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*

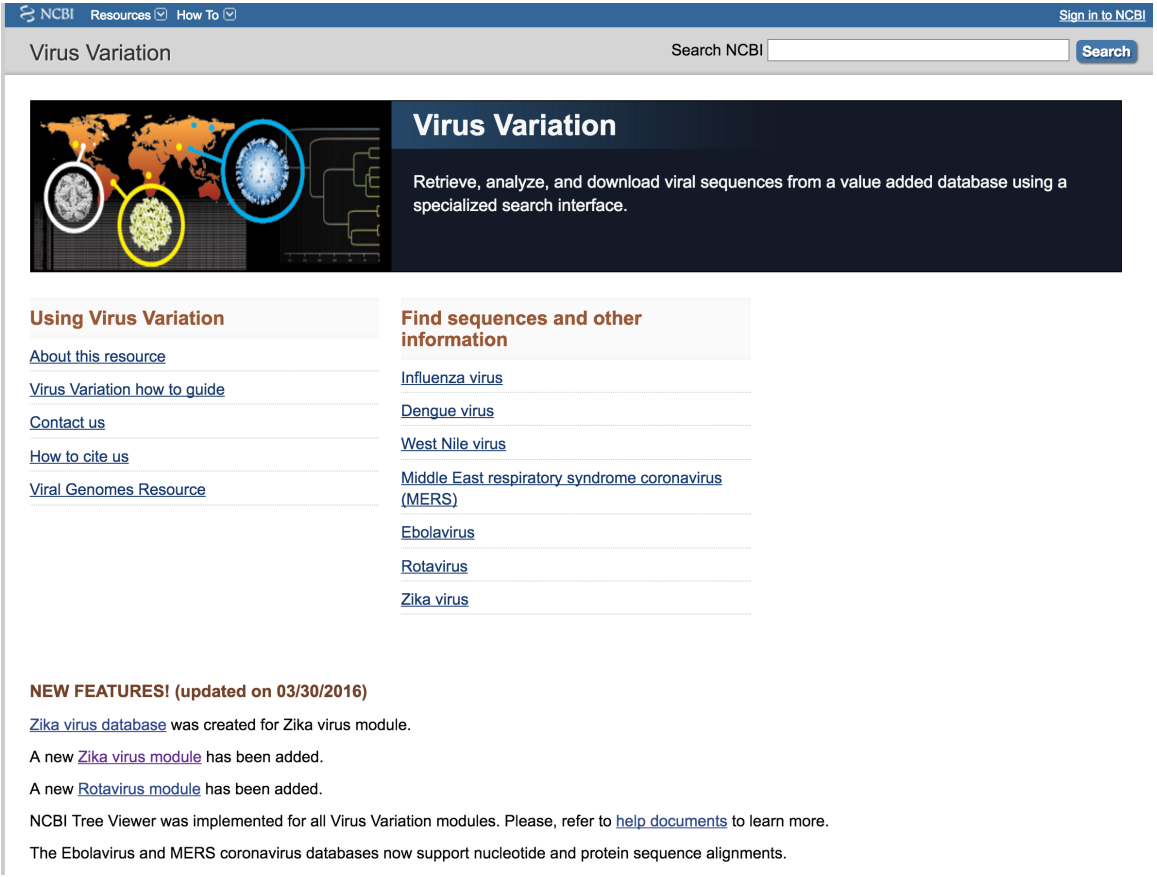
rank	species	#seqs	family	type	host	#cds	#mature peptides
1	Hepatitis B	7061	Hepadnaviridae	dsDNA-RT	humans	7	-
2	Dengue	3875	Flaviviridae	(+)ssRNA	humans	1	14
3	Rotavirus	22536†	Reoviridae	dsRNA	humans	12	-
4	HIV-1	2132	Retroviridae	ssRNA-RT	humans	10	14
5	Porcine circovirus	1559	Circoviridae	ssDNA	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	Picarnoviridae	(+)ssRNA	humans	1	11
10	RSV	623	Paramyxoviridae	(-)ssRNA	humans	11	-
11	Enterovirus C (Polio)	618	Picarnoviridae	(+)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	Calciviridae	(+)ssRNA	humans	3	6

† sum of 11 segments

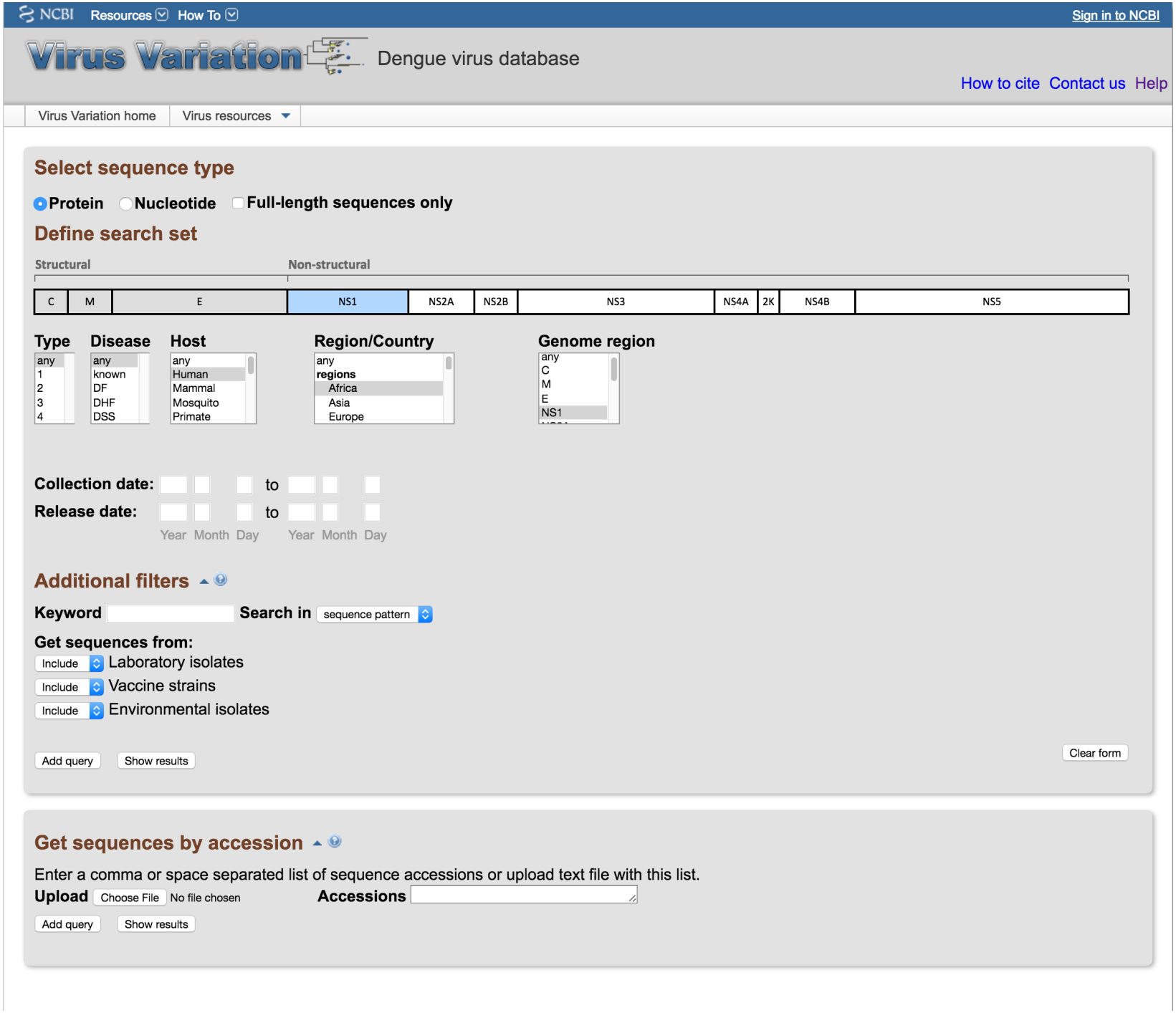
*Influenza is omitted. Current as of April 18, 2016.

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 Virus Variation Resource is a powerful tool for viral research



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...

Outline of talk

1. Selection of four pilot viral species

2. Overview of annotation pipeline and explanation of "error codes"

3. Statistics on pilot phase annotations

4. Implementation of annotation pipeline

5. Interaction with J. Rodney Brister's Virus Variation group

6. Future directions

Outline of talk

1. Selection of four pilot viral species
2. Overview of annotation pipeline and explanation of "error codes"
3. Statistics on pilot phase annotations
4. Implementation of annotation pipeline
5. Interaction with J. Rodney Brister's Virus Variation group
6. Future directions

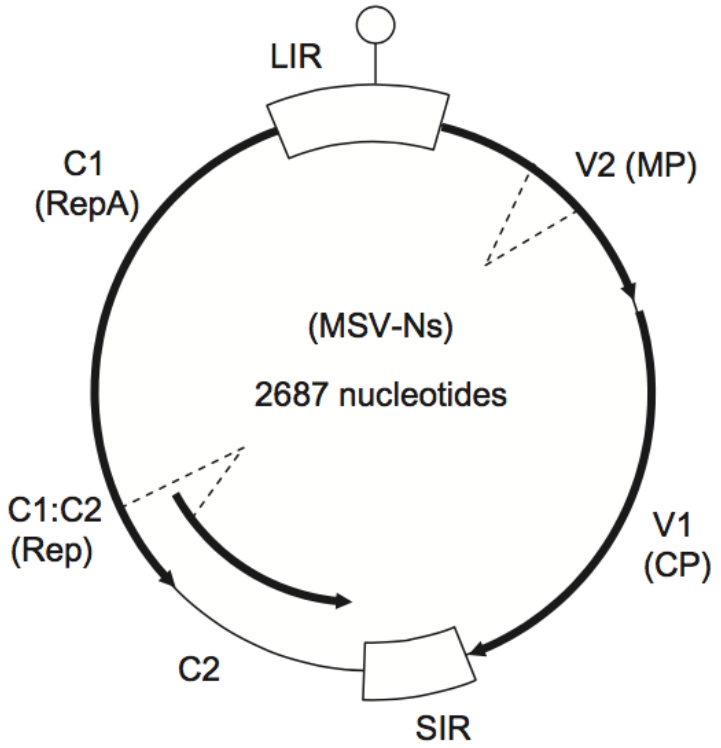
Four pilot species were chosen from the 15 most prevalent

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	Picarnoviridae	(+)ssRNA	humans	1	11
10	RSV	623	Paramyxoviridae	(-)ssRNA	humans	11	-
11	Enterovirus C (Polio)	618	Picarnoviridae	(+)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	Calciviridae	(+)ssRNA	humans	3	6

$*$ sum of 11 segments

Maize Streak Virus

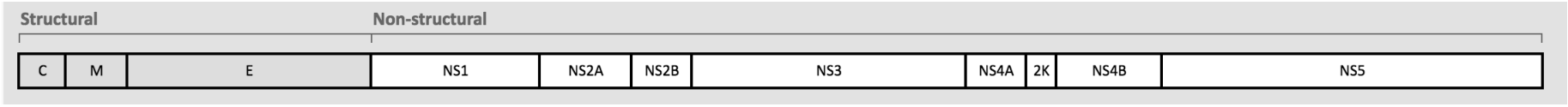
- cause of most serious crop disease in Africa, responsible for between USD$100-$500 million of crop loss per year
- NC_001346 (2689 nt):
  - 4 CDS (V1, V2, C1, C1:C2), 5 total exons
  - both strands have >= 1 CDS
  - one CDS (C1:C2) has two segments (exons)
  - origin sequence: TAATATT|AC
- closed genome
- 508 full genome sequences in GenBank
  - average number of CDS annotated: 3.34

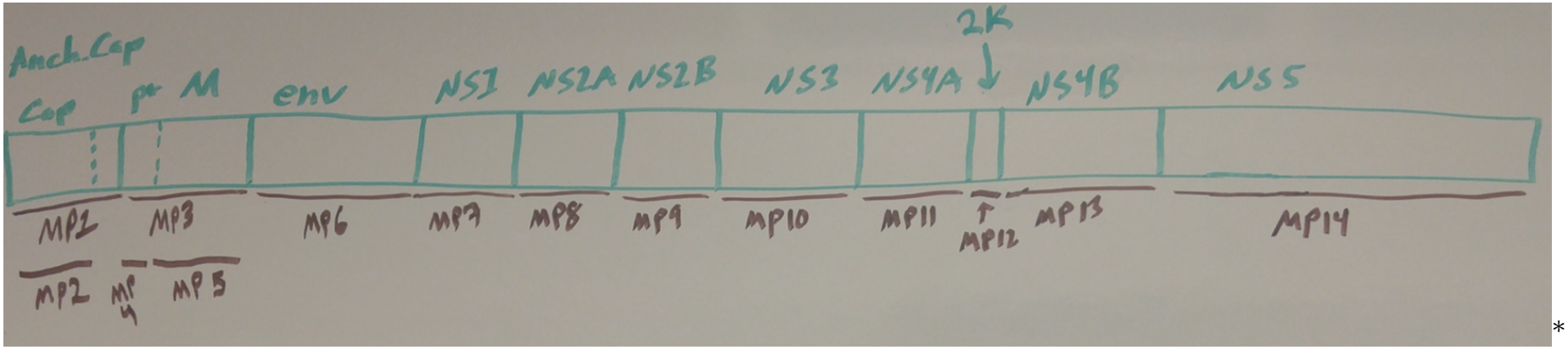


Shepherd et al., Molecular Plant Pathology (2010) 11(1), 112; PMID: 20078771

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



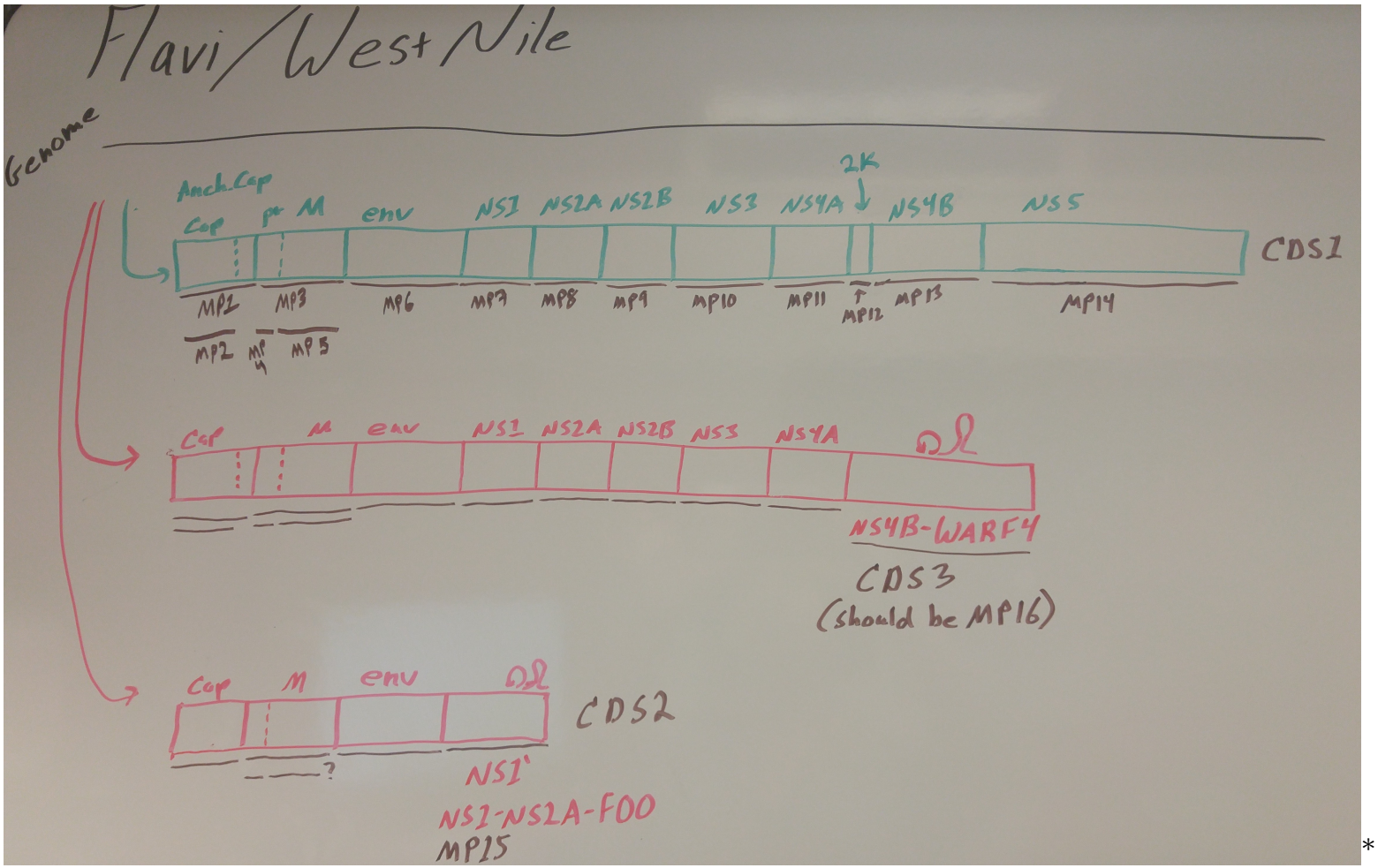


*

*illustration by Eneida Hatcher

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

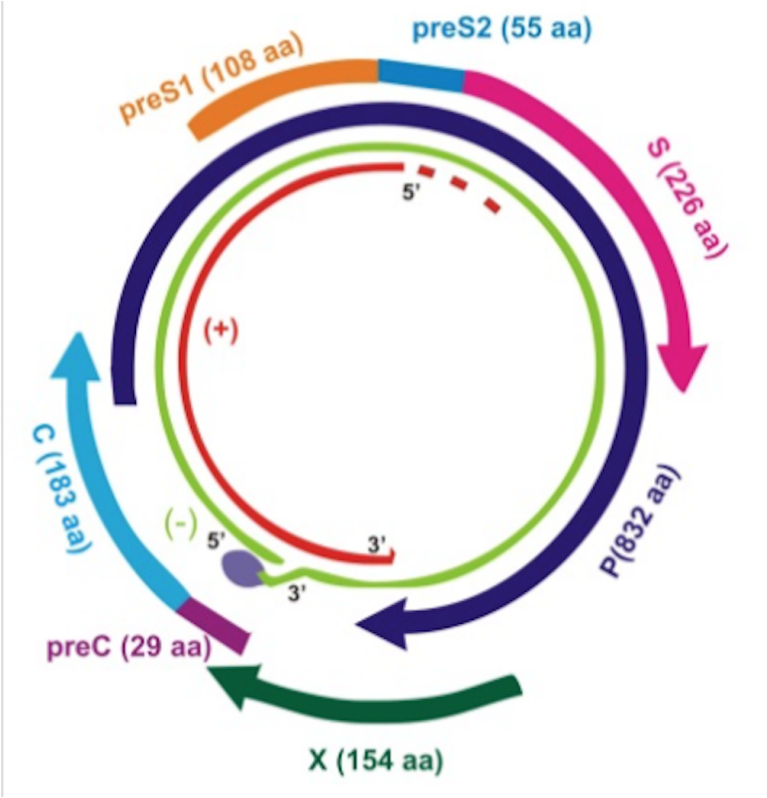


*

*illustration by Eneida Hatcher

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.

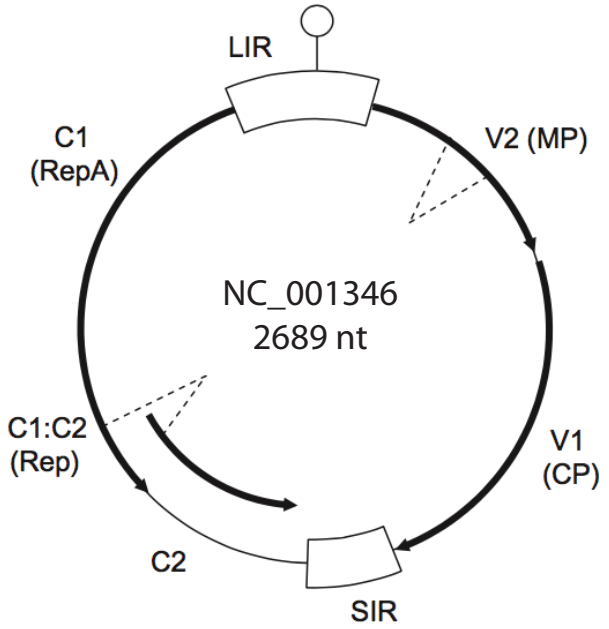
Outline of talk

1. Selection of four pilot viral species
2. Overview of annotation pipeline and explanation of "error codes"
3. Statistics on pilot phase annotations
4. Implementation of annotation pipeline
5. Interaction with J. Rodney Brister's Virus Variation group
6. Future directions

Overview of annotation pipeline for Maize Streak Virus

INPUT:

1. RefSeq annotation

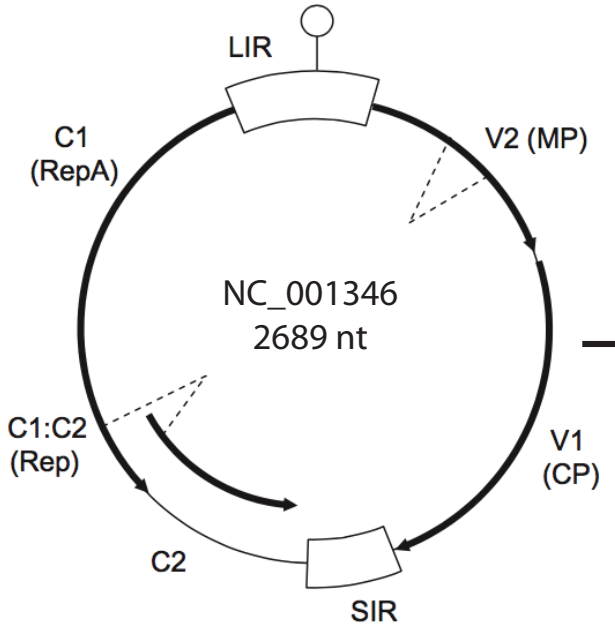


2. Target sequence(s) to annotate:

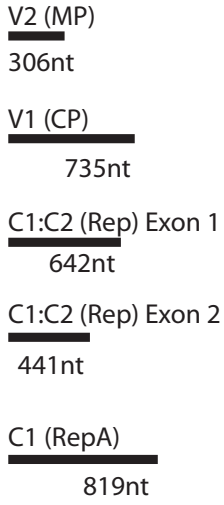
Overview of annotation pipeline for Maize Streak Virus

INPUT:

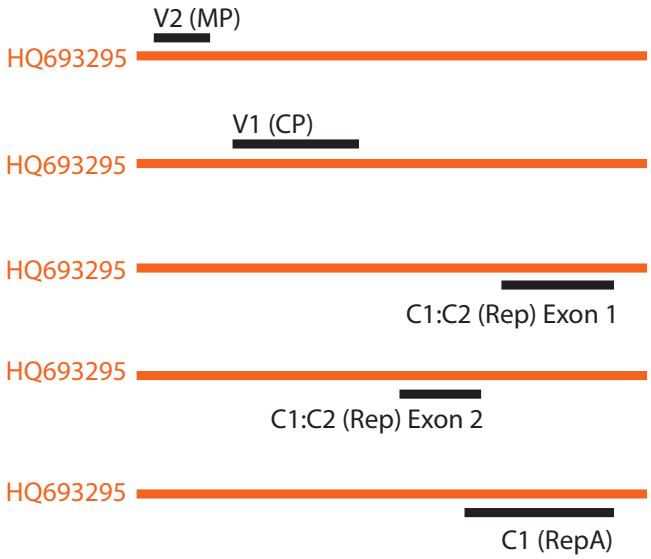
1. RefSeq annotation



Build homology models and search targets:



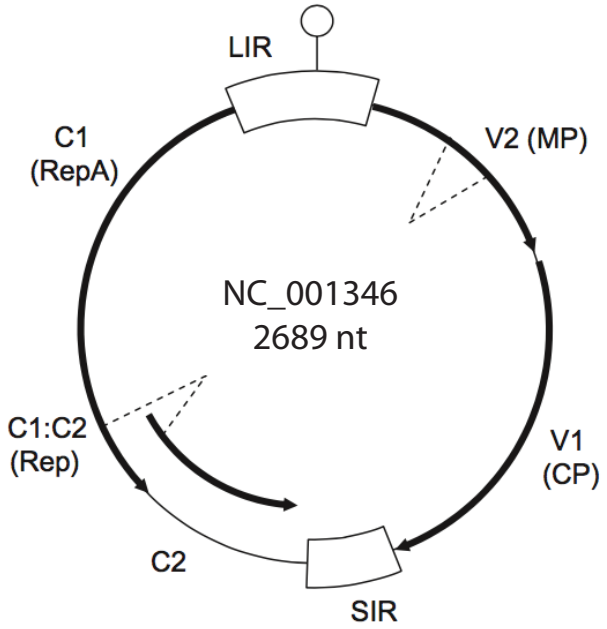
2. Target sequence(s) to annotate:



Overview of annotation pipeline for Maize Streak Virus

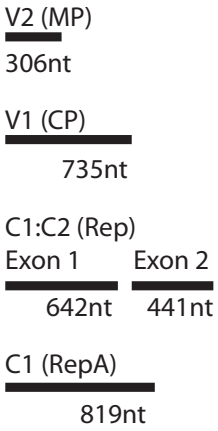
INPUT:

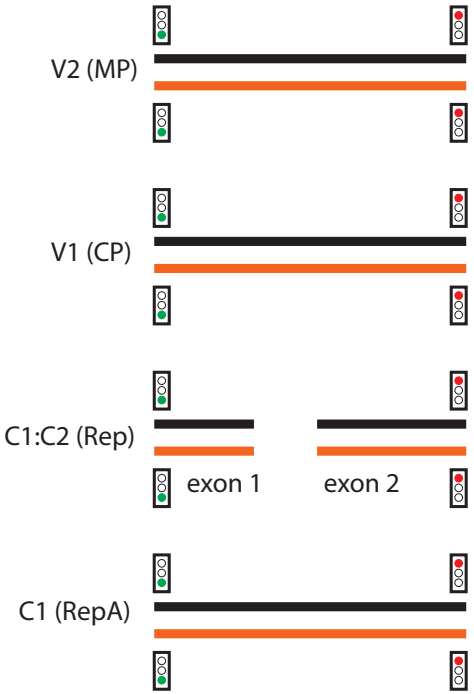
1. RefSeq annotation



2. Target sequence(s) to annotate:

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

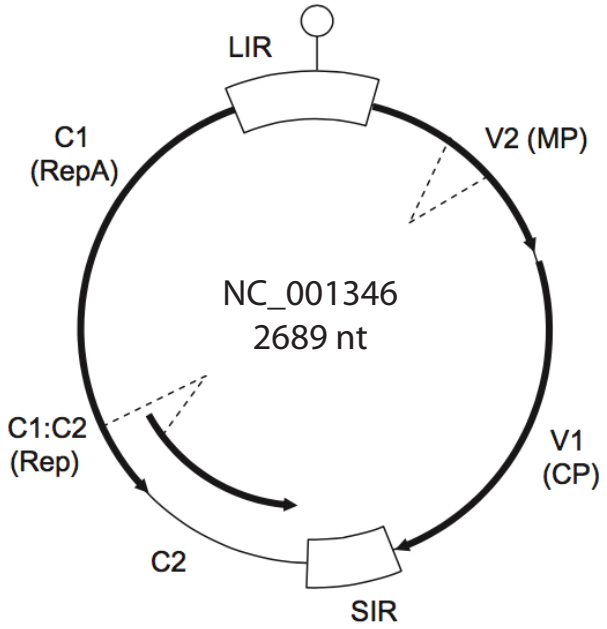




Overview of annotation pipeline for Maize Streak Virus

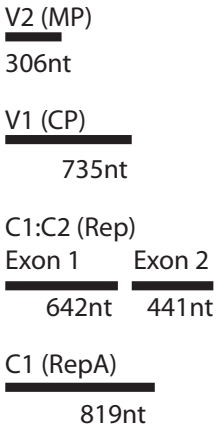
INPUT:

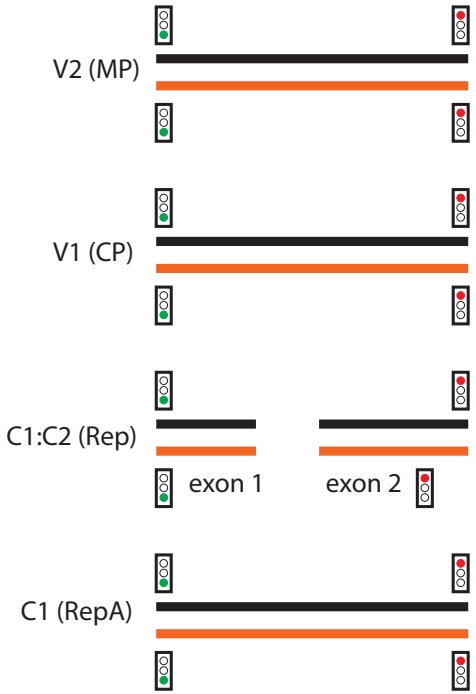
1. RefSeq annotation



2. Target sequence(s) to annotate:

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





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':

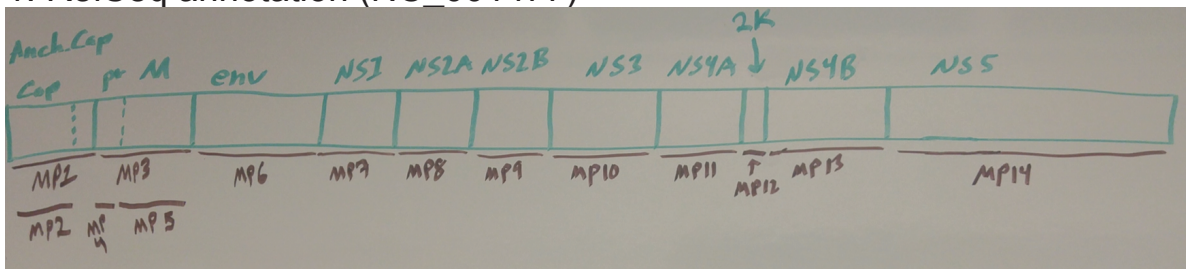
EU628610:Rep:trc

3. Nucleotide and protein multiple alignments (optional)

Overview of annotation pipeline for Dengue Virus

INPUT:

1. RefSeq annotation (NC_001477)

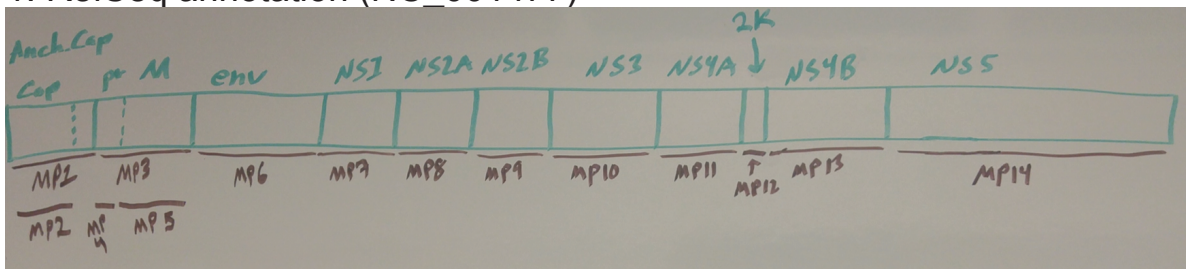


2. Target sequence(s) to annotate:

Overview of annotation pipeline for Dengue Virus

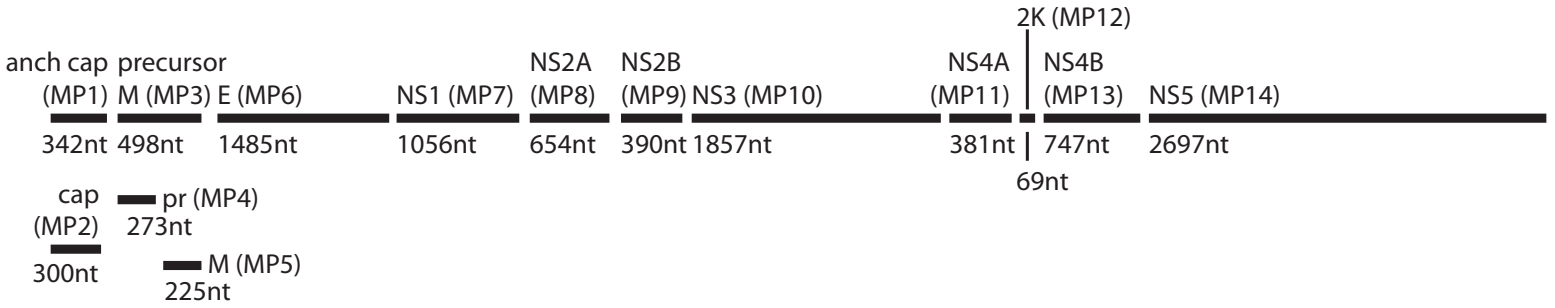
INPUT:

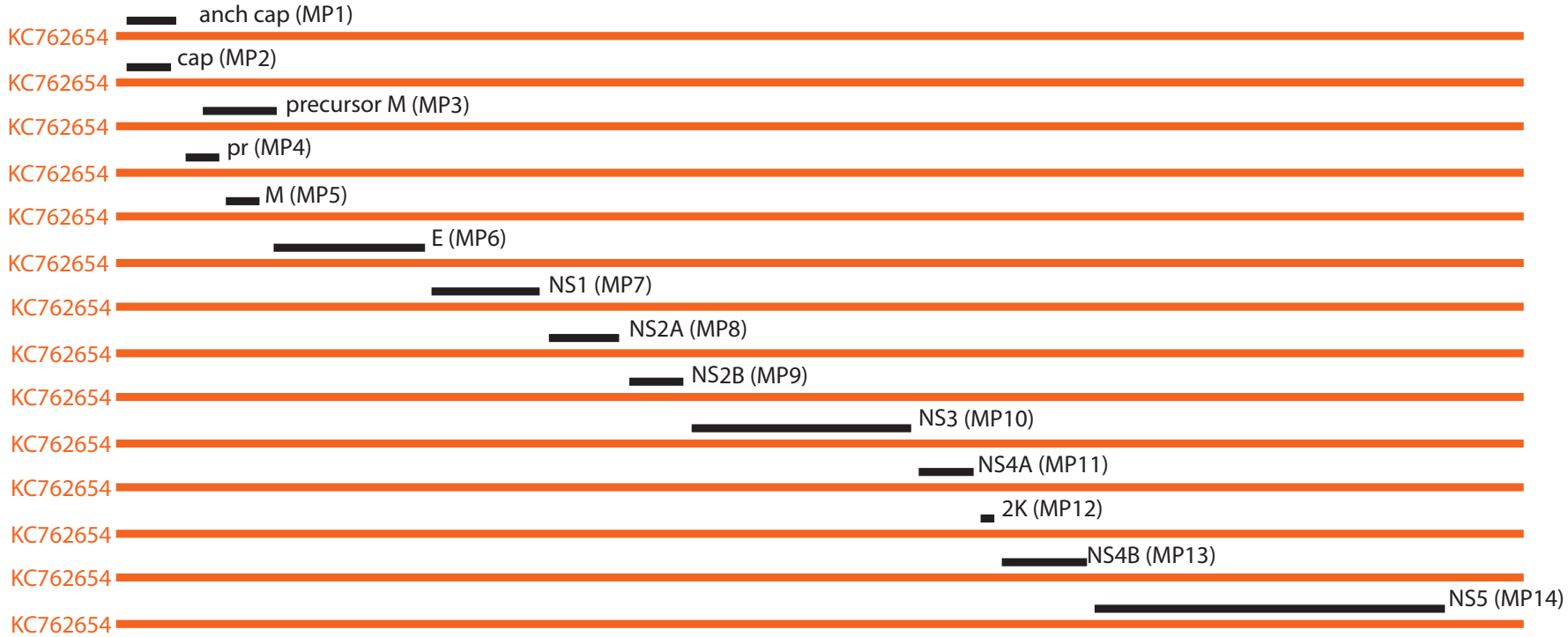
1. RefSeq annotation (NC_001477)



2. Target sequence(s) to annotate:

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

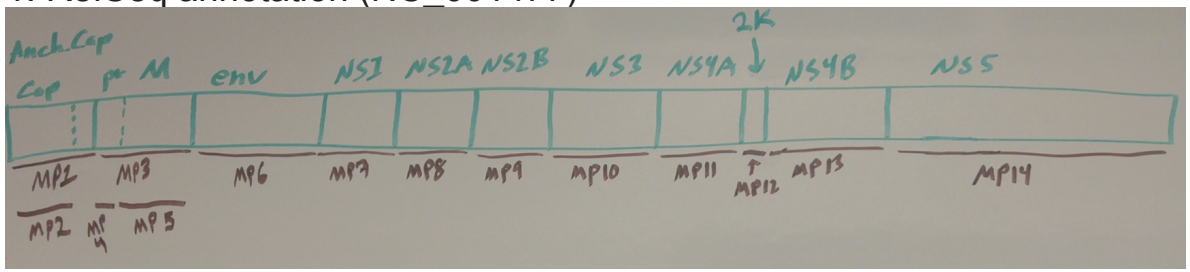




Overview of annotation pipeline for Dengue Virus

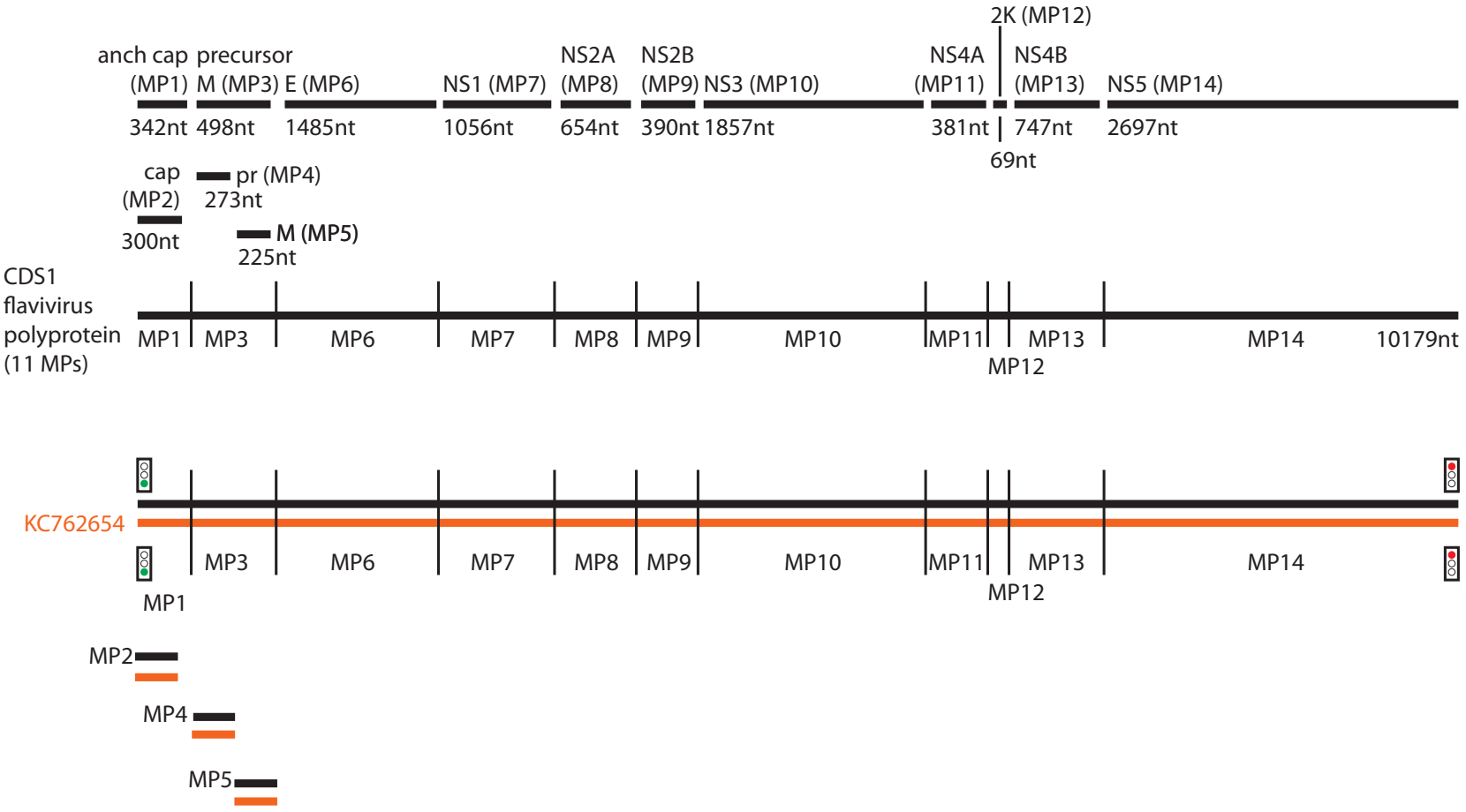
INPUT:

1. RefSeq annotation (NC_001477)



2. Target sequence(s) to annotate:

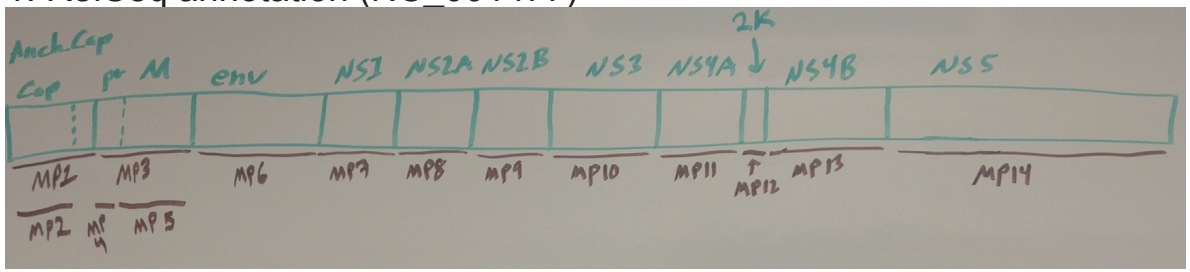
Combine models into 'features' (14 MPs and 1 CDS), translate, find stop codons, and identify errors:



Overview of annotation pipeline for Dengue Virus

INPUT:

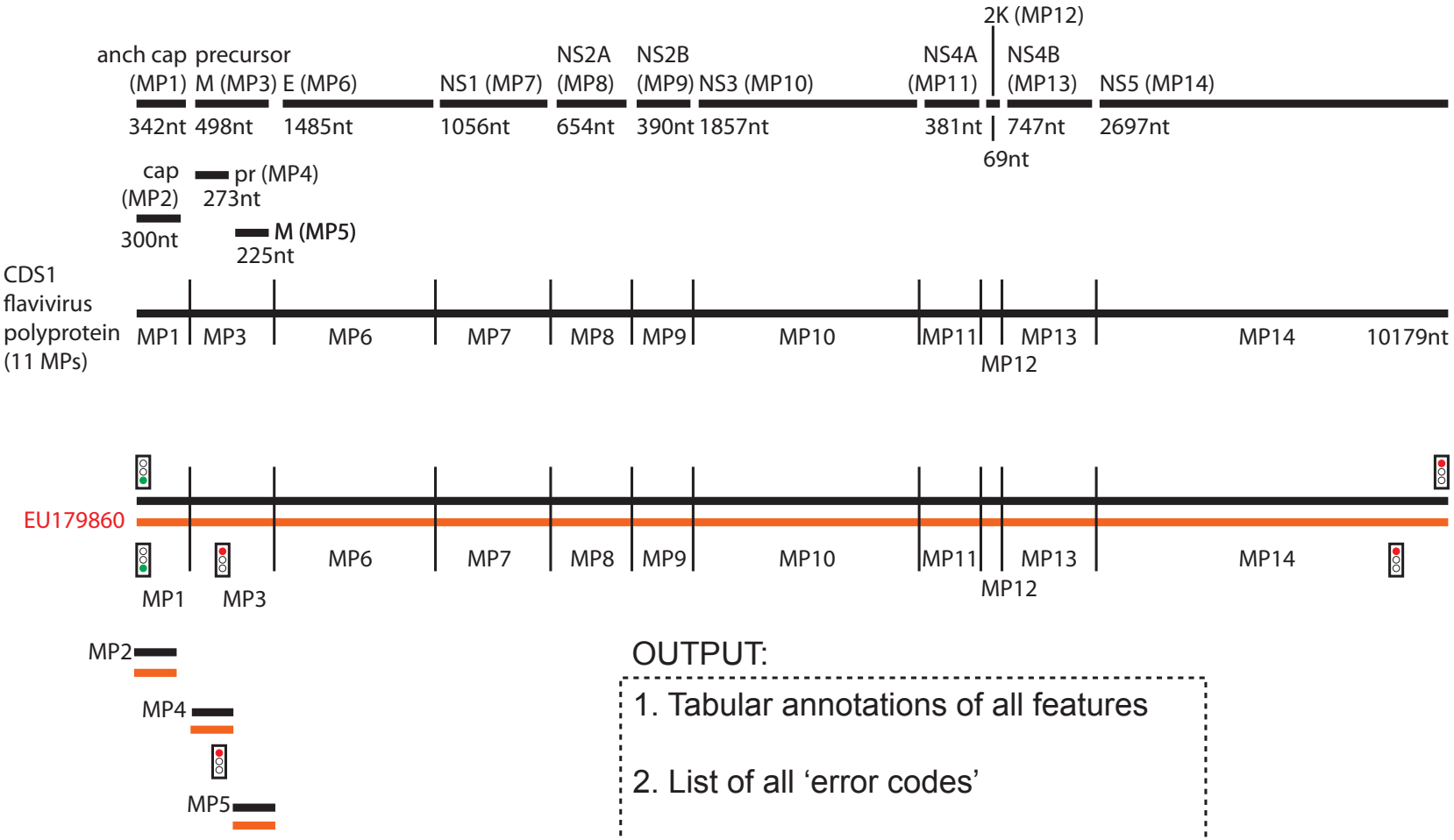
1. RefSeq annotation (NC_001477)



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:



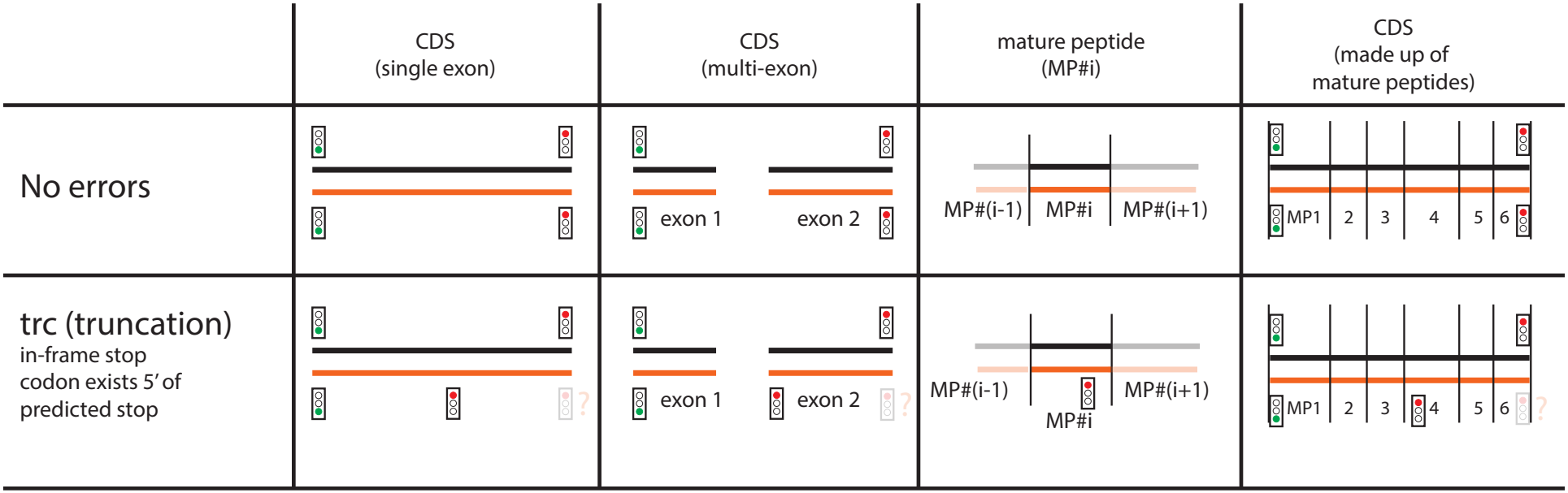
OUTPUT:

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)

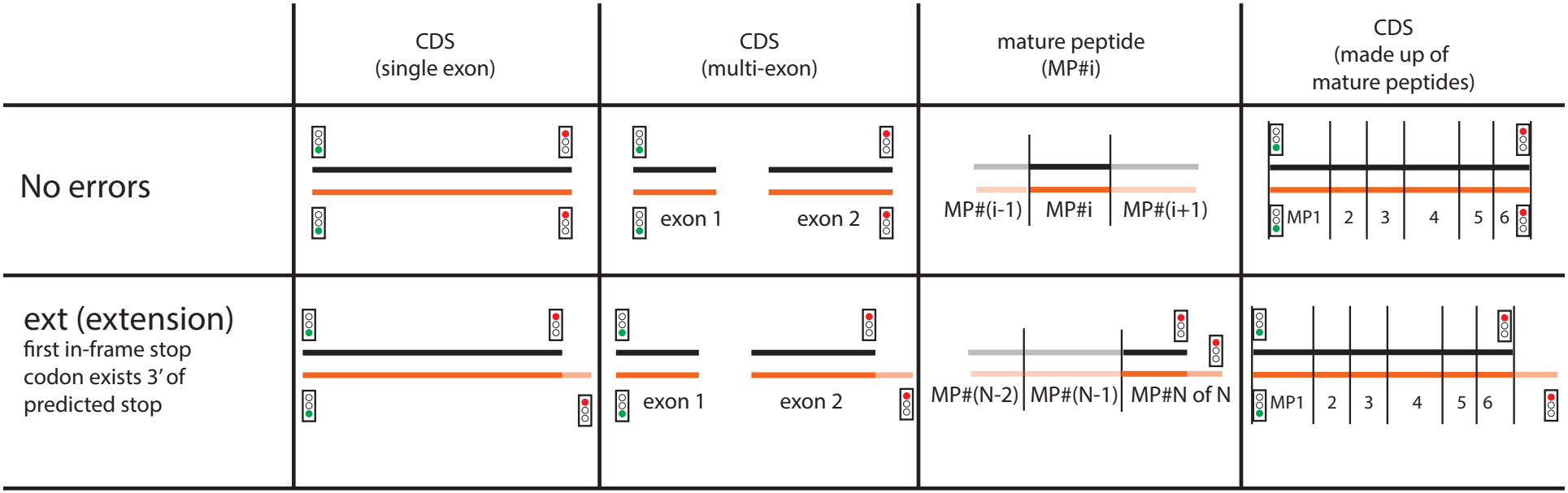
trc error code reports a truncation due to an early stop codon



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

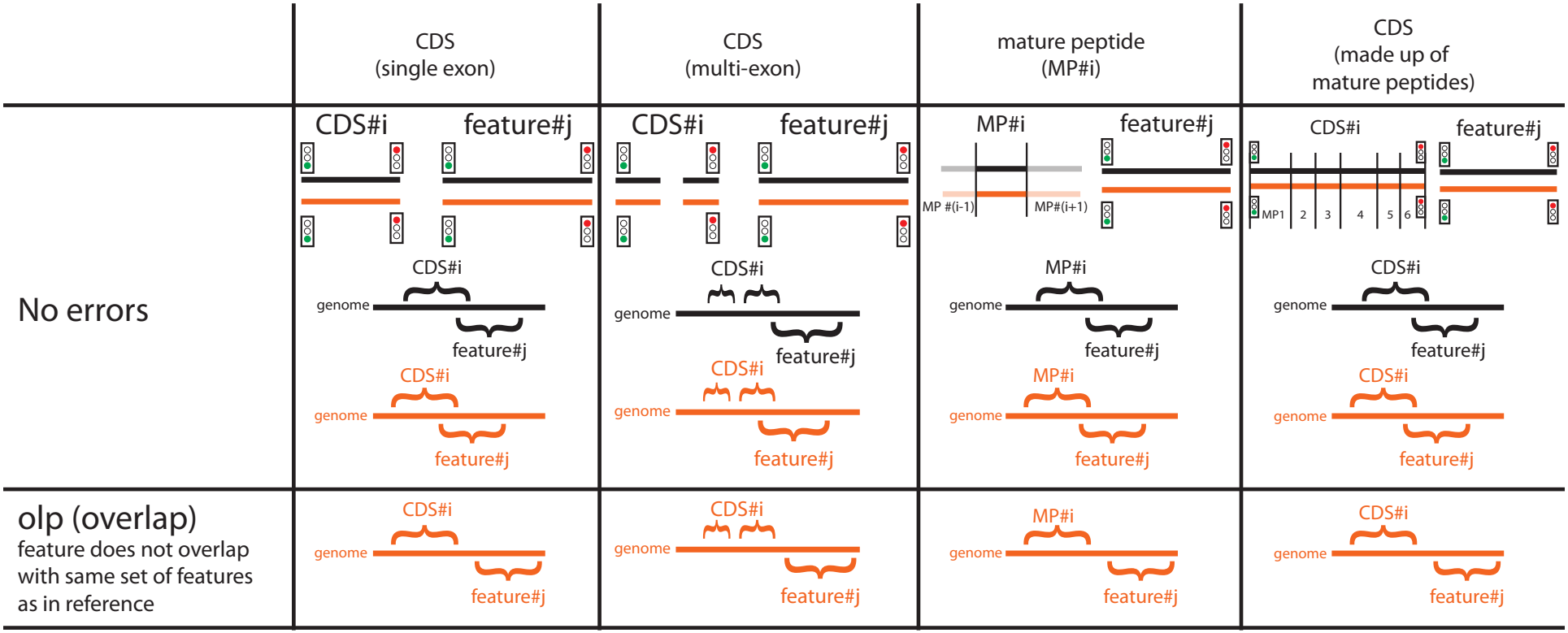


Example output from annotation script:

```
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)]
```

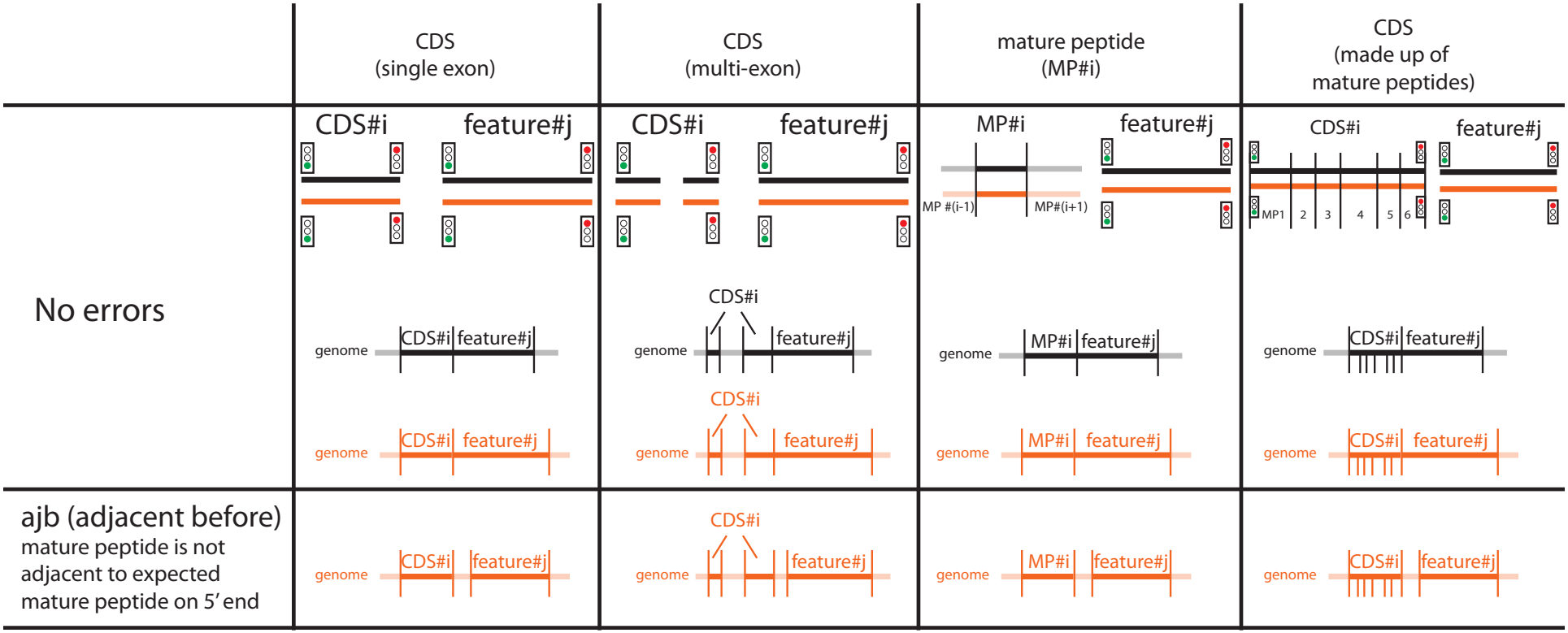
olp: lack of an expected overlap with another feature



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



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)]
```

Outline of talk

1. Selection of four pilot viral species
2. Overview of annotation pipeline and explanation of "error codes"
3. Statistics on pilot phase annotations
4. Implementation of annotation pipeline
5. Interaction with J. Rodney Brister's Virus Variation group
6. Future directions

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)	per-accession averages are in ()
# 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

Maize streak 508 accessions			
code	#tot	#accn	fract
trc	179	177	0.3484
ntr	-	-	-
ext	0	0	0.0000
nst	0	0	0.0000
str	0	0	0.0000
stp	52	52	0.1024
nm3	0	0	0.0000
bd5	0	0	0.0000
bd3	0	0	0.0000
nop	0	0	0.0000
olp	6	3	0.0059
ajb	0	0	0.0000
aja	0	0	0.0000
aji	-	-	-
int	-	-	-
inp	-	-	-
ori	2	2	0.0039
-----	----	-----	------
total	239	-	-
any	178	-	0.3504
none	330	-	0.6496

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		
code	#tot	#accn	fract	#tot	#accn	fract
trc	179	177	0.3484	37	19	0.0050
ntr	-	-	-	60	8	0.0021
ext	0	0	0.0000	6	3	0.0008
nst	0	0	0.0000	1	1	0.0003
str	0	0	0.0000	5	5	0.0013
stp	52	52	0.1024	2	2	0.0005
nm3	0	0	0.0000	15	9	0.0024
bd5	0	0	0.0000	3	2	0.0005
bd3	0	0	0.0000	1	1	0.0003
nop	0	0	0.0000	6	6	0.0016
olp	6	3	0.0059	35	9	0.0024
ajb	0	0	0.0000	65	23	0.0061
aja	0	0	0.0000	60	23	0.0061
aji	-	-	-	15	15	0.0040
int	-	-	-	8	8	0.0021
inp	-	-	-	5	5	0.0013
ori	2	2	0.0039	-	-	-
-----	----	-----	------	----	----	------
total	239	-	-	324	-	-
any	178	-	0.3504	33	-	0.0087
none	330	-	0.6496	3739	-	0.9913

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		
code	#tot	#accn	fract	#tot	#accn	fract	#tot	#accn	fract
trc	179	177	0.3484	37	19	0.0050	91	43	0.0392
ntr	-	-	-	60	8	0.0021	19	2	0.0018
ext	0	0	0.0000	6	3	0.0008	6	3	0.0027
nst	0	0	0.0000	1	1	0.0003	0	0	0.0000
str	0	0	0.0000	5	5	0.0013	3	1	0.0009
stp	52	52	0.1024	2	2	0.0005	3	3	0.0027
nm3	0	0	0.0000	15	9	0.0024	3	3	0.0027
bd5	0	0	0.0000	3	2	0.0005	5	4	0.0036
bd3	0	0	0.0000	1	1	0.0003	0	0	0.0000
nop	0	0	0.0000	6	6	0.0016	4	2	0.0018
olp	6	3	0.0059	35	9	0.0024	18	9	0.0082
ajb	0	0	0.0000	65	23	0.0061	17	7	0.0064
aja	0	0	0.0000	60	23	0.0061	16	7	0.0064
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	2	2	0.0039	-	-	-	-	-	-
-----	----	-----	------	----	----	------	----	----	-------
total	239	-	-	324	-	-	196	-	-
any	178	-	0.3504	33	-	0.0087	47	-	0.0428
none	330	-	0.6496	3739	-	0.9913	1051	-	0.9572

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		
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	2	2	0.0039	-	-	-	-	-	-	5235	5235	0.7481
total	239	-	-	324	-	-	196	-	-	21789	-	-
any	178	-	0.3504	33	-	0.0087	47	-	0.0428	6266	-	0.8954
none	330	-	0.6496	3739	-	0.9913	1051	-	0.9572	732	-	0.1046

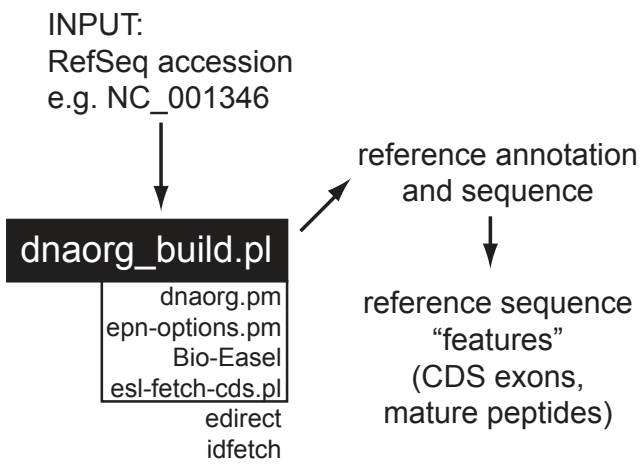
total: total number of error codes reported
  any: number of accessions with >= 1 error code reported
 none: number of accessions with 0 error codes reported

Outline of talk

1. Selection of four pilot viral species
2. Overview of annotation pipeline and explanation of "error codes"
3. Statistics on pilot phase annotations
4. Implementation of annotation pipeline
4. Interaction with J. Rodney Brister's Virus Variation group
6. Future directions

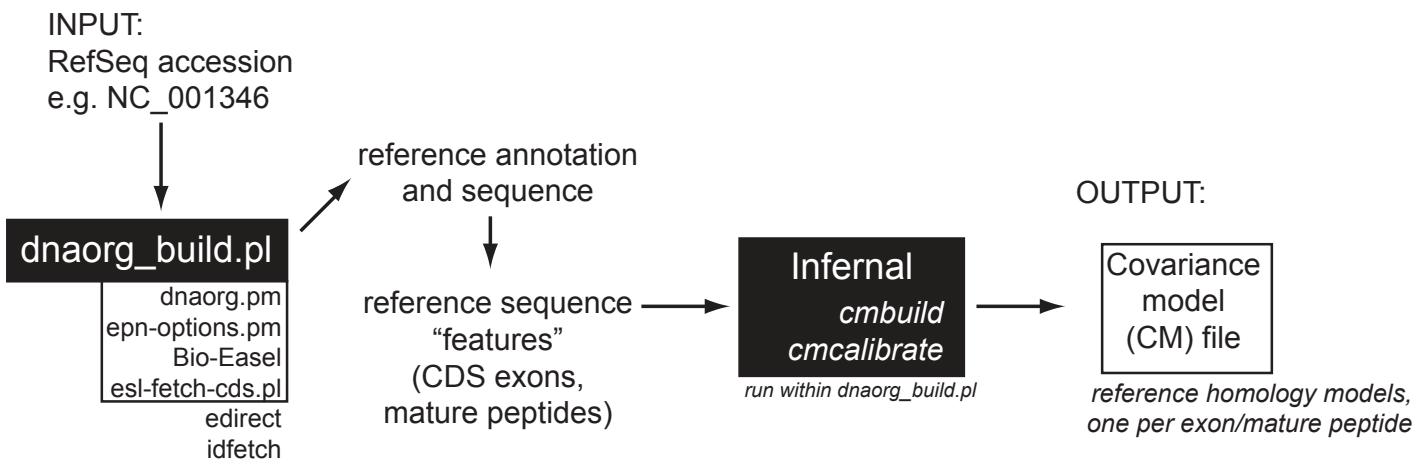
dnaorg_build.pl:

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



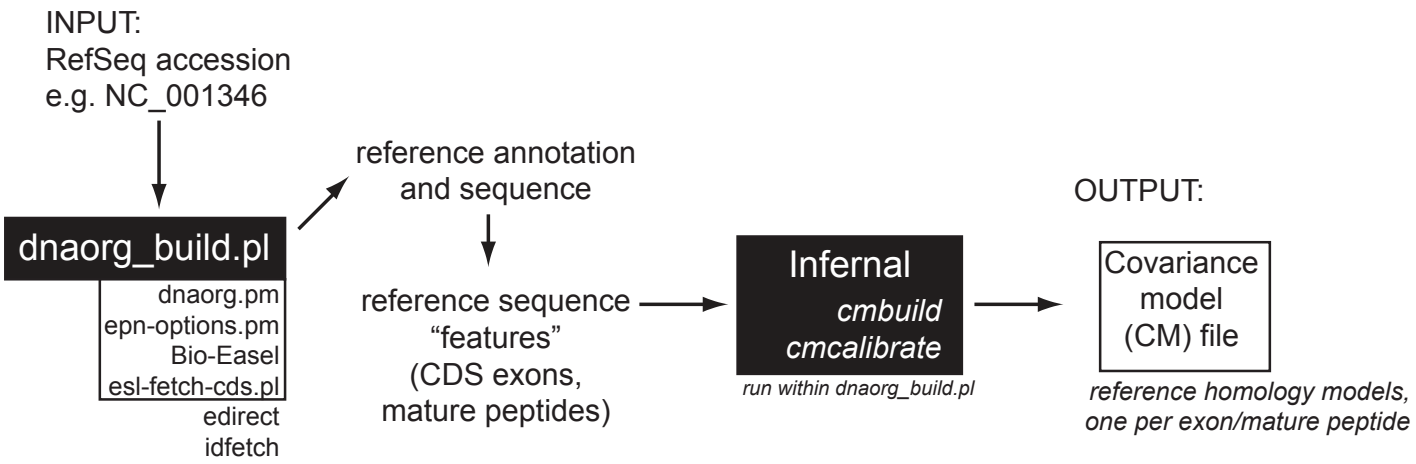
dnaorg_build.pl:

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



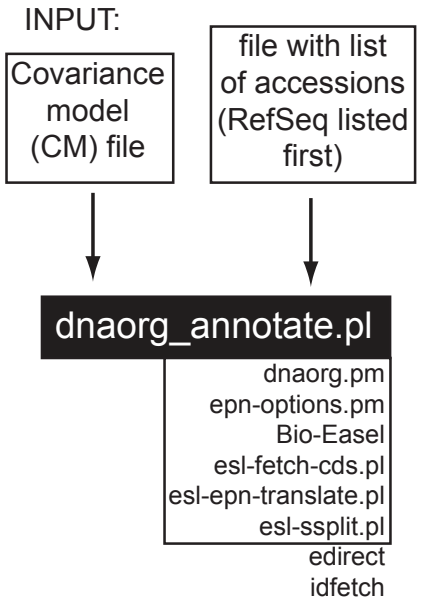
dnaorg_build.pl:

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



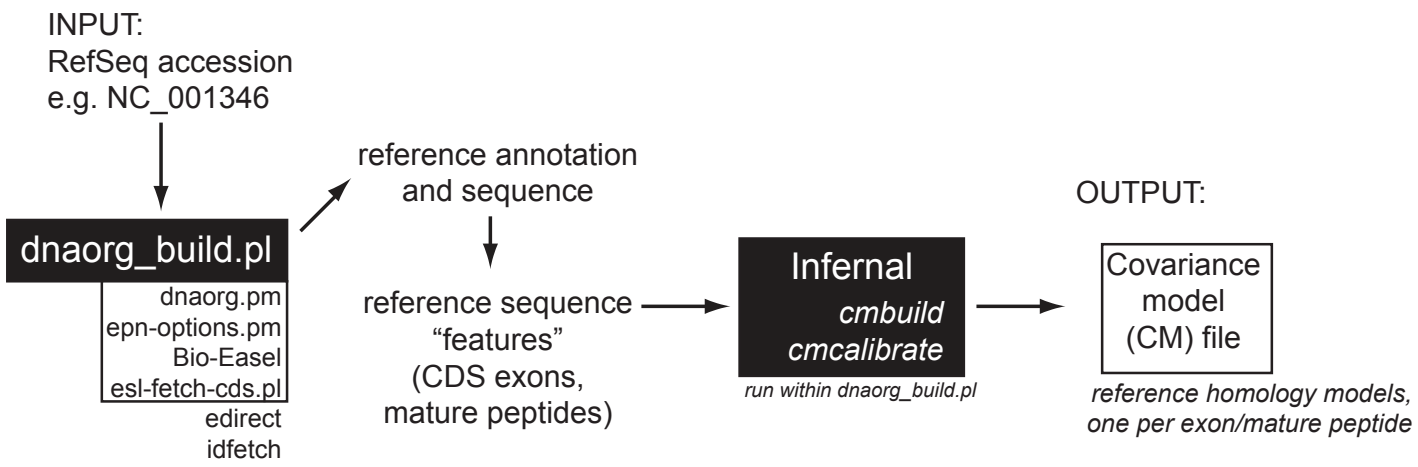
dnaorg_annotate.pl:

Annotate non-reference genomes. Repeat as necessary.



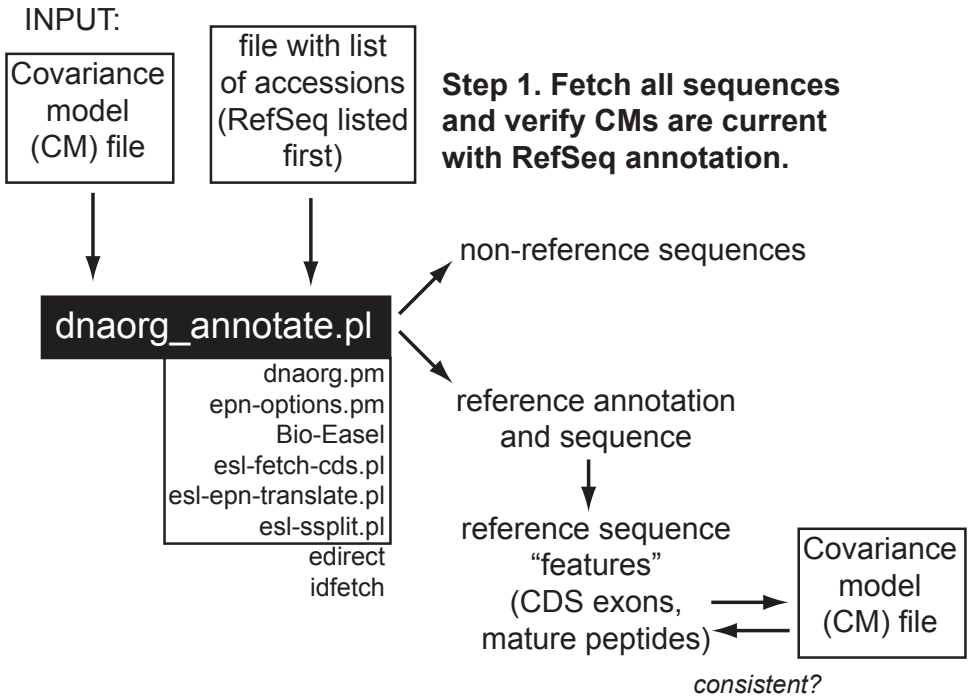
dnaorg_build.pl:

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



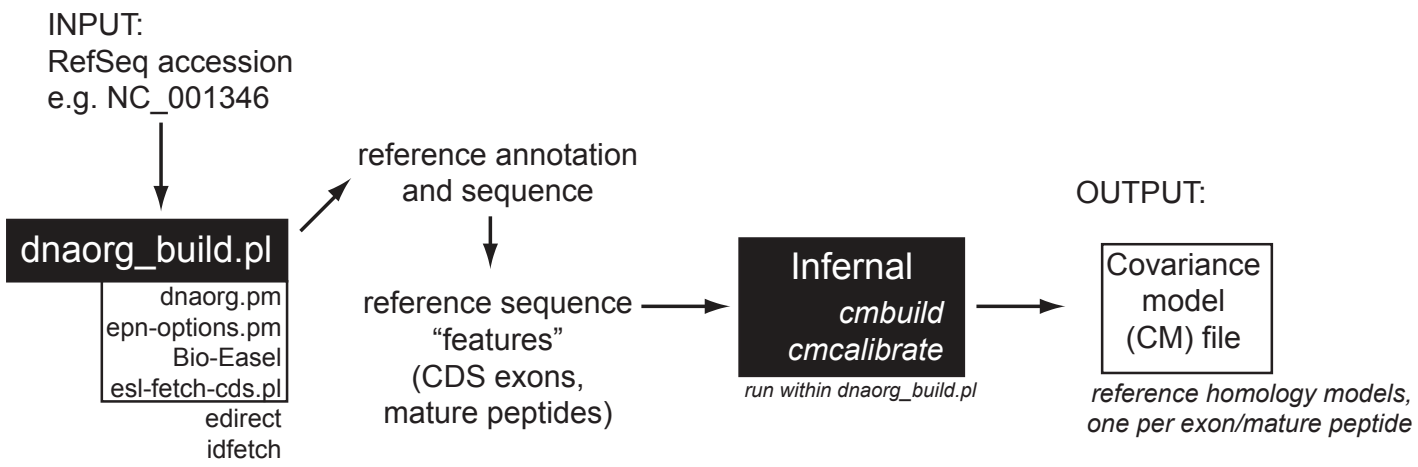
dnaorg_annotate.pl:

Annotate non-reference genomes. Repeat as necessary.



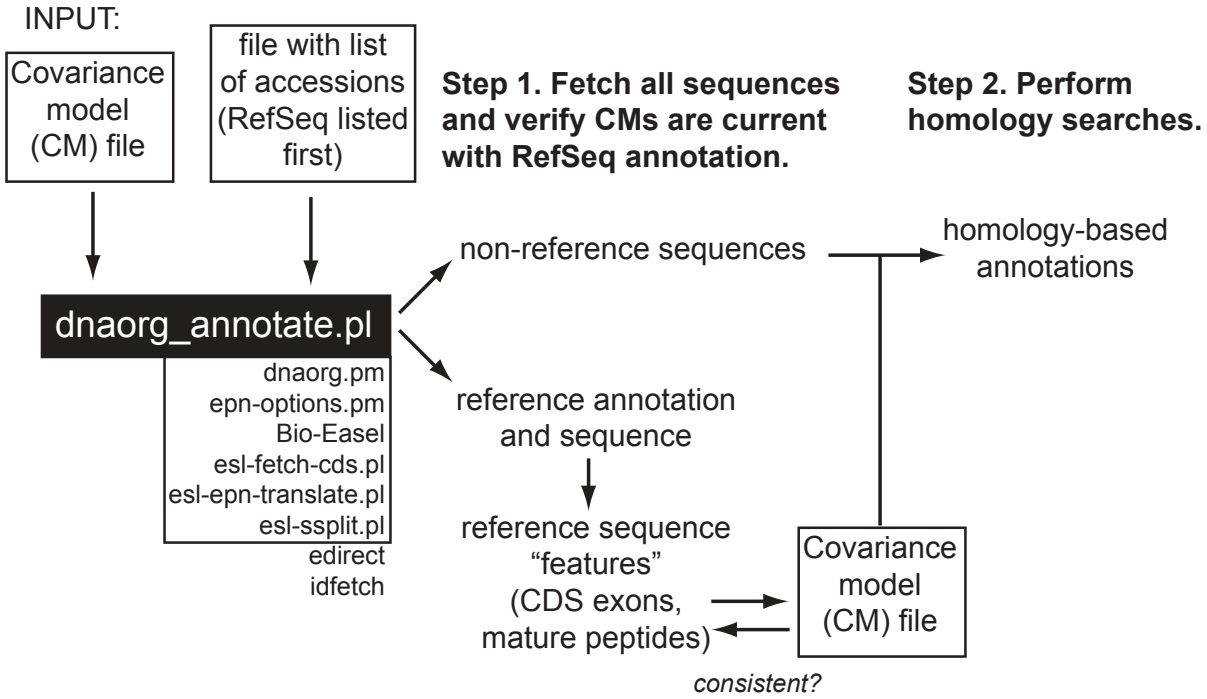
dnaorg_build.pl:

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



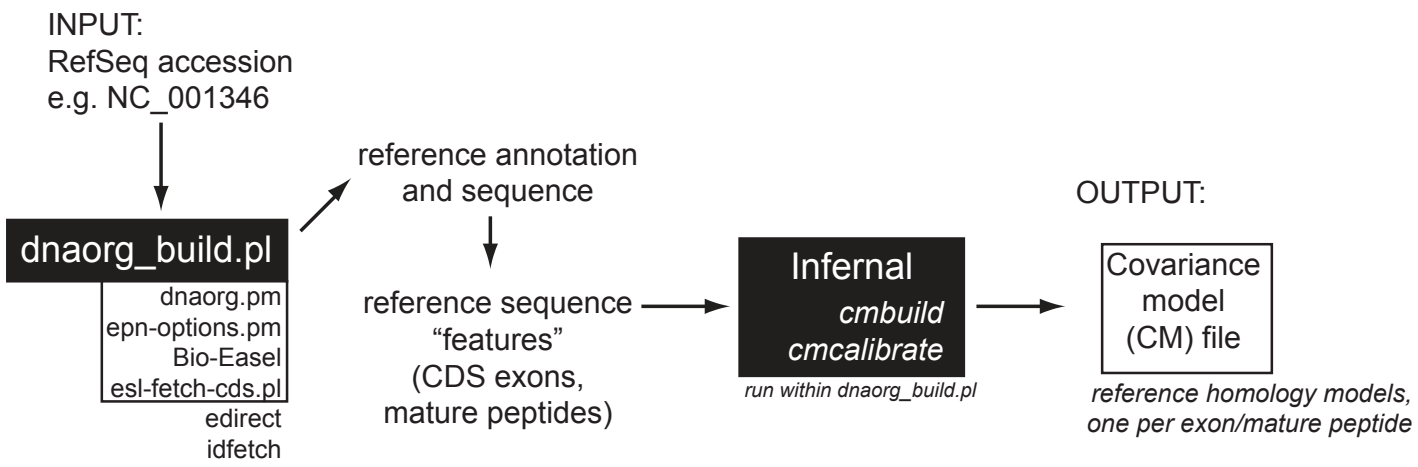
dnaorg_annotate.pl:

Annotate non-reference genomes. Repeat as necessary.



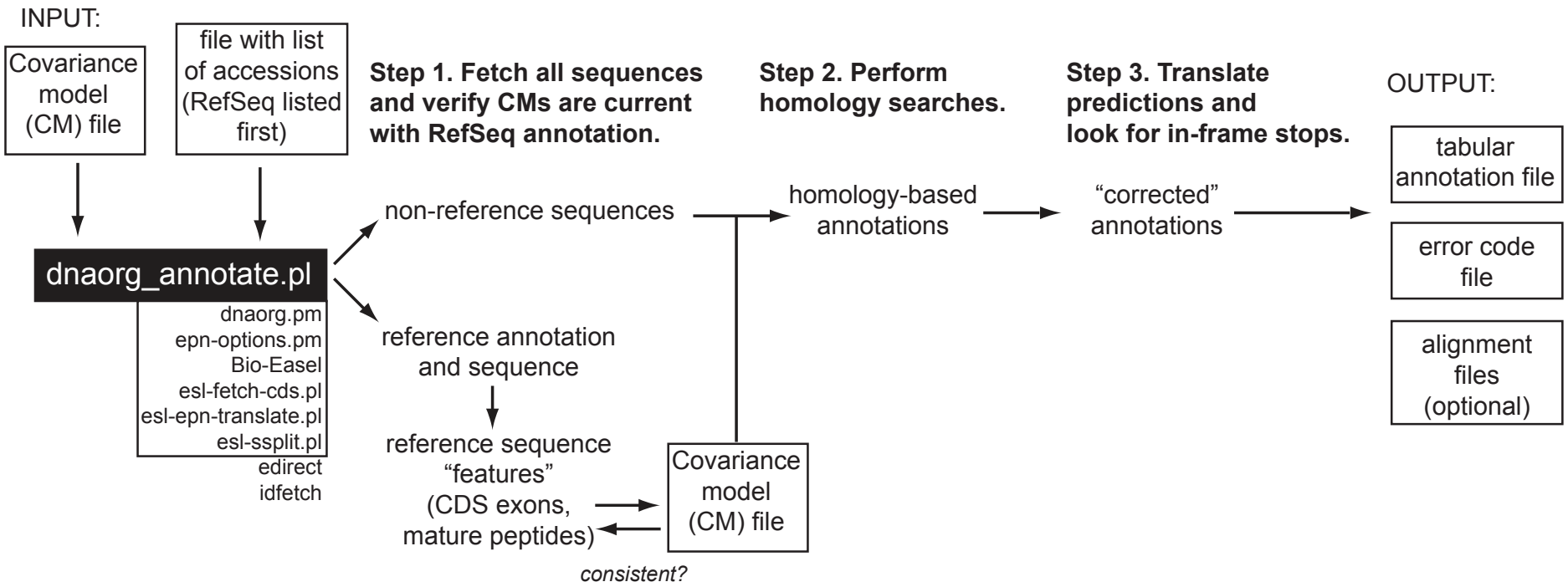
dnaorg_build.pl:

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:                                                          NC_001346.dnaorg_build.log
# List of executed commands saved in:                                                         NC_001346.dnaorg_build.cmd
# List and description of all output files saved in:                                          NC_001346.dnaorg_build.list
# CM file #1, CDS#1 (currently calibrating on the farm) saved in:                             NC_001346.dnaorg_build.0.cm
# CM file #2, CDS#2 (currently calibrating on the farm) saved in:                             NC_001346.dnaorg_build.1.cm
# CM file #3, CDS#3.1 (currently calibrating on the farm) saved in:                           NC_001346.dnaorg_build.2.cm
# CM file #4, CDS#3.2 (currently calibrating on the farm) saved in:                           NC_001346.dnaorg_build.3.cm
# CM file #5, CDS#4 (currently calibrating on the farm) saved in:                             NC_001346.dnaorg_build.4.cm
# Shell script to submit cmcalibrate commands with (already executed, jobs submitted to farm) saved in:   NC_001346.dnaorg_build.ref.cm.qsub
#
# All output files created in directory ./NC_001346/
#
# CPU time:  00:00:58.13
#            hh:mm:ss
#
# DNAORG-SUCCESS
```

Annotating Maize Streak Virus: dnaorg_annotate.pl step

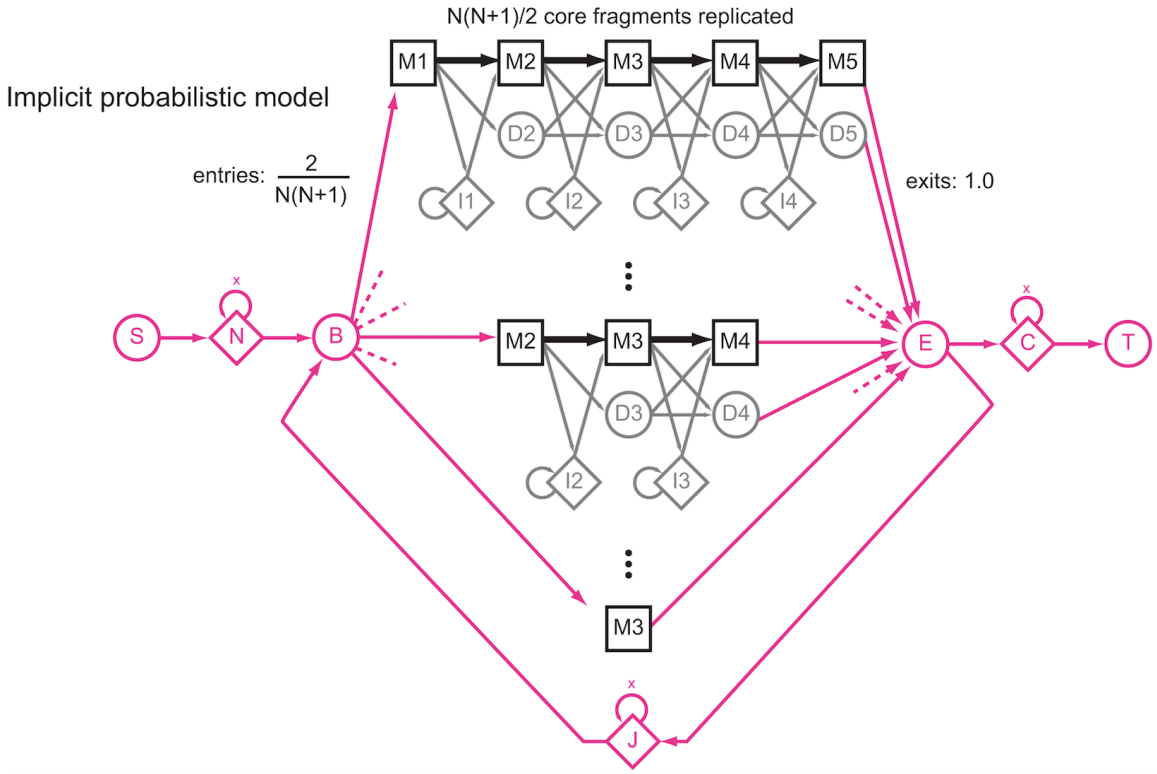
```
<[(example-runs)]> dnaorg_annotate.pl -c --origin TAATATT\|AC NC_001346.ntlist
# dnaorg_annotate.pl :: annotate sequences based on a reference annotation
# dnaorg 0.1 (Apr 2016)
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# date:    Mon Apr 25 15:44:14 2016
#
# file with list of accessions:        NC_001346.ntlist
# genome is closed (a.k.a. circular):  yes [-c]
# identify origin seq <s> in genomes:  TAATATT|AC [--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:                                                 NC_001346.dnaorg_annotate.log
# List of executed commands saved in:                                                NC_001346.dnaorg_annotate.cmd
# List and description of all output files saved in:                                 NC_001346.dnaorg_annotate.list
# CM file (a concatenation of individual files created by dnaorg_build.pl) saved in: NC_001346.dnaorg_annotate.ref.cm
# All annotations in tabular format saved in:                                        NC_001346.dnaorg_annotate.tbl
# Summary of all annotations saved in:                                               NC_001346.dnaorg_annotate.tbl.summary
# Annotations for all sequences with >= 1 failure in tabular format saved in:        NC_001346.dnaorg_annotate.fail.tbl
# Annotations for all sequences with >= 1 error in tabular format saved in:          NC_001346.dnaorg_annotate.error.tbl
# List of errors, one line per sequence saved in:                                    NC_001346.dnaorg_annotate.peraccn.errors
# List of errors, one line per error saved in:                                       NC_001346.dnaorg_annotate.all.errors
# Summary of all errors saved in:                                                    NC_001346.dnaorg_annotate.errors.summary
#
# All output files created in directory ./NC_001346/
#
# CPU time:  00:05:34.66
#            hh:mm:ss
#
# DNAORG-SUCCESS
```

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).

*Eddy, PLoS Comput Biol 4.5 (2008): e1000069.

Outline of talk

1. Selection of four pilot viral species
2. Overview of annotation pipeline and explanation of "error codes"
3. Statistics on pilot phase annotations
4. Implementation of annotation pipeline
5. Interaction with J. Rodney Brister's Virus Variation group
6. Future directions

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

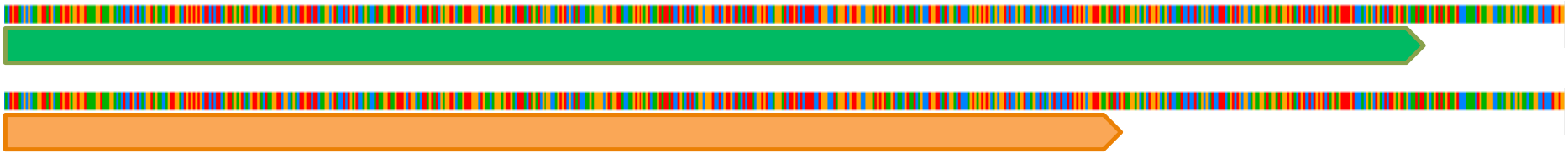
```
160 HM488171    16  MP#16       trc  in-frame stop codon exists 5' of stop position predicted by homology to reference [homolog
161 HM488171    19  CDS(MP)#3   trc  in-frame stop codon exists 5' of stop position predicted by homology to reference [homolog
162 HQ671730    15  MP#15       ext  first 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)]

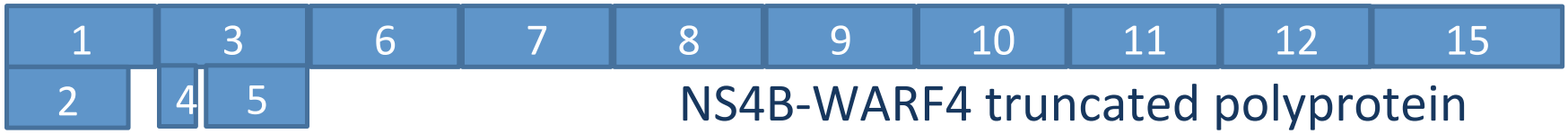
3. Rodney's group provides Alejandro and Eric with a list suggesting improvements that should be made in the code.

slide created by Eneida Hatcher

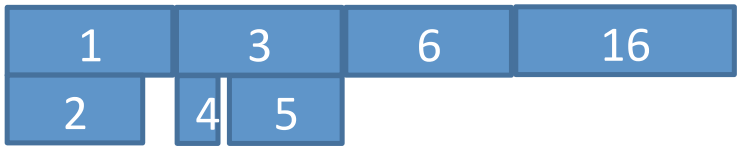
Identifying Rare Errors

West Nile virus, lineage 1

3 typical translations



NS4B-WARF4 truncated polyprotein

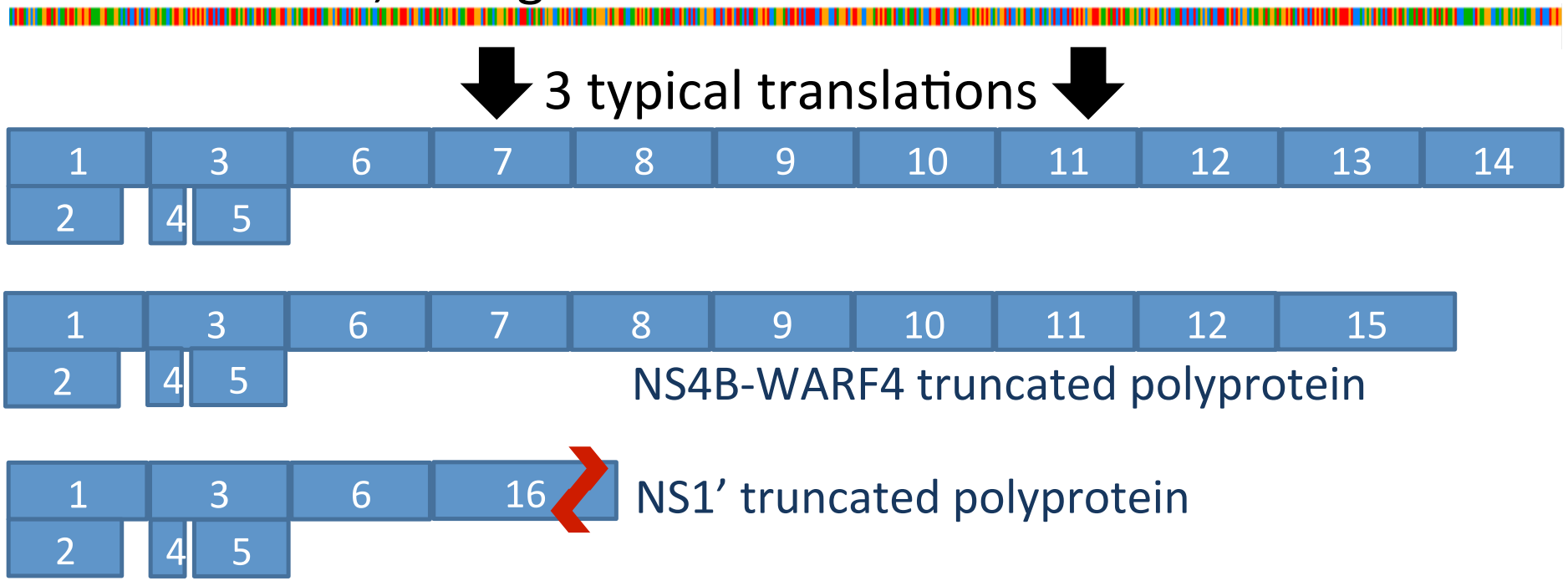


NS1’ truncated polyprotein

slide created by Eneida Hatcher

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

slide created by Eneida Hatcher

Identifying Rare Errors

West Nile virus, lineage 1

JX503096 errors:

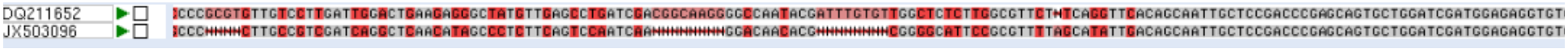
#1 olp feature does not overlap with same set of features as in reference [-(1.1,2.1)]

#2 olp feature does not overlap with same set of features as in reference [-(2.1,1.1)]

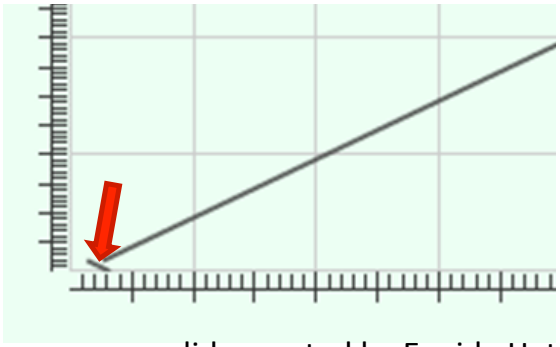
#2 trc in-frame stop codon exists 5' of stop position predicted by homology to reference [homology search predicted 228..1 revised to 228..175 (stop shifted 174 nt)]

CDS str predicted CDS start position is not beginning of ATG start codon [ATT starting at position 51 on strand +]

Poor alignment:



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



slide created by Eneida Hatcher

Future directions

- Annotate more species (Norovirus, Zika, and Ebola are currently being checked).

- Long term:
  - Protein homology searches to supplement or replace nucleotide searches
  - Given any viral sequence, find its nearest RefSeq
  - Annotate structural RNA features (Infernal is well-suited for this)
  - More robust solution to the problem of finding origin sequences
  - Multiple alignment based homology searches

- Short term: we have a 36 item TODO list that we agreed to with J. Rodney Brister's group on April 12, which includes:
  - Completed items (including analysis of West Nile annotations)
  - Loading and using our pipeline's annotations in the Virus Variation database for Dengue and West Nile
  - Reviewing Ebola and Norovirus annotations
  - Comparing our Ebola annotations to existing annotations

Acknowledgements

Alejandro Schäffer

David Landsman

David Lipman

J. Rodney Brister

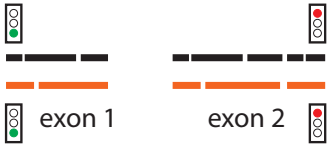
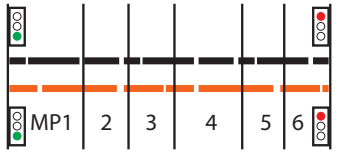
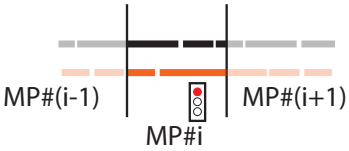
Eneida Hatcher

Olga Blinkova

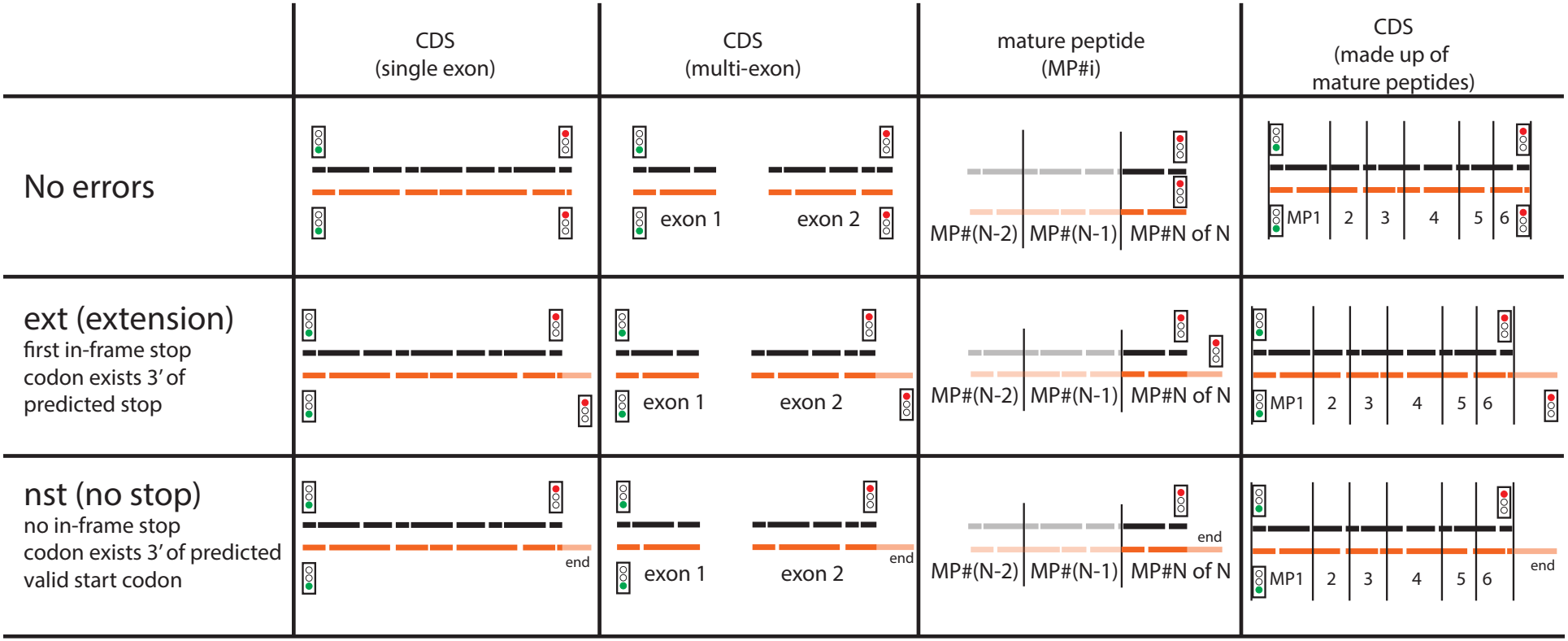
Sergey Zhdanov

Yiming Bao

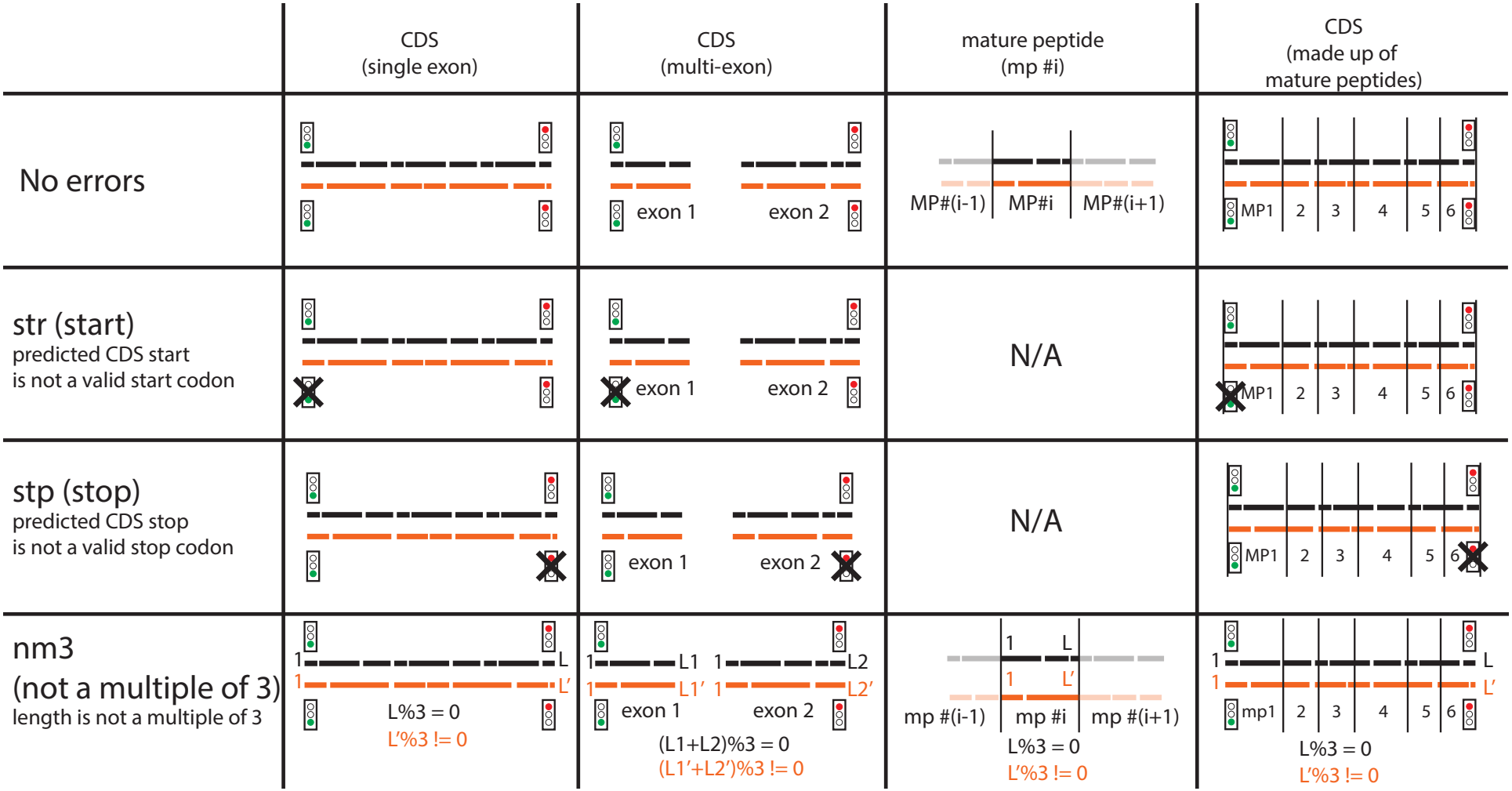
Unexpected stop codon errors (trc,ntr)

	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

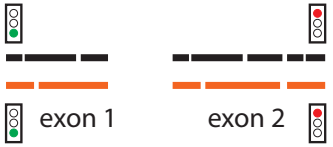
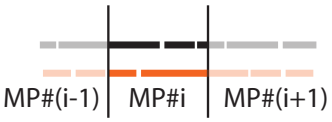
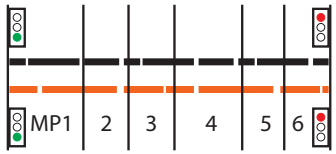
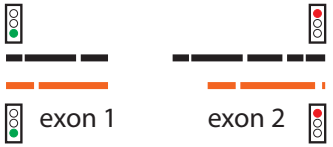
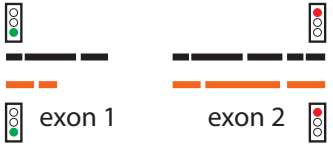
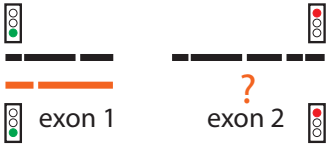
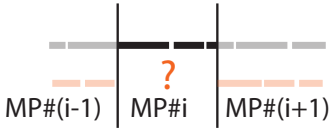
Unexpected stop codon errors (ext,nst)



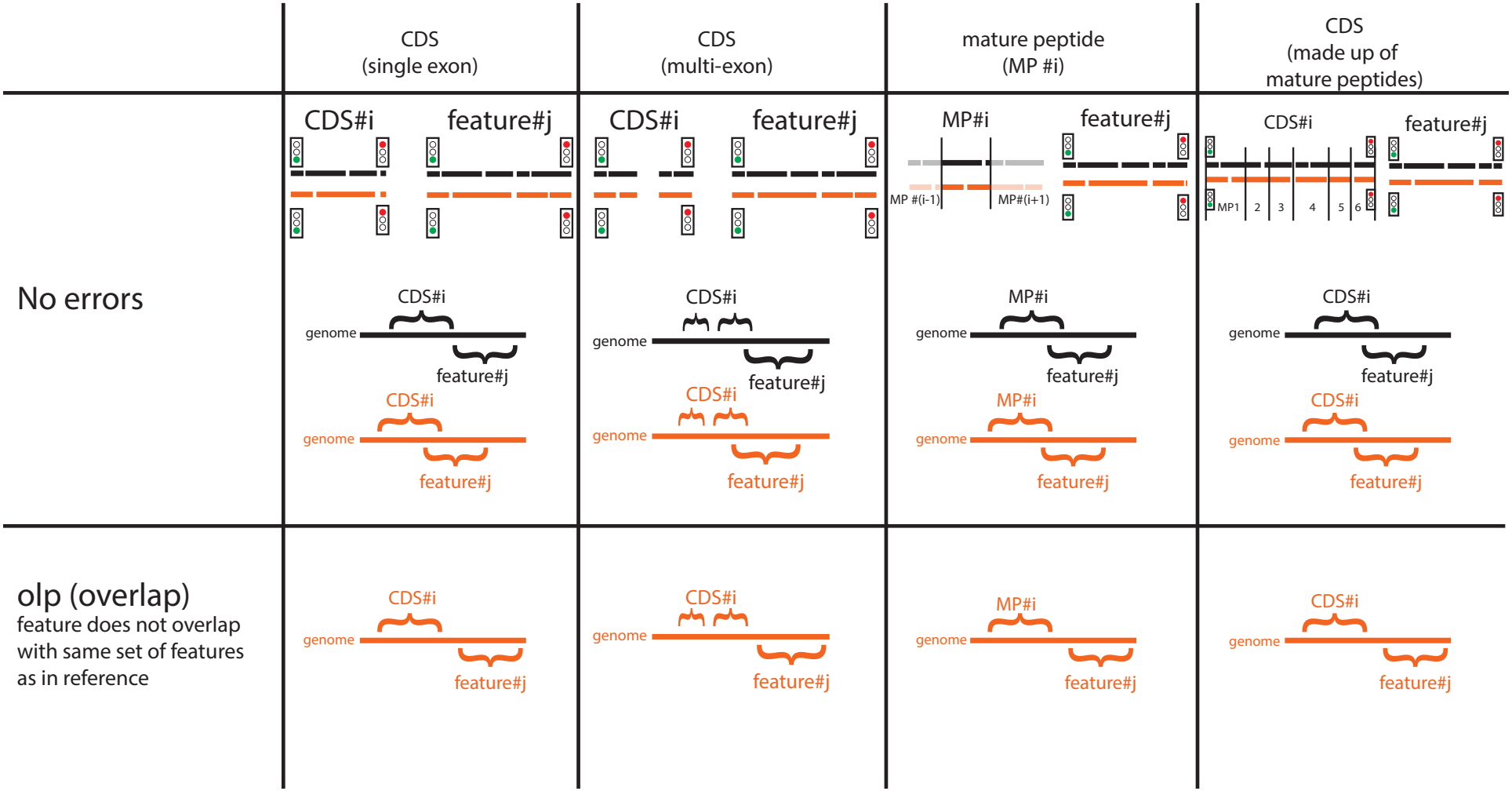
Missing expected features (str,stp,nm3)



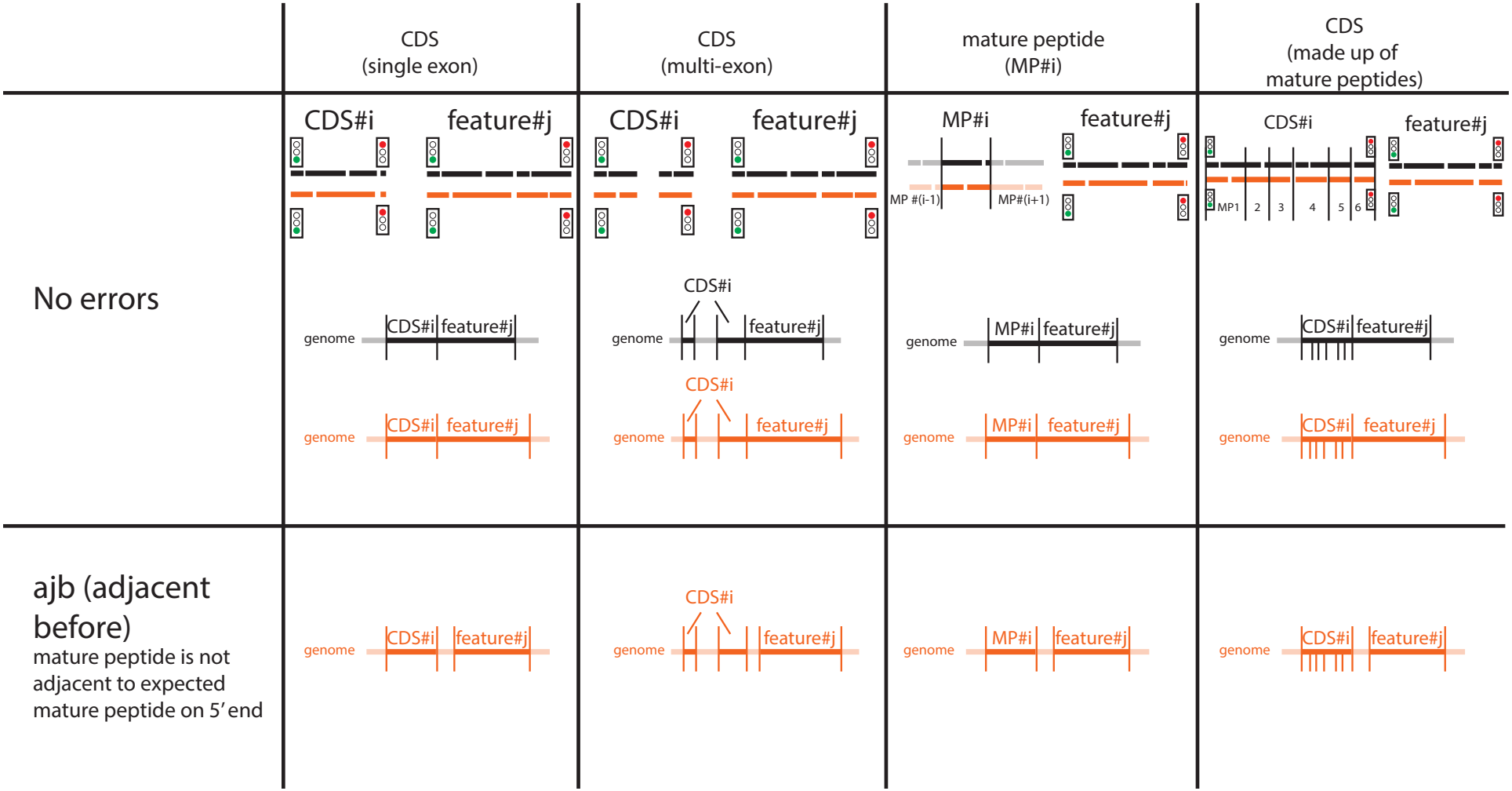
Problem with homology search prediction (bd5,bd3,nop)

	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

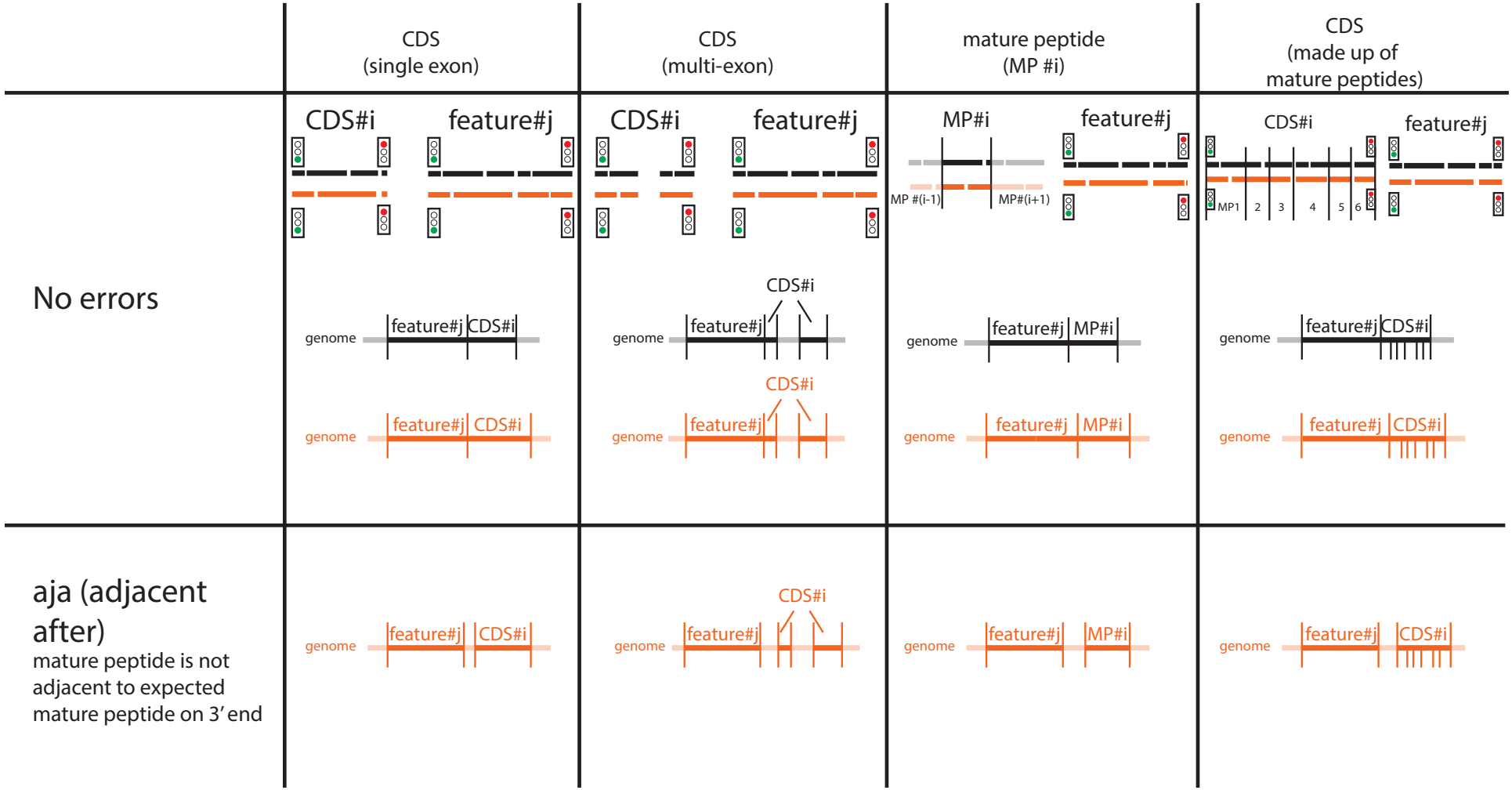
Unexpected relationship to other features (olp)



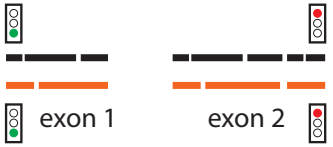
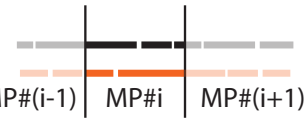
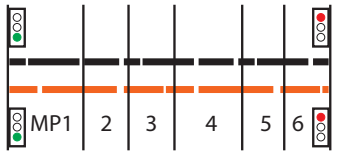
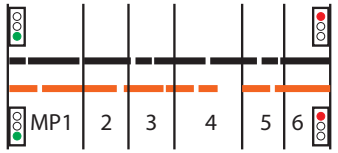
Unexpected relationship to other features (ajb)



Unexpected relationship to other features (aja)



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)