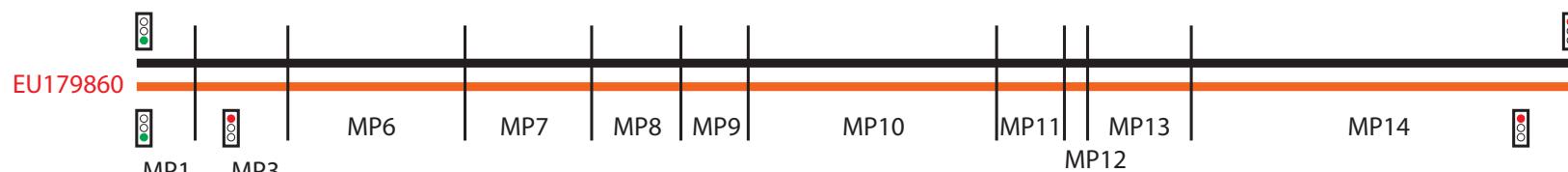
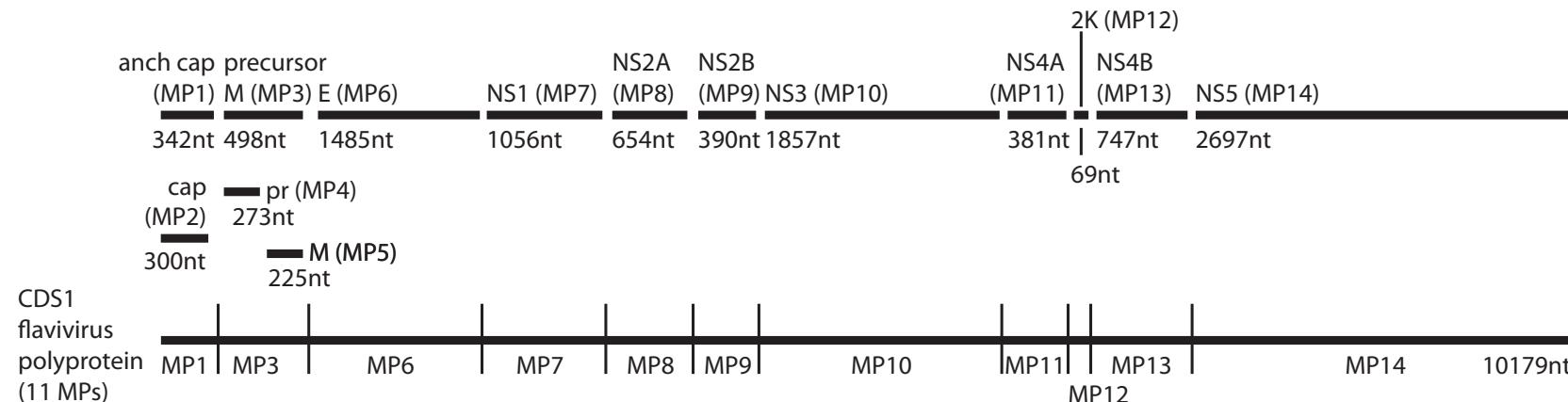
1. RefSeq annotation (NC\_001477)



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)

# Annotation statistics for pilot species and comparison to GenBank annotations

	Maize streak	Dengue (4 serologies summed)	West Nile (2 lineages summed)	Hepatitis B
# accessions	508	3769	1097	6998
# CDS/MP	3 (CDS)	14 (MP)	16 (MP)	7 (CDS)
# Exons/MP	4 (exons)	14 (MP)	16 or 18 (MP)	7 (exons)
# CDS/MP annotated in GenBank	1694 (3.34)	34371 (9.12)	8004 (7.30)	34060 (4.87)
# Exons/MP annotated in GenBank	1772 (3.48)	34371 (9.12)	8006 (7.30)	34080 (4.87)
# Exons/MP annotations that match b/t GenBank & pipeline	1651 (3.25)	32195 (8.54)	7583 (6.91)	29399 (4.20)
# Exons/MP annotations that do not match b/t GenBank & pipeline	121 (0.24)	2176 (0.58)	423 (0.39)	4681 (0.67)
# new Exons/MP from pipeline	768 (1.51)	18395 (4.88)	9546 (8.70)	14906 (2.13)
# Exons/MP annotations that match b/t Virus Variation & pipeline		41050	9822	
# Exons/MP annotations that do not match b/t Virus Variation & pipeline		8	2014	-- 2012/2014 due to same MP boundary difference in 1006 accessions

per-accession  
averages are  
in ()

Maize streak				Dengue			West Nile			Hepatitis B		
508 accessions				3772 accessions from 4 serologies			1098 accessions from 2 lineages			6998 accessions		
code	#tot	#accn	fract	#tot	#accn	fract	#tot	#accn	fract	#tot	#accn	fract
trc	179	177	0.3484	37	19	0.0050	91	43	0.0392	4430	2792	0.3990
ntr	-	-	-	60	8	0.0021	19	2	0.0018	-	-	-
ext	0	0	0.0000	6	3	0.0008	6	3	0.0027	537	422	0.0603
nst	0	0	0.0000	1	1	0.0003	0	0	0.0000	0	0	0.0000
str	0	0	0.0000	5	5	0.0013	3	1	0.0009	1725	1518	0.2169
stp	52	52	0.1024	2	2	0.0005	3	3	0.0027	407	331	0.0473
nm3	0	0	0.0000	15	9	0.0024	3	3	0.0027	353	301	0.0430
bd5	0	0	0.0000	3	2	0.0005	5	4	0.0036	119	70	0.0100
bd3	0	0	0.0000	1	1	0.0003	0	0	0.0000	0	0	0.0000
nop	0	0	0.0000	6	6	0.0016	4	2	0.0018	108	22	0.0031
olp	6	3	0.0059	35	9	0.0024	18	9	0.0082	8873	2804	0.4007
ajb	0	0	0.0000	65	23	0.0061	17	7	0.0064	1	1	0.0001
aja	0	0	0.0000	60	23	0.0061	16	7	0.0064	1	1	0.0001
aji	-	-	-	15	15	0.0040	4	4	0.0036	-	-	-
int	-	-	-	8	8	0.0021	5	2	0.0018	-	-	-
inp	-	-	-	5	5	0.0013	2	2	0.0018	-	-	-
ori	0	0	0.0000	-	-	-	-	-	-	144	144	0.0206
total	239	-	-	324	-	-	196	-	-	16698	-	-
any	178	-	0.3504	33	-	0.0087	47	-	0.0428	3632	-	0.5190
none	330	-	0.6496	3739	-	0.9913	1051	-	0.9572	3366	-	0.4810

total: total number of error codes reported

any: number of accessions with >= 1 error code reported

none: number of accessions with 0 error codes reported

Maize streak 508 accessions				Dengue 3772 accessions from 4 serologies			West Nile 1098 accessions from 2 lineages			Hepatitis B 6998 accessions			Hepatitis B 178 accessions deposited by CDC in October 2013*		
code	#tot	#accn	fract	#tot	#accn	fract	#tot	#accn	fract	#tot	#accn	fract	#tot	#accn	fract
trc	179	177	0.3484	37	19	0.0050	91	43	0.0392	4430	2792	0.3990	38	26	0.1461
ntr	-	-	-	60	8	0.0021	19	2	0.0018	-	-	-	-	-	-
ext	0	0	0.0000	6	3	0.0008	6	3	0.0027	537	422	0.0603	175	171	0.9607
nst	0	0	0.0000	1	1	0.0003	0	0	0.0000	0	0	0.0000	0	0	0.0000
str	0	0	0.0000	5	5	0.0013	3	1	0.0009	1725	1518	0.2169	185	175	0.9850
stp	52	52	0.1024	2	2	0.0005	3	3	0.0027	407	331	0.0473	179	178	1.0000
nm3	0	0	0.0000	15	9	0.0024	3	3	0.0027	353	301	0.0430	132	132	0.7416
bd5	0	0	0.0000	3	2	0.0005	5	4	0.0036	119	70	0.0100	3	3	0.0169
bd3	0	0	0.0000	1	1	0.0003	0	0	0.0000	0	0	0.0000	0	0	0.0000
nop	0	0	0.0000	6	6	0.0016	4	2	0.0018	108	22	0.0031	0	0	0.0000
olp	6	3	0.0059	35	9	0.0024	18	9	0.0082	8873	2804	0.4007	375	171	0.9607
ajb	0	0	0.0000	65	23	0.0061	17	7	0.0064	1	1	0.0001	0	0	0.0000
aja	0	0	0.0000	60	23	0.0061	16	7	0.0064	1	1	0.0001	0	0	0.0000
aji	-	-	-	15	15	0.0040	4	4	0.0036	-	-	-	-	-	-
int	-	-	-	8	8	0.0021	5	2	0.0018	-	-	-	-	-	-
inp	-	-	-	5	5	0.0013	2	2	0.0018	-	-	-	-	-	-
ori	0	0	0.0000	-	-	-	-	-	-	144	144	0.0206	0	0	0.0000
total	239	-	-	324	-	-	196	-	-	16698	-	-	1087	-	-
any	178	-	0.3504	33	-	0.0087	47	-	0.0428	3632	-	0.5190	178	-	1.000
none	330	-	0.6496	3739	-	0.9913	1051	-	0.9572	3366	-	0.4810	0	-	0.000

total: total number of error codes reported

any: number of accessions with  $\geq 1$  error code reported

none: number of accessions with 0 error codes reported

\* From "Recent population expansions of hepatitis B virus in the United States.", Ramachandran et. al, J Virol. 2014 Dec; 88(24) 13971-80.

# Output list of all insertions and deletions

# Counts of all gaps:

#

CDS#1 D1546:954 1	CDS#2 D1170:1 3	CDS#5 D190:276 1	CDS#5 I465:601 33	CDS#7 D355:198 1
CDS#1 D1600:900 1	CDS#2 D143:1028 3	CDS#5 D403:63 2	CDS#5 I465:603 1	CDS#7 D400:153 1
CDS#1 D1712:788 1	CDS#2 D199:972 1	CDS#5 D419:2 4	CDS#5 I465:607 123	CDS#7 D502:51 1
CDS#1 D1807:693 1	CDS#2 D3:1 1	CDS#5 D420:5 2	CDS#5 I465:609 2	CDS#7 D505:48 1
CDS#1 D226:2274 2	CDS#2 D619:552 1	CDS#5 D421:45 3	CDS#5 I465:625 1	CDS#7 D547:6 2
CDS#1 D556:1944 1	CDS#2 D625:546 1	CDS#5 D423:43 160	#	CDS#7 D550:2 1
CDS#1 D684:1816 3	CDS#2 D9:1 1	CDS#5 D424:3 3	CDS#6 D1:10 5	CDS#7 I453:6 1
CDS#1 I47:6 1	CDS#2 I10:3 1	CDS#5 D425:5 3	CDS#6 D1:11 1	CDS#7 I454:6 131
CDS#1 I48:6 132	CDS#2 I1170:789 3	CDS#5 D430:2 4	CDS#6 D1:13 78	CDS#7 I455:6 3
CDS#1 I49:6 3	CDS#2 I5:1 3	CDS#5 D434:3 3	CDS#6 D1:19 1	CDS#7 I549:18 1
CDS#1 I539:33 33	CDS#2 I5:2 1	CDS#5 D435:3 2	CDS#6 D1:68 3	CDS#7 I551:2 1
CDS#1 I546:1 3	CDS#2 I5:3 3	CDS#5 D437:29 4	CDS#6 D1:8 49	
CDS#1 I546:2 1	#	CDS#5 D440:26 3	CDS#6 D1:9 41	
CDS#1 I546:3 3	CDS#3 D295:552 1	CDS#5 D446:20 2	CDS#6 D48:1 1	
CDS#1 I546:33 102	CDS#3 D301:546 1	CDS#5 I463:11 3	CDS#6 I15:2 2	
CDS#1 I548:30 4	CDS#3 D64:783 1	CDS#5 I465:403 1	CDS#6 I26:7 7	
CDS#1 I549:36 2	CDS#3 D745:102 2	CDS#5 I465:434 1	CDS#6 I26:9 165	
CDS#1 I551:3 1	#	CDS#5 I465:556 1	CDS#6 I28:7 2	
CDS#1 I553:33 1	CDS#4 D130:552 1	CDS#5 I465:559 1	CDS#6 I540:6 1	
#	CDS#4 D136:546 1	CDS#5 I465:587 2	CDS#6 I541:6 132	
	CDS#4 D580:102 2	CDS#5 I465:592 2	CDS#6 I92:36 4	

Format: D<x>:<y> <z> deletion relative to RefSeq starting at reference position <x> of <y> nucleotides occurs in <z> seqs  
: I<x>:<y> <z> insertion relative to RefSeq starting after reference position <x> of <y> nucleotides occurs in <z> seqs

## Next step: Annotate “All” Flaviviruses

RefSeq accession	name	num seqs	num MPs	num CDS	avg number of failed annotations		average number of errors
					MP	CDS	
NC_001477	Dengue virus serology 1	1594	14	1	0.002	0.007	0.104
NC_001474	Dengue virus serology 2	1171	14	1	0.002	0.005	0.059
NC_001475	Dengue virus serology 3	856	14	1	0.001	0.002	0.033
NC_002640	Dengue virus serology 4	151	14	1	0	0.026	0.404
NC_009942	West Nile virus lineage 1	1027	16	3	0	0.005	0.124
NC_001563	West Nile virus lineage 2	71	15	2	0.056	0.028	1.014
NC_001564	Cell fusing agent virus	3	15	2	0	0.667	7.667
NC_002657	Classical swine fever virus	95	13	2	0.021	0.347	3.653
NC_004102	Hepatitis C virus genotype 1	778	11	2	0.030	0.675	8.104
NC_005064	Kamiti River virus	3	15	2	0	0	0
NC_008604	Culex flavivirus	26	15	2	0	0	0.231
NC_012671	Quang Binh virus	3	15	2	0	0	1.667
NC_012932	Aedes flavivirus	4	14	2	0	0	0
NC_017086	Chaoyang virus	4	15	2	0	0	0
NC_020902	Equine pegivirus	2	10	1	0	0	0
NC_021154	Rodent pegivirus	2	11	1	0	0	0
NC_024299	Nienokoue virus	2	15	2	0	0	0
NC_024805	Ilomantsi virus	2	15	2	0	0	0
NC_026797	Bovine hepacivirus	11	10	1	0	0	0
NC_027817	Parramatta River virus	2	15	2	0	0	0
NC_027819	Mercadeo virus	3	15	2	0	0	0
NC_029055	Spondweni virus	2	14	1	0	0	0
NC_030400	Nakiwogo virus	2	15	2	0	0	0
NC_030401	Hanko virus	2	15	2	0	0	0

# Automatic Classification of Flaviviruses

- Given a viral sequence:
  - Is it a flavivirus?
  - If yes: what species it is (closest RefSeq)?
  - Annotate based on closest RefSeq.
- Trial implementation (Lara Shonkwiler):
  - Correctly classifies 1100 out of 1102 flavivirus seqs

## **Future directions**

- Protein homology searches to supplement or replace nucleotide searches
- Multiple alignment based homology searches
- Given any viral sequence, find its nearest RefSeq
- Annotate structural RNA features (Infernal)

# Acknowledgements

Lara Shonkwiler

Alejandro Schäffer

David Landsman

David Lipman

J. Rodney Brister

Eneida Hatcher

Olga Blinkova

Sergey Zhdanov

Yiming Bao

# 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				
trc (truncation) in-frame stop codon exists 5' of predicted stop				

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	

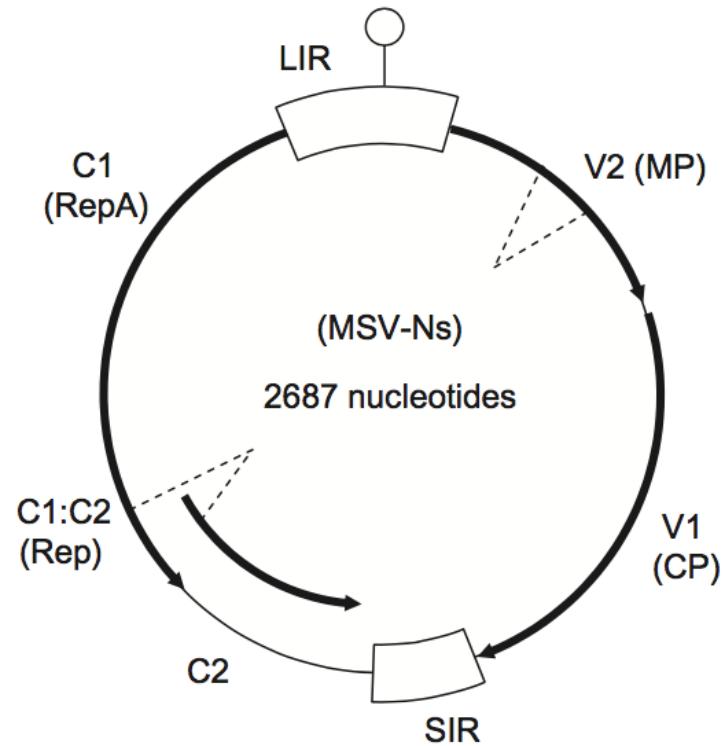
Example output from annotation script:

```
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)]
```

# Maize Streak Virus

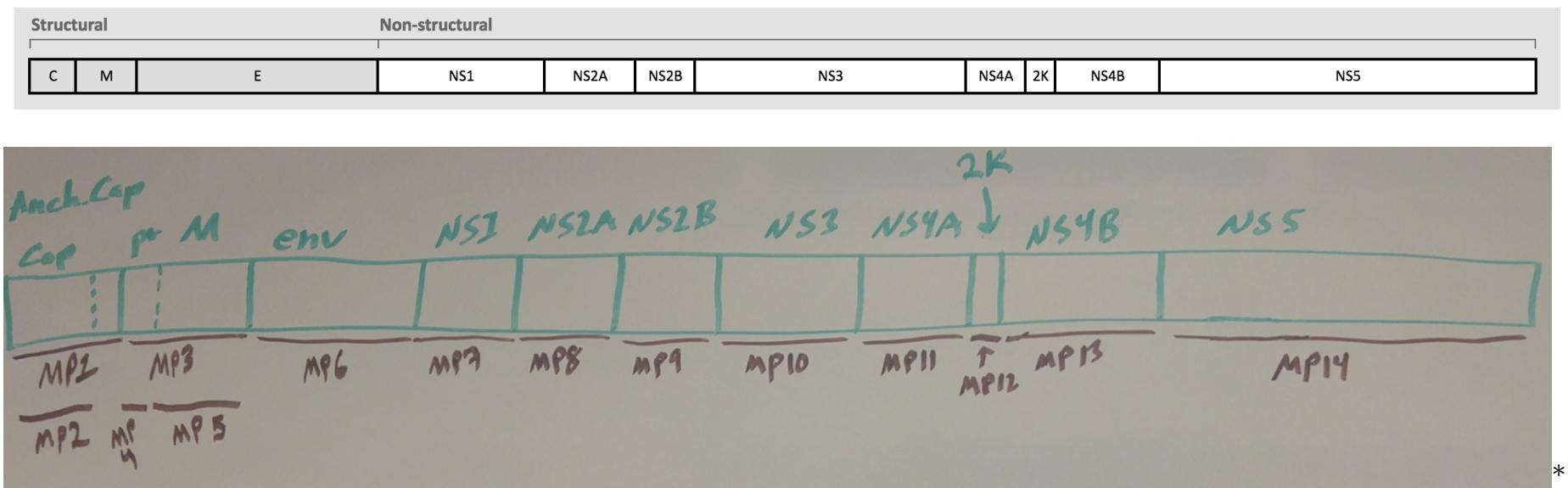
- cause of most serious crop disease in Africa, responsible for between USD\$100-\$500 million of crop loss per year
- NC\_001346 (2689 nt):
  - 4 CDS (V1, V2, C1, C1:C2), 5 total exons
  - both strands have  $\geq 1$  CDS
  - one CDS (C1:C2) has two segments (exons)
  - origin sequence: TAATATT|AC
- closed genome
- 508 full genome sequences in GenBank
  - average number of CDS annotated: 3.34



Shepherd et al., Molecular Plant Pathology (2010) 11(1), 112;  
PMID: 20078771

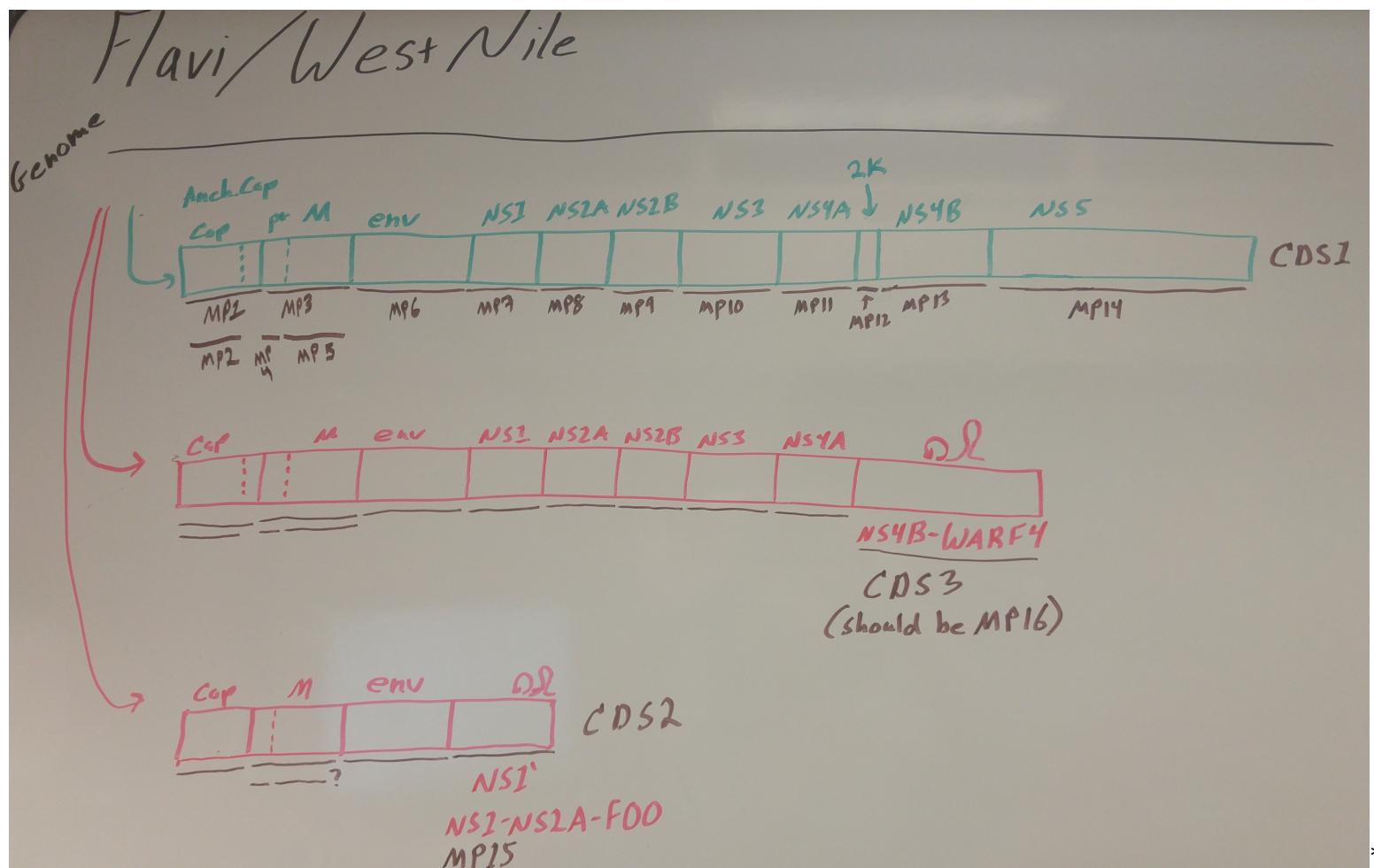
# Dengue virus

- up to 400 million people infected per year worldwide (CDC)
- a single CDS encodes the flavivirus polyprotein gene
- the polyprotein is cleaved in two rounds into 14 smaller 'mature peptides'
- annotations exist in Virus Variation to compare against our own



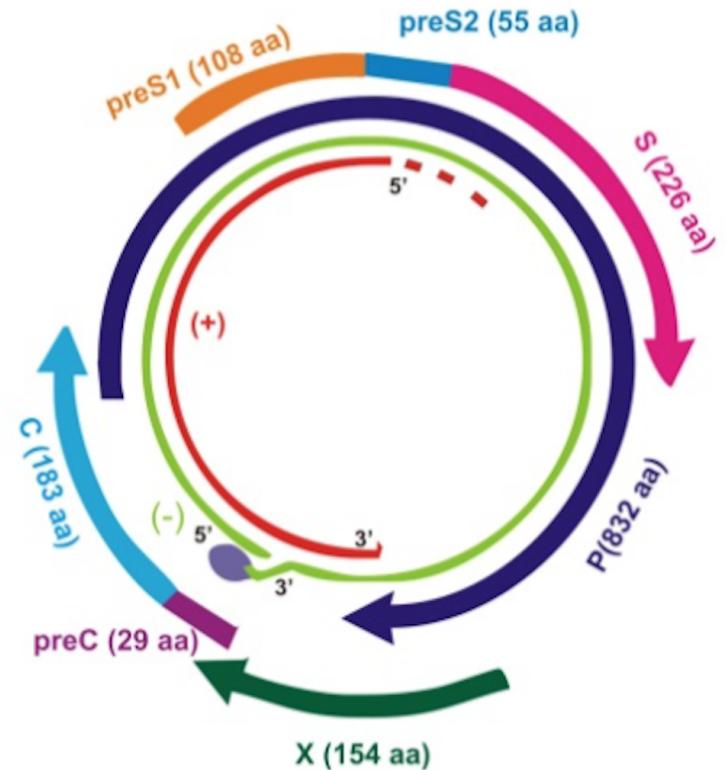
# West Nile virus

- another flavivirus, similar to Dengue
- 3 CDS, including:
  - two CDS that utilize ribosomal frameshift to change reading frame, possibly complicating annotation
  - annotations exist in Virus Variation to compare against our own



# Hepatitis B Virus

- causes a serious liver infection in more than 200,000 people in the US per year
- NC\_003977 (3182 nt):
  - 7 single exon CDS, all on positive strand
- closed genome
- 7061 full genome sequences in GenBank
  - average number of CDS annotated: 4.87



Jayalakshmi et. al, Hepatitis B virus genetic diversity: disease pathogenesis. INTECH Open Access Publisher, 2013.

# 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)]
```

# ajb and aja error codes indicates lack of an expected adjacent feature

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors	 	 	 	 
ajb (adjacent before) mature peptide is not adjacent to expected mature peptide on 5' end				

Example output from annotation script:

KJ501413

5 MP#5

ajb mature peptide is not adjacent to same set of  
mature peptides before it as in reference [-(5.1,4.1)]

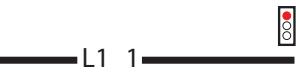
# Unexpected stop codon errors (trc,ntr)

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors				
trc (truncation) in-frame stop codon exists 5' of predicted stop				
ntr (not translated) mature peptide is not translated due to early stop 5' of predicted start	N/A	N/A		N/A

# Unexpected stop codon errors (ext,nst)

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors	 	 	 	 
ext (extension) first in-frame stop codon exists 3' of predicted stop	 	 	 	 
nst (no stop) no in-frame stop codon exists 3' of predicted valid start codon	 	 	 	 

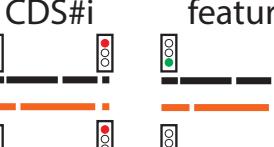
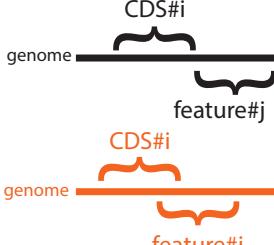
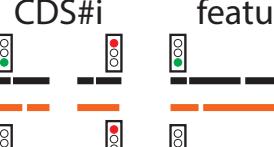
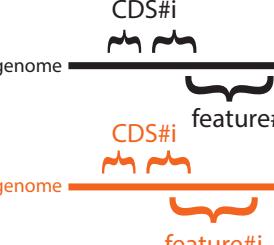
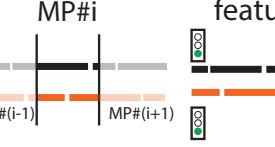
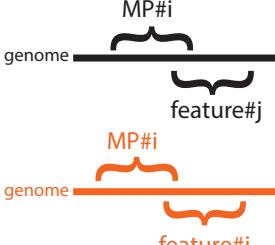
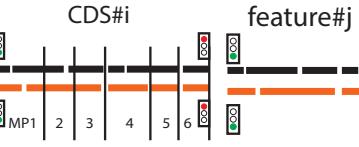
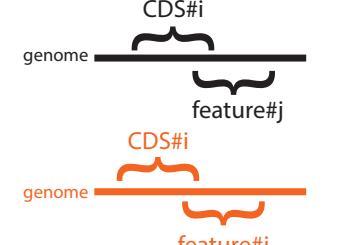
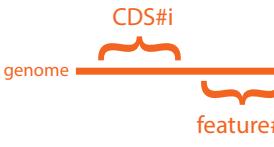
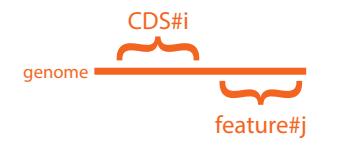
# Missing expected features (str,stp,nm3)

	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)	 MP1 2 3 4 5 6
str (start) predicted CDS start is not a valid start codon			N/A	 MP1 2 3 4 5 6
stp (stop) predicted CDS stop is not a valid stop codon			N/A	 MP1 2 3 4 5 6
nm3 (not a multiple of 3) length is not a multiple of 3	 L % 3 = 0 L' % 3 != 0	 (L1 + L2) % 3 = 0 (L1' + L2') % 3 != 0	 mp #(i-1) mp #i mp #(i+1) L % 3 = 0 L' % 3 != 0	 mp1 2 3 4 5 6 L % 3 = 0 L' % 3 != 0

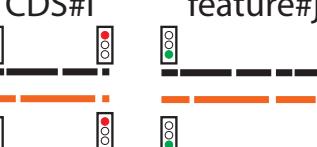
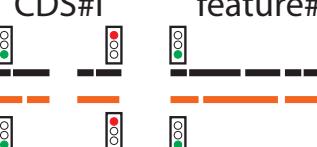
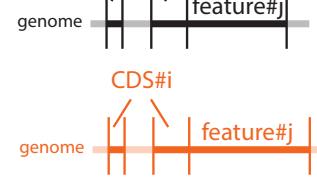
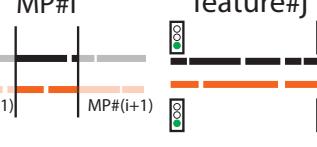
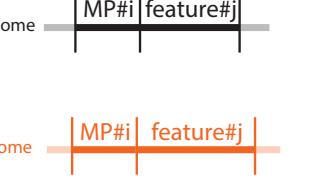
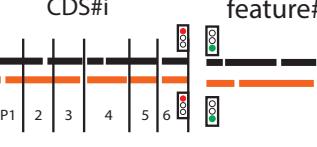
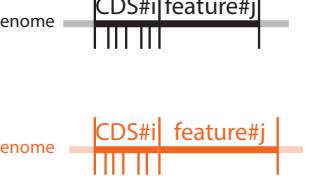
# Problem with homology search prediction (bd5,bd3,nop)

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors				
bd5 (5' boundary) alignment does not extend to 5' boundary of reference				N/A
bd3 (3' boundary) alignment does not extend to 3' boundary of reference				N/A
nop (no prediction) homology search yielded no prediction				N/A

# Unexpected relationship to other features (olp)

	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				

# Unexpected relationship to other features (ajb)

	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors	 	 	 	 
ajb (adjacent before)	 <p>genome   CDS#i   feature#j  </p>	 <p>genome   CDS#i   feature#j  </p>	 <p>genome   MP#i   feature#j  </p>	 <p>genome   CDS#i   feature#j  </p>
mature peptide is not adjacent to expected mature peptide on 5' end				

# Unexpected relationship to other features (aja)

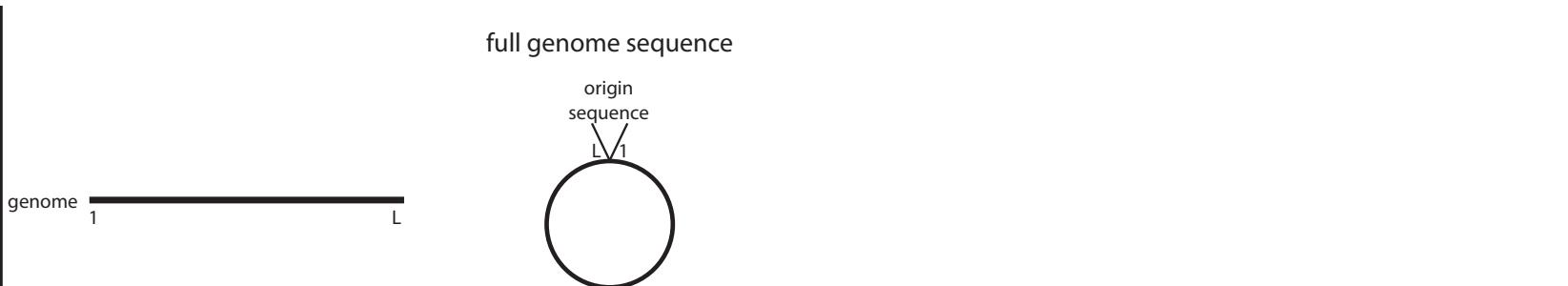
	CDS (single exon)	CDS (multi-exon)	mature peptide (MP #i)	CDS (made up of mature peptides)
No errors	   	   	   	   
aja (adjacent after) mature peptide is not adjacent to expected mature peptide on 3' end	   	   	   	   

# Problem annotating CDS due to mature peptide errors (aji,int,inp)

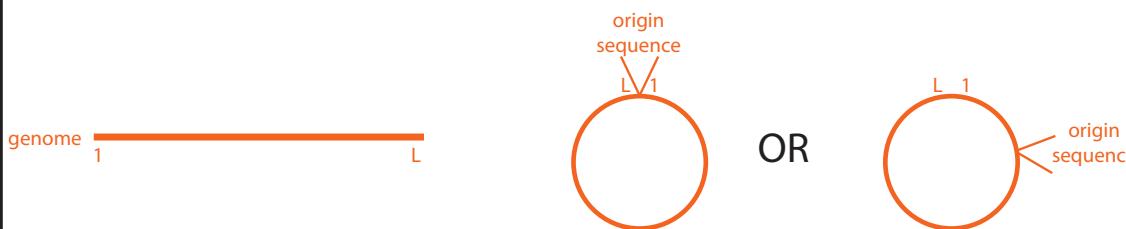
	CDS (single exon)	CDS (multi-exon)	mature peptide (MP#i)	CDS (made up of mature peptides)
No errors				
aji (adjacency inconsistency) CDS has at least one inconsistent adjacency relative to reference	N/A	N/A	N/A	
int (interrupted translation) CDS has at least one mature peptide that is not translated due to early stop (trc)	N/A	N/A	N/A	
inp (interrupted prediction) CDS has at least one mature peptide that is not translated due to lack of prediction (nop)	N/A	N/A	N/A	

# Lack of exactly one origin sequence (ori)

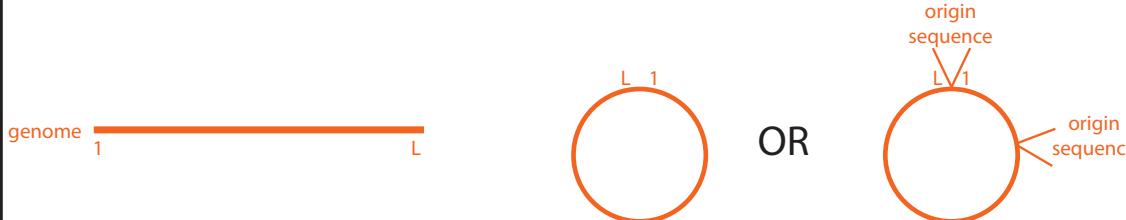
Reference



No errors

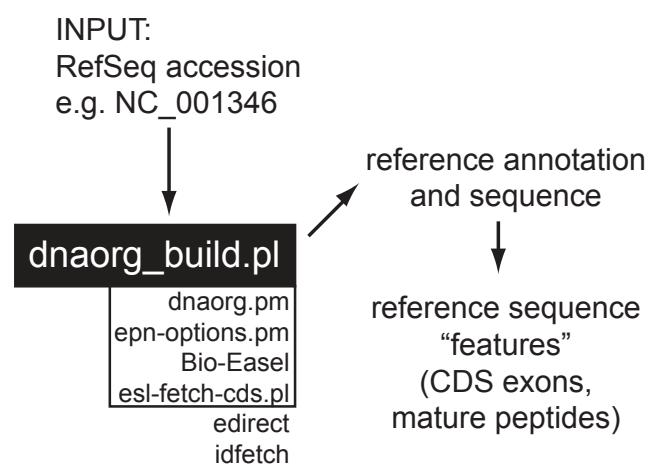


ori(origin)  
sequence does not  
contain exactly  
1 occurrence of the  
origin sequence



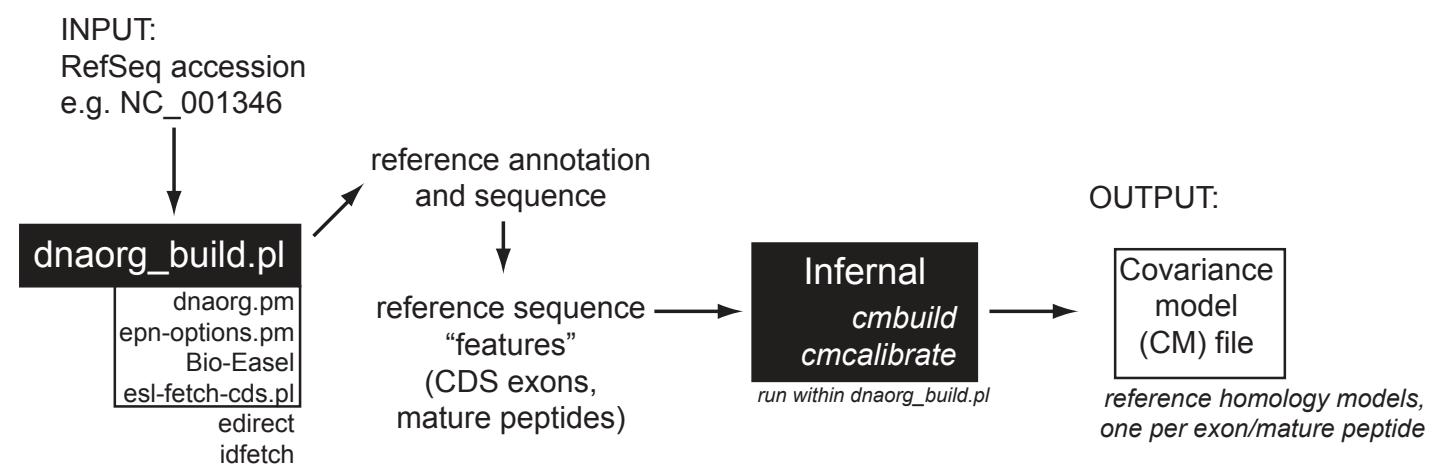
## dnaorg\_build.pl:

Build homology models of reference genome features.  
Run once per reference genome. (Repeat if reference is updated.)



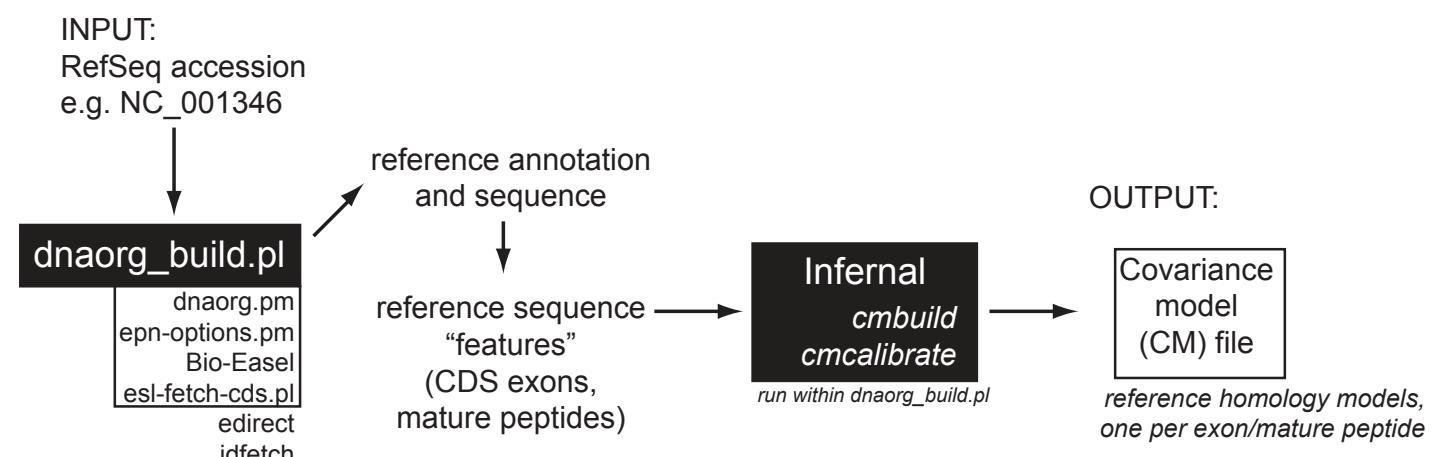
## **dnaorg\_build.pl:**

Build homology models of reference genome features.  
Run once per reference genome. (Repeat if reference is updated.)



## **dnaorg\_build.pl:**

Build homology models of reference genome features.  
Run once per reference genome. (Repeat if reference is updated.)



## **dnaorg\_annotate.pl:**

Annotate non-reference genomes.  
Repeat as necessary.

INPUT:

Covariance model (CM) file

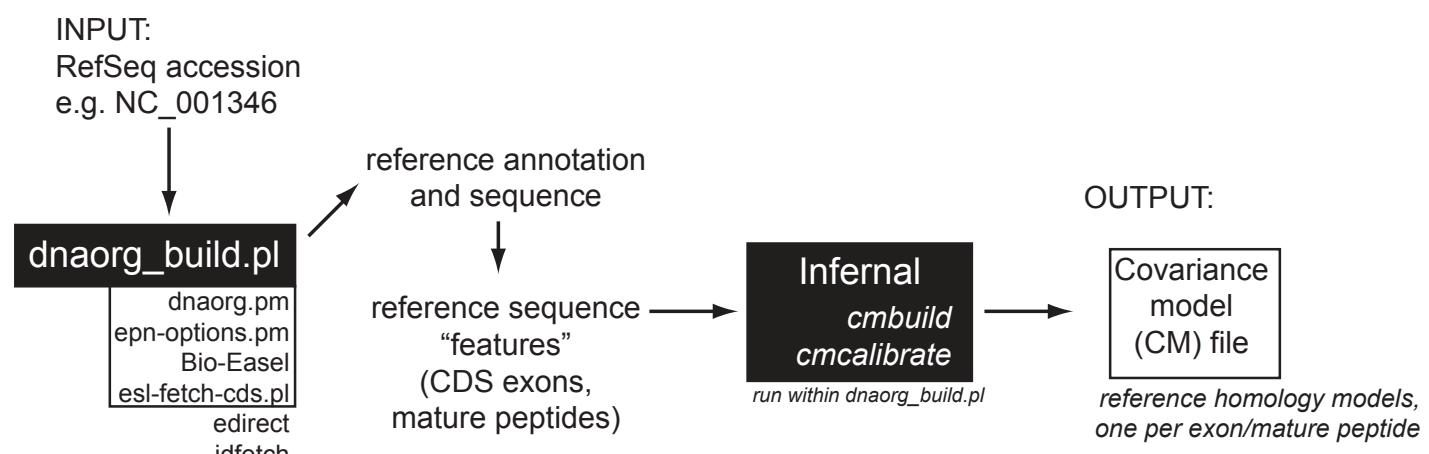
file with list of accessions (RefSeq listed first)

dnaorg\_annotate.pl

dnaorg.pm  
epn-options.pm  
Bio-Easel  
esl-fetch-cds.pl  
esl-epn-translate.pl  
esl-ssplit.pl  
edirect  
idfatch

## dnaorg\_build.pl:

Build homology models of reference genome features.  
Run once per reference genome. (Repeat if reference is updated.)



## dnaorg\_annotate.pl:

Annotate non-reference genomes.  
Repeat as necessary.

INPUT:

Covariance model (CM) file

file with list of accessions (RefSeq listed first)

**Step 1. Fetch all sequences and verify CMs are current with RefSeq annotation.**

dnaorg\_annotate.pl

dnaorg.pm  
epn-options.pm  
Bio-Easel  
esl-fetch-cds.pl  
esl-epn-translate.pl  
esl-ssplit.pl  
edirect  
idfatch

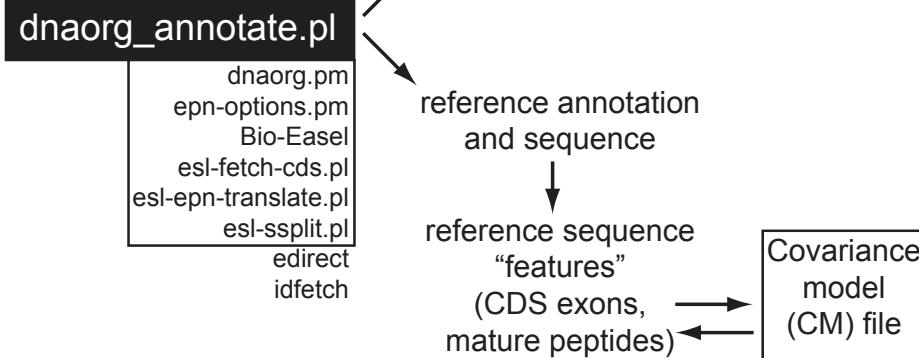
non-reference sequences

reference annotation and sequence

reference sequence "features" (CDS exons, mature peptides)

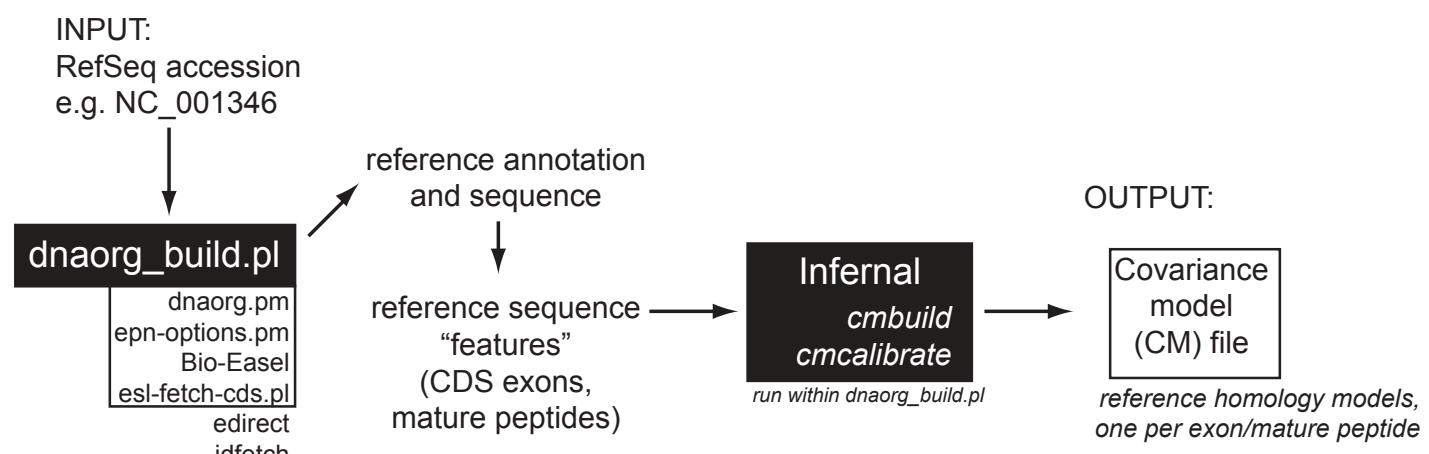
Covariance model (CM) file

consistent?



## dnaorg\_build.pl:

Build homology models of reference genome features.  
Run once per reference genome. (Repeat if reference is updated.)



## dnaorg\_annotation.pl:

Annotate non-reference genomes.

Repeat as necessary.

### INPUT:

Covariance model (CM) file

file with list of accessions (RefSeq listed first)

**Step 1. Fetch all sequences and verify CMs are current with RefSeq annotation.**

**Step 2. Perform homology searches.**

### dnaorg\_annotation.pl

dnaorg.pm  
epn-options.pm  
Bio-Easel  
esl-fetch-cds.pl  
esl-epn-translate.pl  
esl-ssplit.pl  
edirect  
idfetch

non-reference sequences

homology-based annotations

reference annotation and sequence

reference sequence "features" (CDS exons, mature peptides)

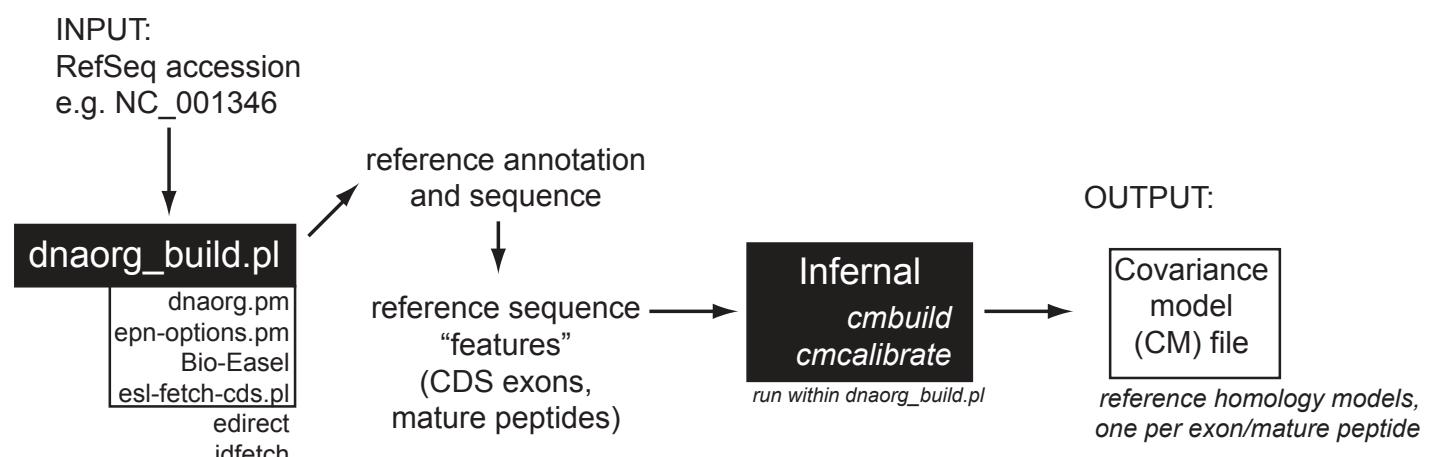
Covariance model (CM) file

consistent?

reference sequence "features" (CDS exons, mature peptides)

## dnaorg\_build.pl:

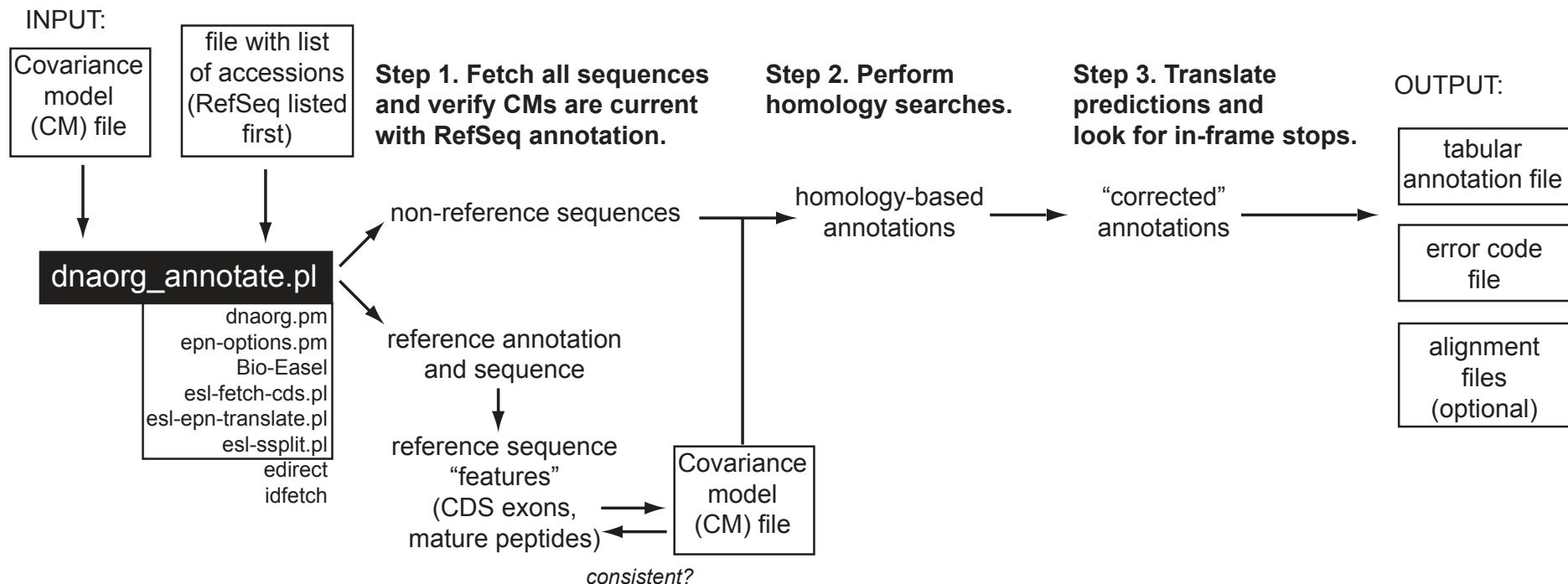
Build homology models of reference genome features.  
Run once per reference genome. (Repeat if reference is updated.)



## dnaorg\_annotate.pl:

Annotate non-reference genomes.

Repeat as necessary.



# Annotating Maize Streak Virus: dnaorg\_build.pl step

```
<[(example-runs)]> dnaorg_build.pl -c NC_001346
# dnaorg_build.pl :: build homology models for features of a reference sequence
# dnaorg 0.1 (Apr 2016)
#
# date:    Mon Apr 25 15:14:54 2016
#
# reference accession:  NC_001346
# genome is circular:   yes [-c]
#
# Gathering information on reference using edirect      ... done. [20.9 seconds]
# Fetching and processing the reference genome         ... done. [3.5 seconds]
# Building models and submitting calibration jobs to the farm ... done. [33.7 seconds]
#
# When the 5 cmcalibrate jobs on the farm finish, you can use dnaorg_annotate.pl
# to use them to annotate genomes.
#
#
# Output printed to screen saved in:
# List of executed commands saved in:
# List and description of all output files saved in:
# CM file #1, CDS#1 (currently calibrating on the farm) saved in:
# CM file #2, CDS#2 (currently calibrating on the farm) saved in:
# CM file #3, CDS#3.1 (currently calibrating on the farm) saved in:
# CM file #4, CDS#3.2 (currently calibrating on the farm) saved in:
# CM file #5, CDS#4 (currently calibrating on the farm) saved in:
# Shell script to submit cmcalibrate commands with (already executed, jobs submitted to farm) saved in:
#
# All output files created in directory ./NC_001346/
#
# CPU time: 00:00:58.13
#           hh:mm:ss
#
# DNAORG-SUCCESS
```

NC_001346.dnaorg_build.log
NC_001346.dnaorg_build.cmd
NC_001346.dnaorg_build.list
NC_001346.dnaorg_build.0.cm
NC_001346.dnaorg_build.1.cm
NC_001346.dnaorg_build.2.cm
NC_001346.dnaorg_build.3.cm
NC_001346.dnaorg_build.4.cm
NC_001346.dnaorg_build.ref.cm.qsub

# Annotating Maize Streak Virus: dnaorg\_annotate.pl step

```
<[example-runs]> dnaorg_annotate.pl -c --origin TAATATT\IAC NC_001346.ntlist
# dnaorg_annotate.pl :: annotate sequences based on a reference annotation
# dnaorg 0.1 (Apr 2016)
# -----
# date: Mon Apr 25 15:44:14 2016
#
# file with list of accessions:      NC_001346.ntlist
# genome is closed (a.k.a. circular): yes [-c]
# identify origin seq <s> in genomes: TAATATT\IAC [--origin]
# -----
# Gathering information on 508 sequences using edirect          ... done. [86.7 seconds]
# Fetching all sequences and processing the reference genome ... done. [78.7 seconds]
# Creating CM database by concatenating individual CM files   ... done. [0.5 seconds]
# Preparing the CM database for homology search using cmpress ... done. [0.2 seconds]
# Verifying CM database created for current reference NC_001346 ... done. [0.0 seconds]
# Submitting 425 cmscan jobs to the farm                      ... done. [82.9 seconds]
# Waiting a maximum of 500 minutes for all farm jobs to finish ... done. [60.8 seconds]
# Parsing cmscan results                                      ... done. [0.1 seconds]
# Calculating predicted feature lengths                      ... done. [0.0 seconds]
# Fetching cmscan predicted hits into fasta files           ... done. [0.1 seconds]
# Combining predicted exons into CDS                        ... done. [0.1 seconds]
# Combining predicted mature peptides into CDS              ... done. [0.0 seconds]
# Identifying internal starts/stops in coding sequences     ... done. [10.5 seconds]
# Correcting homology search stop codon predictions to account for observed stop codons ... done. [0.0 seconds]
# Identifying overlap and adjacency errors                  ... done. [0.1 seconds]
# Finalizing annotations and validating error combinations ... done. [0.5 seconds]
# Fetching corrected matches into fasta files               ... done. [0.1 seconds]
# Combining corrected exons into CDS                      ... done. [0.9 seconds]
# Combining corrected mature peptides into CDS            ... done. [0.0 seconds]
# Translating corrected nucleotide features into protein sequences ... done. [7.7 seconds]
# Generating error code output                            ... done. [0.0 seconds]
# Generating tabular annotation output                   ... done. [0.1 seconds]
#
```

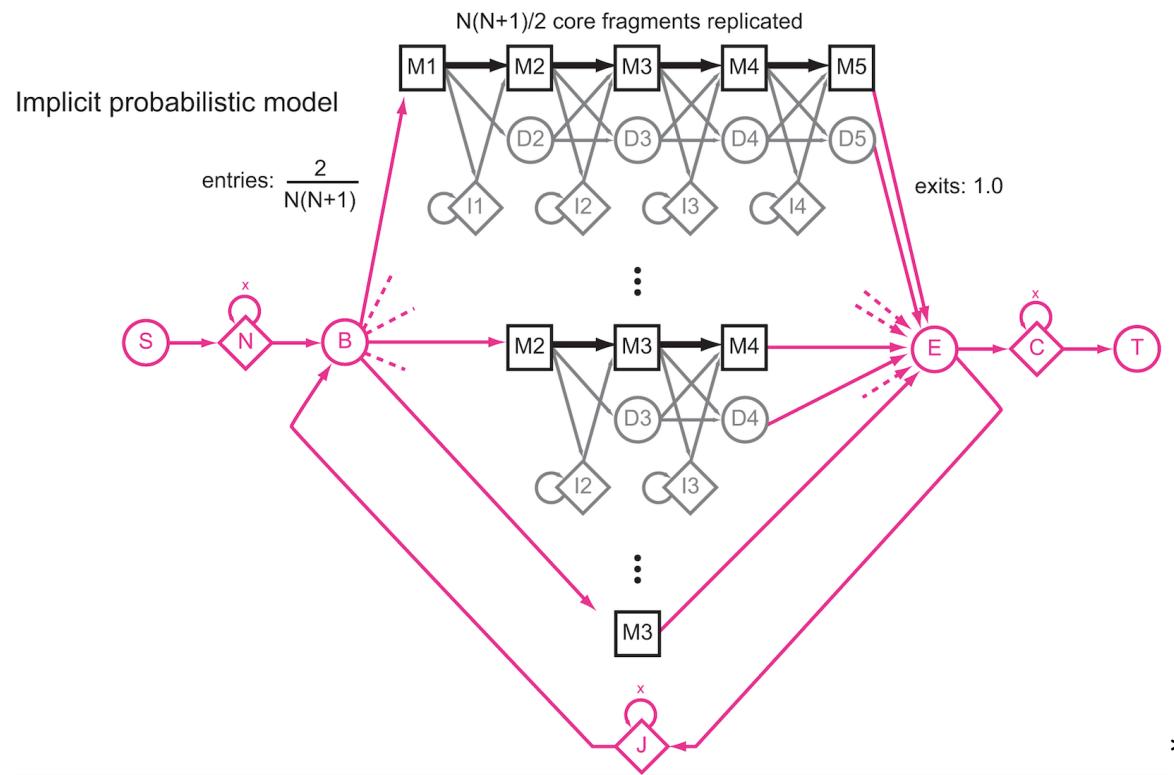
# Annotating Maize Streak Virus: dnaorg\_annotate.pl step

```
#  
# Output printed to screen saved in:  
# List of executed commands saved in:  
# List and description of all output files saved in:  
# CM file (a concatenation of individual files created by dnaorg_build.pl) saved in:  
# All annotations in tabular format saved in:  
# Summary of all annotations saved in:  
# Annotations for all sequences with >= 1 failure in tabular format saved in:  
# Annotations for all sequences with >= 1 error in tabular format saved in:  
# List of errors, one line per sequence saved in:  
# List of errors, one line per error saved in:  
# Summary of all errors saved in:  
#  
# All output files created in directory ./NC_001346/  
#  
# CPU time: 00:05:34.66  
#           hh:mm:ss  
#  
# DNAORG-SUCCESS
```

NC_001346.dnaorg_annotate.log
NC_001346.dnaorg_annotate.cmd
NC_001346.dnaorg_annotate.list
NC_001346.dnaorg_annotate.ref.cm
NC_001346.dnaorg_annotate.tbl
NC_001346.dnaorg_annotate.tbl.summary
NC_001346.dnaorg_annotate.fail.tbl
NC_001346.dnaorg_annotate.error.tbl
NC_001346.dnaorg_annotate.peraccn.errors
NC_001346.dnaorg_annotate.all.errors
NC_001346.dnaorg_annotate.errors.summary

# Infernal was chosen for homology search tool over HMMER because it encourages global alignment

- HMMER3 alignments are local and often trim ends due to its probabilistic model



Infernal encourages full length alignment by setting probability of global alignment to 0.5 (not shown).

# Annotation Review Process

1. Randomly select sequences with and without errors from the pipeline output OR list of differences between annotation output and Virus Variation Resource database

```
160 HM488171 16 MP#16 trc in-frame stop codon exists 5' of stop position predicted by homology to reference [homolog
161 HM488171 19 CDS(MP)#3 trc in-frame stop codon exists 5' of stop position predicted by homology to reference [homolog
162 HQ671730 15 MP#15 ext first in-frame stop codon exists 3' of stop position predicted by homology to reference [h
163 HQ671730 18 CDS(MP)#2 stp predicted CDS stop position is not end of valid stop codon (TAG|TAA|TGA) [CAG ending at p
164 HQ671730 18 CDS(MP)#2 ext first in-frame stop codon exists 3' of stop position predicted by homology to reference [h
165 JX503098 14 MP#14 trc in-frame stop codon exists 5' of stop position predicted by homology to reference [homolog
```

OR

```
1 FJ461303: NS2B 4108-4500 (file) None-None (DB)
2 FJ461303: NS3 4501-6357 (file) None-None (DB)
3 KF955440: NS5 7552-9873 (file) None-None (DB)
4 HQ166035: NS3 4465-4572 (file) 4464-6320 (DB)
5 KF955446: NS4A 6341-6720 (file) 6341-6721 (DB)
6 JQ922552: NS4B 6799-7539 (file) 6799-7542 (DB)
```

2. Use alignments of the RefSeq(s) and the selected sequences to verify or correct each error, and identify errors that may have been missed.

NC\_009942 MP#16 (NS1')



HM488171 MP#16 trc in-frame stop codon exists 5' of stop position predicted by homology to reference [homology search predicted 3553..3681 revised to 3553..3603 (stop shifted 78 nt)]

3. Rodney's group provides Alejandro and Eric with a list suggesting improvements that should be made in the code.

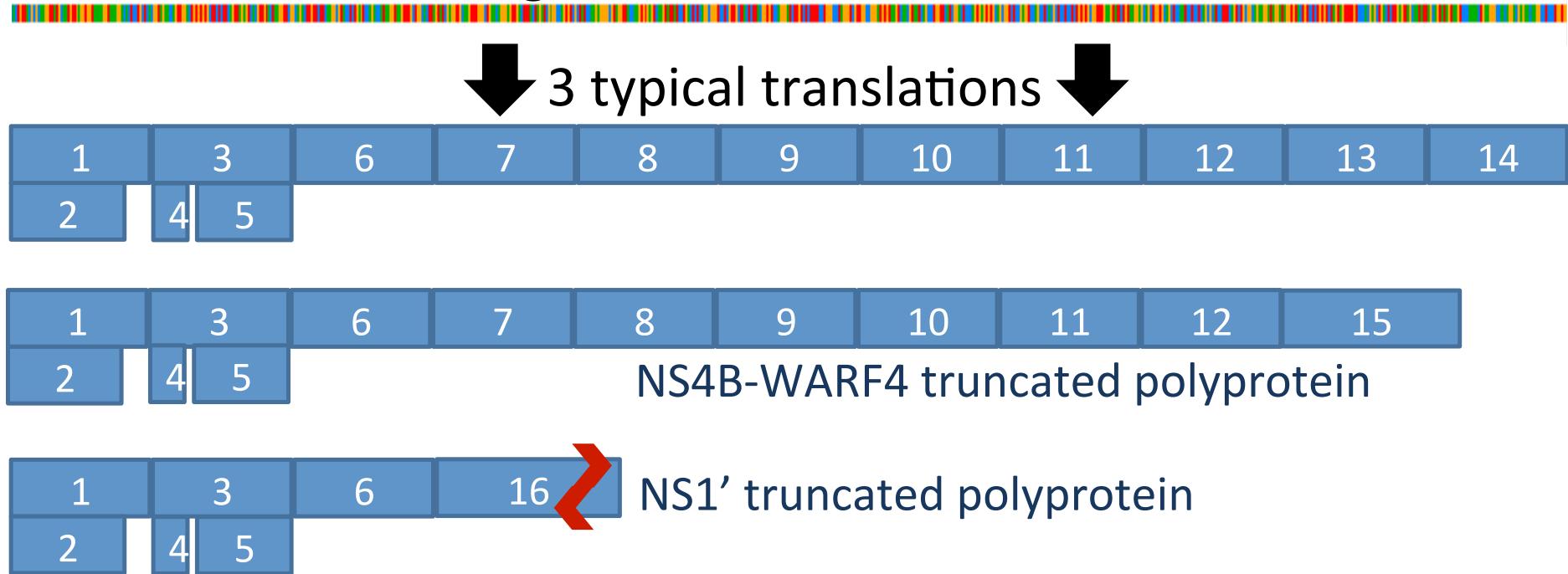
# Identifying Rare Errors

West Nile virus, lineage 1



# Identifying Rare Errors

West Nile virus, lineage 1



HM488171 errors:

#16 truncated, correct error, leads to

#15 not translated due to an earlier truncation, incorrect error

Code was edited, removing the incorrect error

# Identifying Rare Errors

## West Nile virus, lineage 1



### JX503096 errors:

- #1 olp feature does not overlap with same set of features as in reference [-(1.1,2.1)]
- #2 olp feature does not overlap with same set of features as in reference [-(2.1,1.1)]
- #2 trc in-frame stop codon exists 5' of stop position predicted by homology to reference [homology search predicted 228..1 revised to 228..175 (stop shifted 174 nt)]
- CDS str predicted CDS start position is not beginning of ATG start codon [ATT starting at position 51 on strand +]



### Poor alignment:

DQ211652 ►□ CCCGGCTTCTCTTGATTCAGCTCAACAGGGCTATCTTGAGCTCTGATCAGCGCCARGGGCCAAATACGATTCTGTGTTGGCTCTCTTGGCTTCTHTCAGGTTCACAGCAATTGCTCCGACCCGAGCAGTGCCTGGATCGATGGAGAGCTGT  
JX503096 ►□ CCCHHHHHCTTCCTCTGATTCAGCTCAACAGGGCTATCTTGAGCTCTGATCAGCGCCARGGGCCAAATACGHHHHHHHHHNGCCTAAACAGGHHHHHHHHHNGCCTGGCGCTTCTACGATATTTCACACCAATTGCTCCGACCCGAGCAGTGCCTGGATCGATGGAGAGCTGT



Code searches for homology in forward and reverse directions, and found that JX503096 has a reverse-complement section

