

Reference-based viral sequence annotation using VADR

Eric Nawrocki

National Center for Biotechnology Information
National Institutes of Health



INSDC (GenBank/ENA/DDBJ) has a lot of sequence data

D94–D99 *Nucleic Acids Research*, 2019, Vol. 47, Database issue
doi: 10.1093/nar/gky989

Published online 26 October 2018

GenBank

Eric W. Sayers[✉], Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt and Ilene Karsch-Mizrachi

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

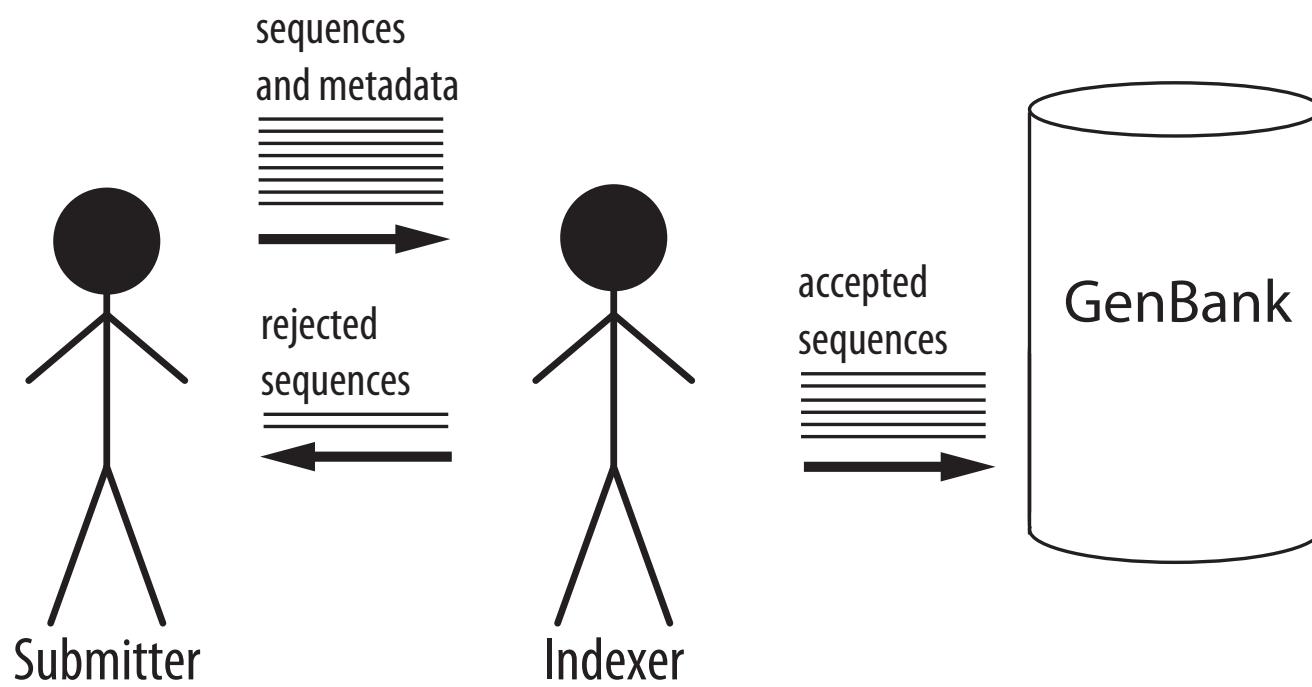
Table 1. Growth of GenBank divisions (nucleotide base-pairs)

| Division | Description | Release 227 (August 2018) | Annual increase (%) ^a |
|----------|----------------------------|---------------------------|----------------------------------|
| MAM | Other mammals | 6 214 774 850 | 60.47% |
| WGS | Whole genome shotgun data | 3 204 855 013 281 | 42.93% |
| UNA | Unannotated | 296 706 | 42.25% |
| PLN | Plants | 23 027 832 426 | 37.21% |
| BCT | Bacteria | 53 541 127 504 | 36.93% |
| TSA | Transcriptome shotgun data | 225 520 004 678 | 35.01% |
| PHG | Phages | 463 029 085 | 34.38% |
| VRL | Viruses | 4 073 816 676 | 16.99% |
| PAT | Patent sequences | 22 019 723 131 | 14.57% |
| VRT | Other vertebrates | 10 441 689 546 | 12.90% |
| ENV | Environmental samples | 5 818 999 756 | 4.09% |
| HTC | High-throughput cDNA | 721 454 983 | 3.57% |
| PRI | Primates | 8 262 441 252 | 2.96% |
| SYN | Synthetic | 1 192 279 390 | 1.62% |
| GSS | Genome survey sequences | 26 339 143 098 | 1.40% |
| EST | Expressed sequence tags | 42 988 632 150 | 0.82% |
| HTG | High-throughput genomic | 27 770 730 435 | 0.45% |
| ROD | Rodents | 4 534 815 151 | 0.31% |
| STS | Sequence tagged sites | 640 879 986 | 0.00% |
| INV | Invertebrates ^b | 8 597 126 159 | −50.09% |
| TOTAL | All GenBank sequences | 3 677 023 810 243 | 39.52% |

^aMeasured relative to Release 221 (August 2017).

^bThe decrease in INV data resulted from the suppression of 36 nematode-related genomes. See the release notes for Release 227 for more details (<ftp.ncbi.nlm.nih.gov/genbank/release.notes/gb227.release.notes>).

Manual NCBI GenBank indexing does not scale

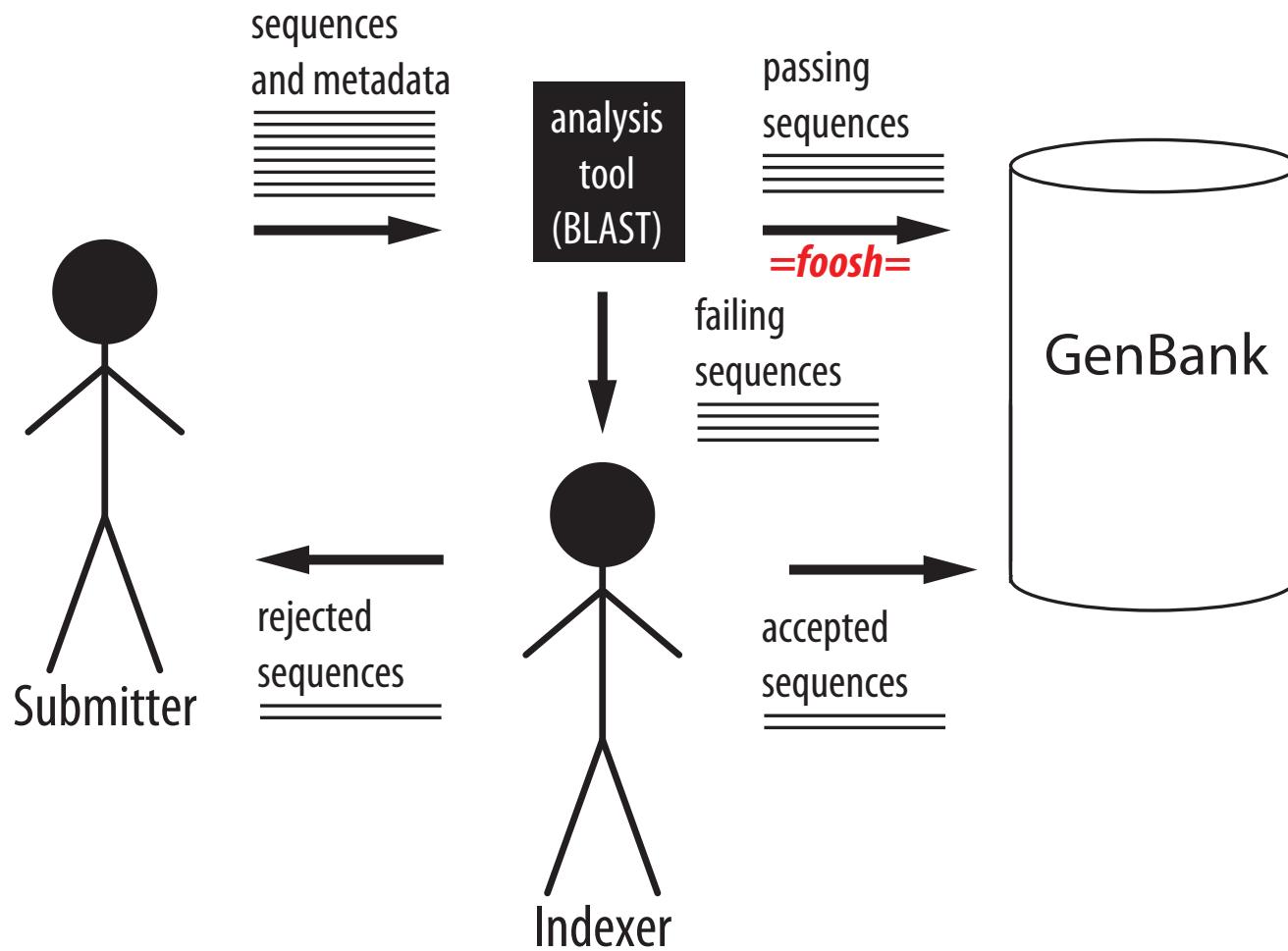


Sequence submissions are handled by expert NCBI indexers

- Indexers check submissions for quality
- Many submissions are of *marker genes*, used to characterize environments (microbiome, soil), which are automatically analyzed by BLAST or specialized tools.

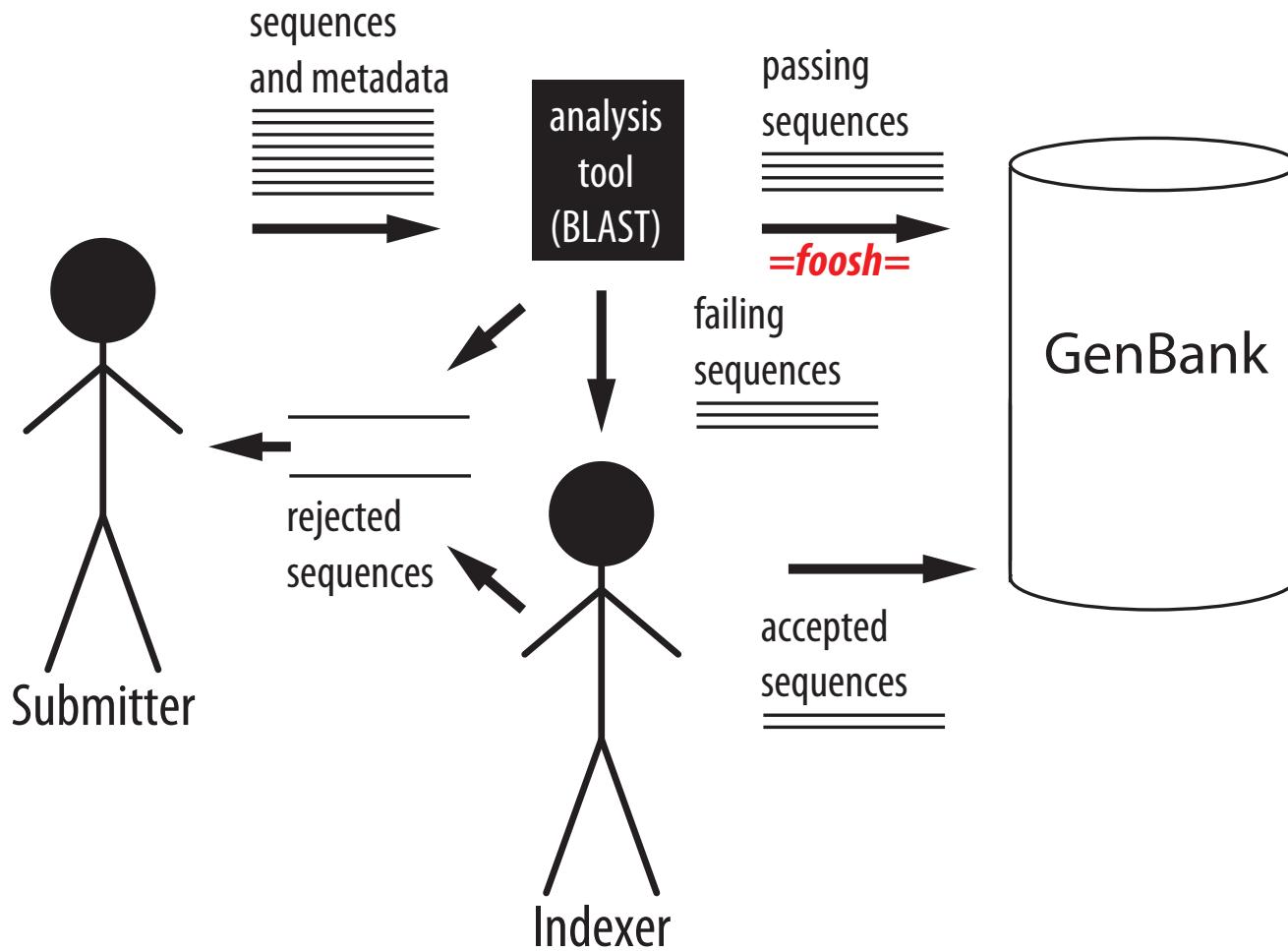
| marker gene/ sequence type | 2018 # seqs | total # seqs |
|-------------------------------|----------------|-----------------|
| 16S rRNA | 333,121 | 8,015,297 |
| COX1 | 35,517 | 1,349,957 |
| 23S rRNA | 74,287 | 275,014 |
| ITS1 | 27,279 | 359,380 |
| ITS2 | 24,144 | 184,515 |
| ITS1+ITS2 | 26,734 | 445,721 |
| Influenza | 74,868 | 665,464 |

NCBI GenBank Indexers use BLAST



- Foosh pipelines exist for 16S, 23S, ITS (BLAST-based) and Influenza (FLAN)

NCBI GenBank Indexers use BLAST



- Foosh pipelines exist for 16S, 23S, ITS (BLAST-based) and Influenza (FLAN)

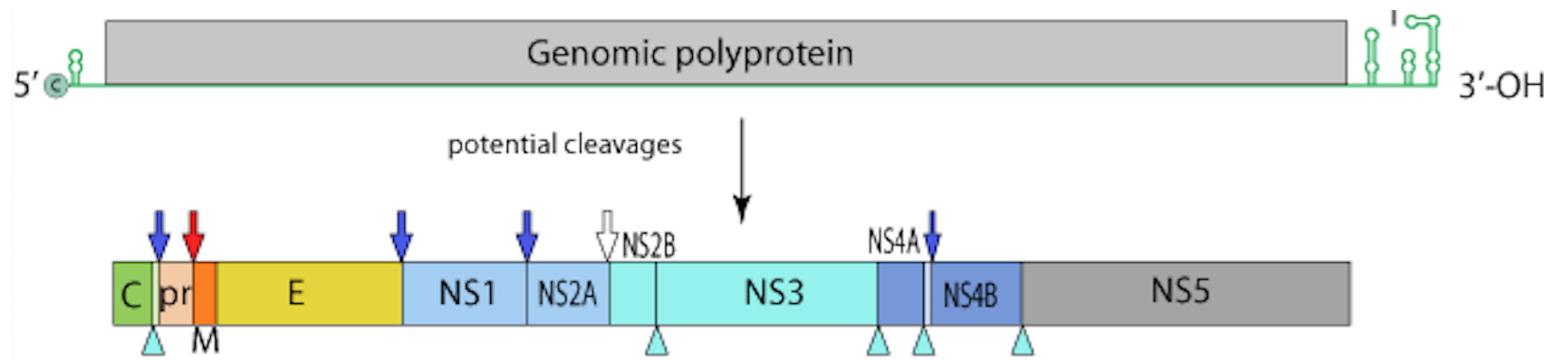
Viruses with highest number of sequences in INSDC*

| species | #seqs | family |
|---------------------------|---------|-------------------------|
| HIV-1 | 850,115 | <i>Retroviridae</i> |
| Influenza A virus | 684,026 | <i>Orthomyxoviridae</i> |
| Hepacivirus C | 244,533 | <i>Flaviviridae</i> |
| Hepatitis B virus | 114,306 | <i>Hepadnaviridae</i> |
| Influenza B virus | 100,373 | <i>Orthomyxoviridae</i> |
| Rotavirus A | 73,375 | <i>Reoviridae</i> |
| SIV | 44,374 | <i>Retroviridae</i> |
| Norovirus (Norwalk virus) | 40,925 | <i>Caliciviridae</i> |
| Enterovirus A | 31,478 | <i>Picornaviridae</i> |
| PRRSV | 29,081 | <i>Arteriviridae</i> |
| Dengue virus | 28,564 | <i>Flaviviridae</i> |
| Human orthopneumovirus | 24,384 | <i>Pneumoviridae</i> |
| Enterovirus B | 23,865 | <i>Picornaviridae</i> |
| Rabies lyssavirus | 23,771 | <i>Rhabdoviridae</i> |
| West Nile virus | 21,563 | <i>Flaviviridae</i> |
| Measles morbillivirus | 17,233 | <i>Paramyxoviridae</i> |

*as of October, 2019.

Viral sequences are not systematically or thoroughly annotated

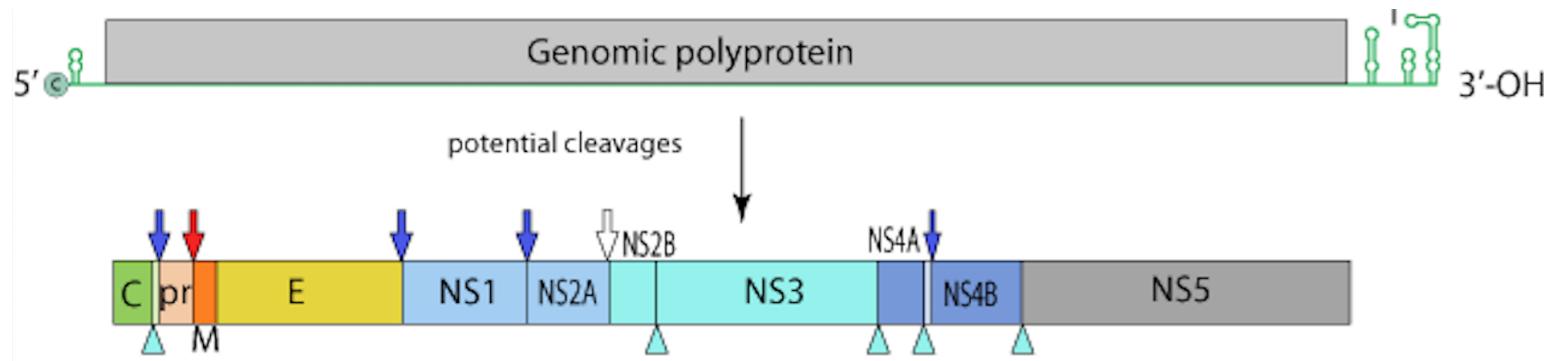
- Genome annotation of the Zika virus:



- Zika's genome encodes a single polyprotein that is cleaved into 14 mature peptides.
- Zika RefSeq annotation (NC_012532) includes CDS and mature peptide annotation.

Viral sequences are not systematically or thoroughly annotated

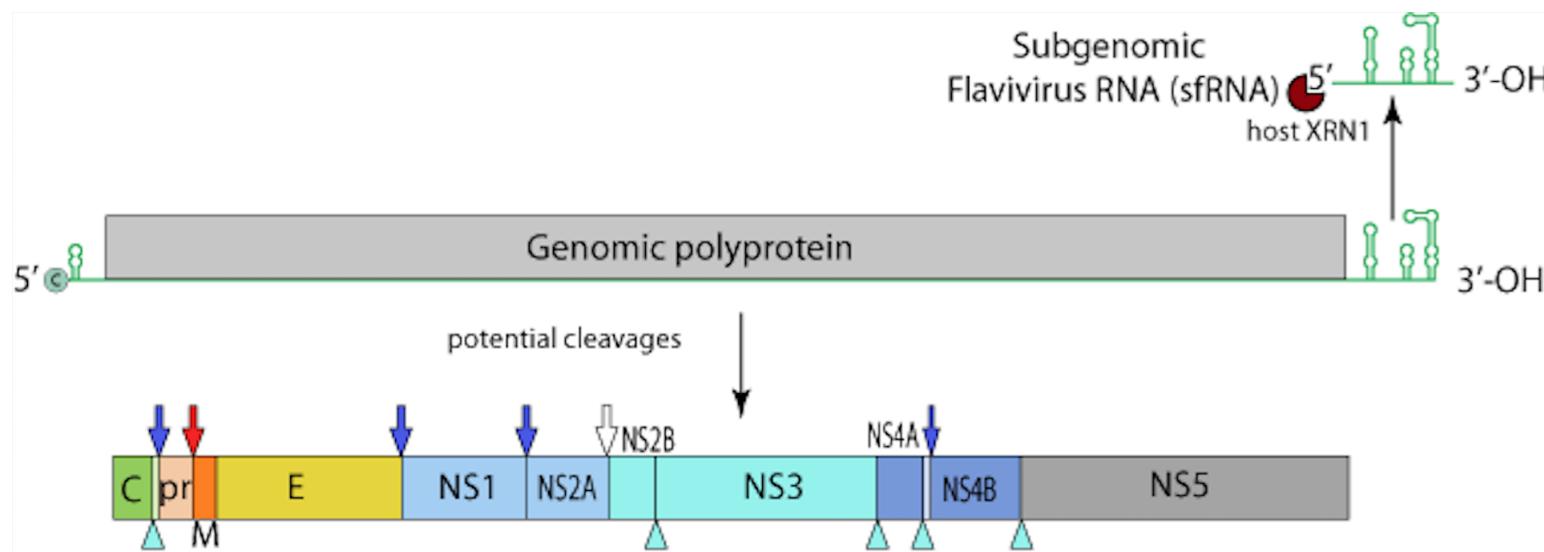
- Genome annotation of the Zika virus:



- Zika's genome encodes a single polyprotein that is cleaved into 14 mature peptides.
- Zika RefSeq annotation (NC_012532) includes CDS and mature peptide annotation.
- About 84% of Zika virus sequences have CDS annotation.
- Less than 25% of Zika virus sequences have mature peptide annotation.
- Less than 7% of Dengue virus sequences have mature peptide annotation.
- Less than 2% of Norovirus sequences have mature peptide annotation.

Viral sequences are not systematically or thoroughly annotated

- Genome annotation of the Zika virus:



- RNA structures in the 3' UTR halt host exonuclease leading to an accumulation of 300-500nt subgenomic flavivirus RNAs (sfRNAs) are related to pathogenicity.

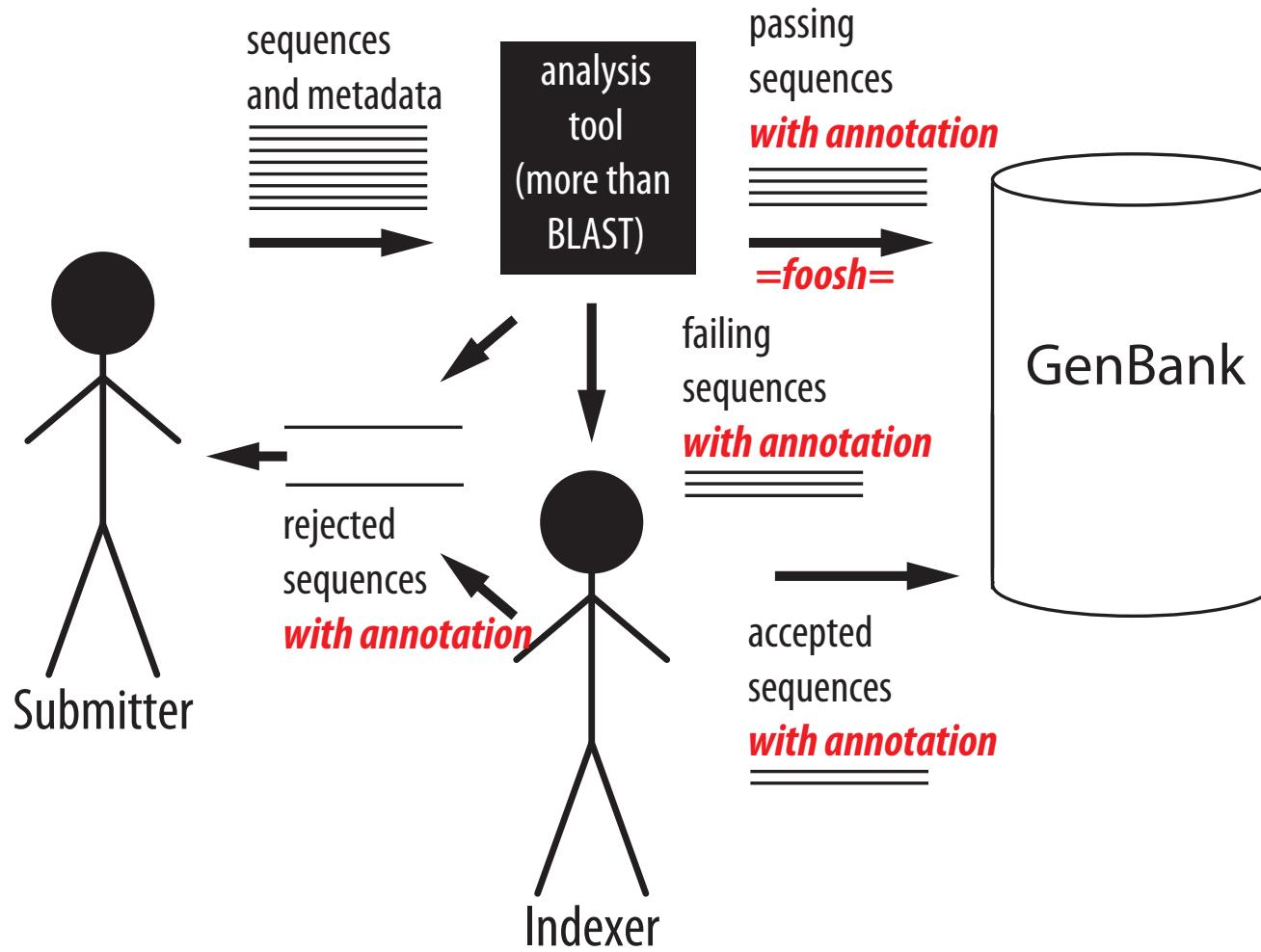
These RNA structures are not annotated in the Zika genome RefSeq (NC_012532)

Viral sequences are not systematically or thoroughly annotated

- CDS are not always annotated
- Mature peptides are rarely annotated
- Rfam families are rarely to never annotated in viral genomes (roughly 200 families)

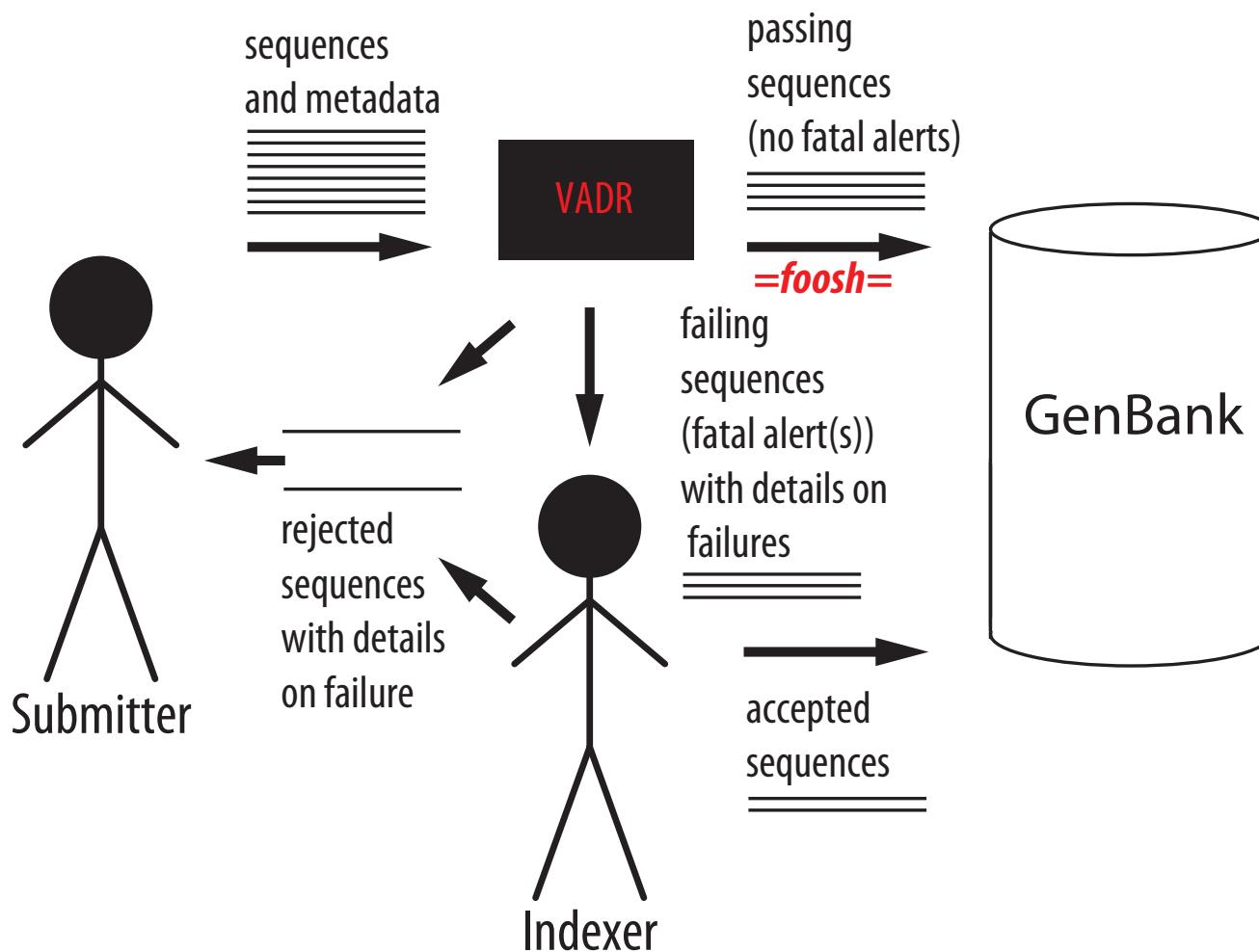
Systematic and complete annotation would benefit viral researchers (facilitate comparative analyses)

Annotation and validation should be coupled



VADR (Viral Annotation DefineR)

uses RefSeqs to validate and annotate viral sequences



- Unexpected characteristics are reported as *alerts* (e.g. early stop codon)
- Some alerts are *fatal* and cause sequences to *fail*

Norovirus and Dengue virus chosen as first viruses for VADR testing

| species | #seqs | family |
|---------------------------|---------|-------------------------|
| HIV-1 | 850,115 | <i>Retroviridae</i> |
| Influenza A virus | 684,026 | <i>Orthomyxoviridae</i> |
| Hepacivirus C | 244,533 | <i>Flaviviridae</i> |
| Hepatitis B virus | 114,306 | <i>Hepadnaviridae</i> |
| Influenza B virus | 100,373 | <i>Orthomyxoviridae</i> |
| Rotavirus A | 73,375 | <i>Reoviridae</i> |
| SIV | 44,374 | <i>Retroviridae</i> |
| Norovirus (Norwalk virus) | 40,925 | <i>Caliciviridae</i> |
| Enterovirus A | 31,478 | <i>Picornaviridae</i> |
| PRRSV | 29,081 | <i>Arteriviridae</i> |
| Dengue virus | 28,564 | <i>Flaviviridae</i> |
| Human orthopneumovirus | 24,384 | <i>Pneumoviridae</i> |
| Enterovirus B | 23,865 | <i>Picornaviridae</i> |
| Rabies lyssavirus | 23,771 | <i>Rhabdoviridae</i> |
| West Nile virus | 21,563 | <i>Flaviviridae</i> |
| Measles morbillivirus | 17,233 | <i>Paramyxoviridae</i> |

VADR build step (v-build.pl) builds a homology model (covariance model (CM)) of a RefSeq and stores feature information

The screenshot shows the NCBI Nucleotide search interface. The search term "Dengue virus 1, complete genome" has been entered into the search field. Below the search results, detailed information about the sequence is provided:

Dengue virus 1, complete genome
NCBI Reference Sequence: NC_001477.1
[FASTA](#) [Graphics](#)

Go to:

| | | | | |
|------------|---|-----------------|--------|-----------------|
| LOCUS | NC_001477 | 10735 bp ss-RNA | linear | VRL 03-MAY-2019 |
| DEFINITION | Dengue virus 1, complete genome. | | | |
| ACCESSION | NC_001477 | | | |
| VERSION | NC_001477.1 | | | |
| DBLINK | BioProject: PRJNA485481 | | | |
| KEYWORDS | RefSeq. | | | |
| SOURCE | Dengue virus 1 | | | |
| ORGANISM | Dengue virus 1 | | | |
| | Viruses; Riboviria; Flaviviridae; Flavivirus. | | | |

VADR build step (v-build.pl) builds a homology model (covariance model (CM)) of a RefSeq and stores feature information

| FEATURES | Location/Qualifiers |
|------------|---|
| source | 1..10735 /organism="Dengue virus 1" /mol_type="genomic RNA" /db_xref="taxon: 11053 " /clone="45AZ5" /type="1" |
| 5' UTR | 1..94 |
| stem loop | 2..69 /note="stem-loop A (SLA)" |
| regulatory | 70..78 /regulatory_class="other" /note="oligo U track spacer" |
| regulatory | 79..94 /regulatory_class="promoter" /note="5' upstream AUG region (UAR)" |
| stem loop | 79..93 /note="stem-loop B (SLB)" |
| gene | 95..10273 /gene="POLY" /locus_tag="DV1_gp1" /gene_synonym="polyprotein gene" /db_xref="GeneID: 5075725 " |
| CDS | 95..10273 /gene="POLY" /locus_tag="DV1_gp1" /gene_synonym="polyprotein gene" /codon_start=1 /product="polyprotein" /protein_id=" NP_059433.1 " /db_xref="GeneID: 5075725 " /translation="MNNQRKKTGRPSFNMLKRARNRVSTVSQLAKRFSKGLLSGQGPM KLVMAFIAFLRFLAIAPP TAGILARWGSFKKNGAIKVLRGFKKEISNMLNIMNRRKRSV |

VADR build step (v-build.pl) builds a homology model (covariance model (CM)) of a RefSeq and stores feature information

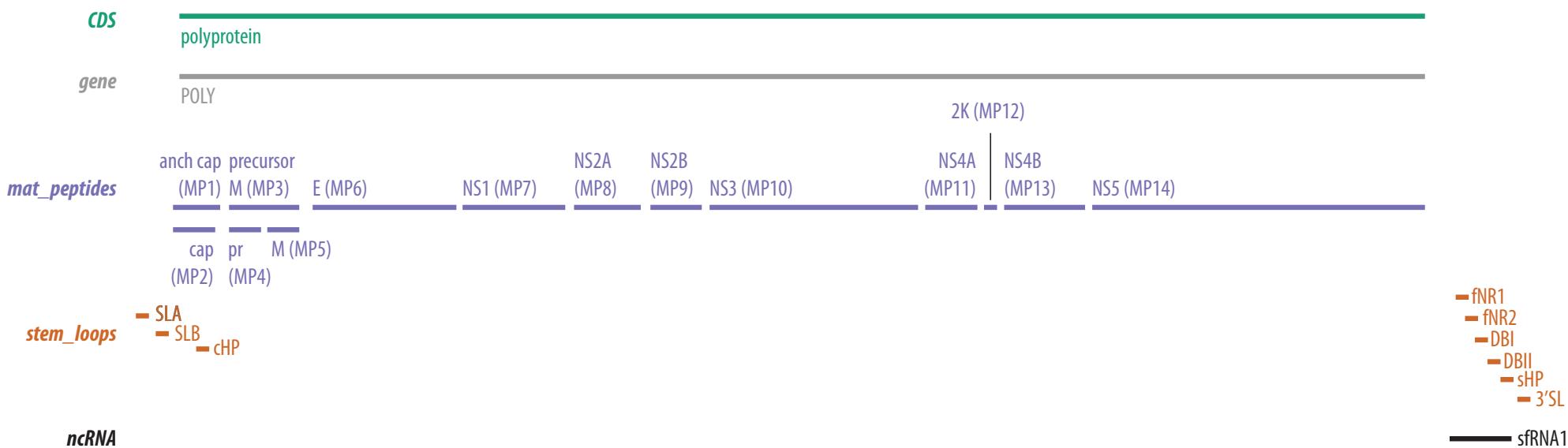
```
mat_peptide    95..436
               /gene="POLY"
               /locus_tag="DV1_gp1"
               /gene_synonym="polyprotein gene"
               /product="anchored capsid protein ancC"
               /protein_id="NP\_722457.2"
               /db_xref="VBRC:35735"

mat_peptide    95..394
               /gene="POLY"
               /locus_tag="DV1_gp1"
               /gene_synonym="polyprotein gene"
               /product="capsid protein C"
               /note="added by NCBI staff following more recent
annotations of this virus sequence"
               /protein_id="NP\_722466.2"
               /db_xref="VBRC:67793"

mat_peptide    437..934
               /gene="POLY"
               /locus_tag="DV1_gp1"
               /gene_synonym="polyprotein gene"
               /product="membrane glycoprotein precursor prM"
               /protein_id="NP\_733807.2"

mat_peptide    437..709
               /gene="POLY"
               /locus_tag="DV1_gp1"
               /gene_synonym="polyprotein gene"
               /product="protein pr"
               /note="peptide pr"
               /protein_id="YP\_009164956.1"
```

VADR build step (v-build.pl) builds a homology model (covariance model (CM)) of a RefSeq and stores feature information



NC_001477 MODEL

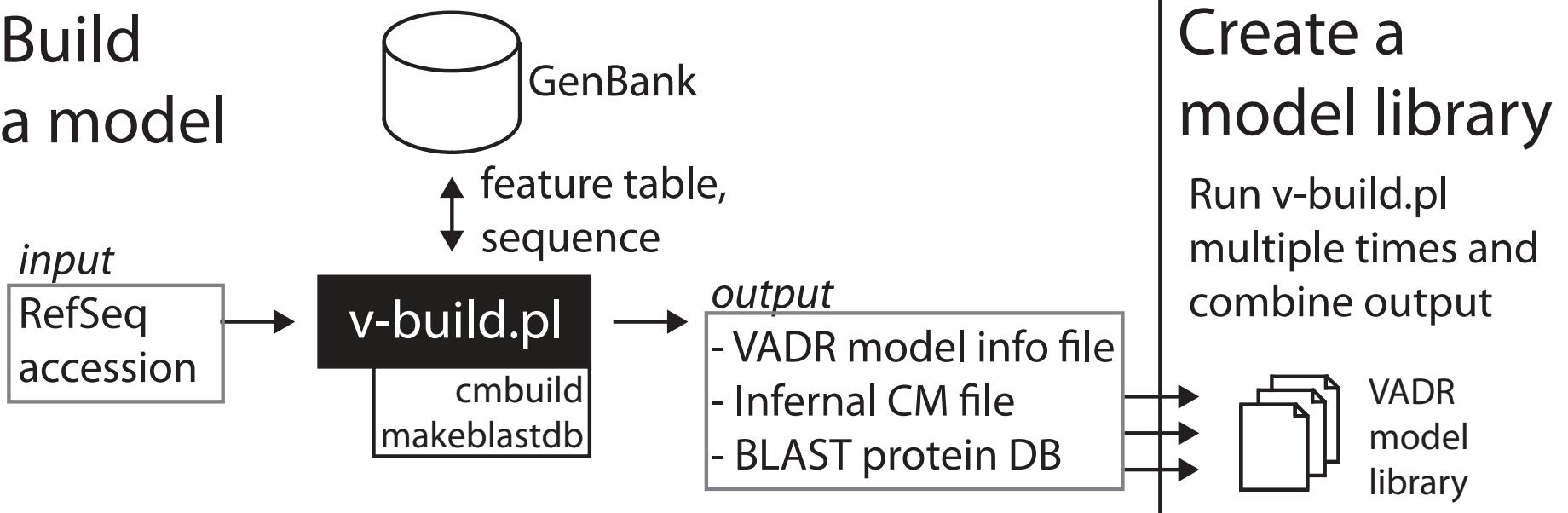


Group: Dengue; Subgroup: 1

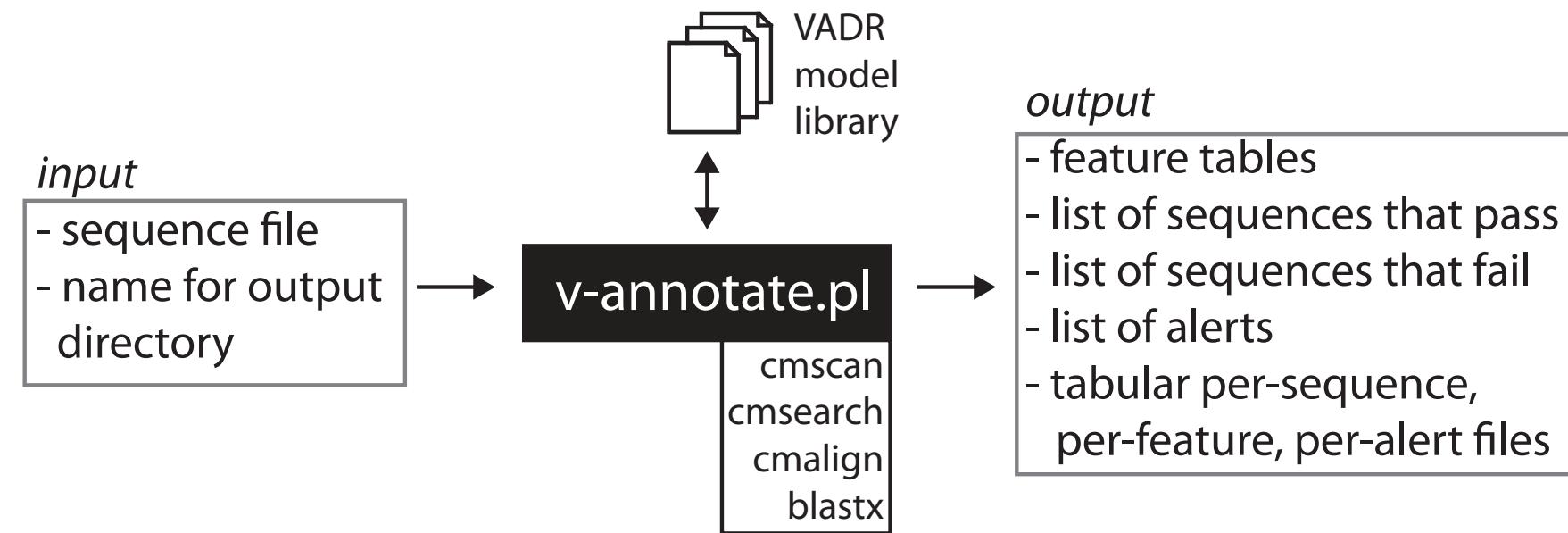
VADR build step (v-build.pl) builds a homology model (covariance model (CM)) of a RefSeq and stores feature information

```
MODEL NC_001477 blastdb:"NC_001477.vadr.protein.fa" cmfile:"NC_001477.vadr.cm" group:"Dengue" length:"10735" subgroup:"1"
FEATURE NC_001477 type:"stem_loop" coords:"2..69:+" parent_idx_str:"GBNULL" note:"stem-loop A (SLA)"
FEATURE NC_001477 type:"stem_loop" coords:"79..93:+" parent_idx_str:"GBNULL" note:"stem-loop B (SLB)"
FEATURE NC_001477 type:"gene" coords:"95..10273:+" parent_idx_str:"GBNULL" gene:"POLY"
FEATURE NC_001477 type:"CDS" coords:"95..10273:+" parent_idx_str:"GBNULL" gene:"POLY" product:"polyprotein"
FEATURE NC_001477 type:"stem_loop" coords:"114..130:+" parent_idx_str:"GBNULL" note:"capsid region hairpin (cHP)"
FEATURE NC_001477 type:"ncRNA" coords:"10320..10735:+" parent_idx_str:"GBNULL" note:"subgenomic flavivirus RNA" product:"sfRNA1" ncRNA_class:"lncRNA"
FEATURE NC_001477 type:"stem_loop" coords:"10324..10389:+" parent_idx_str:"GBNULL" note:"flaviviral nuclease-resistant RNA 1 (fNR1); also called stem-loop 1 or xrRNA1"
FEATURE NC_001477 type:"stem_loop" coords:"10397..10458:+" parent_idx_str:"GBNULL" note:"flaviviral nuclease-resistant RNA 2 (fNR2); also called stem-loop 2 or xrRNA2"
FEATURE NC_001477 type:"stem_loop" coords:"10467..10549:+" parent_idx_str:"GBNULL" note:"dumbbell 1 (DBI); also called xrRNA3"
FEATURE NC_001477 type:"stem_loop" coords:"10551..10633:+" parent_idx_str:"GBNULL" note:"dumbbell 2 (DBII); also called xrRNA4"
FEATURE NC_001477 type:"stem_loop" coords:"10643..10656:+" parent_idx_str:"GBNULL" note:"short hairpin (sHP)"
FEATURE NC_001477 type:"stem_loop" coords:"10657..10735:+" parent_idx_str:"GBNULL" note:"3' stem-loop (3'SL)"
FEATURE NC_001477 type:"mat_peptide" coords:"95..436:+" parent_idx_str:"3" product:"anchored capsid protein ancC"
FEATURE NC_001477 type:"mat_peptide" coords:"95..394:+" parent_idx_str:"3" product:"capsid protein C"
FEATURE NC_001477 type:"mat_peptide" coords:"437..934:+" parent_idx_str:"3" product:"membrane glycoprotein precursor prM"
FEATURE NC_001477 type:"mat_peptide" coords:"437..709:+" parent_idx_str:"3" product:"protein pr"
FEATURE NC_001477 type:"mat_peptide" coords:"710..934:+" parent_idx_str:"3" product:"membrane glycoprotein M"
FEATURE NC_001477 type:"mat_peptide" coords:"935..2419:+" parent_idx_str:"3" product:"envelope protein E"
FEATURE NC_001477 type:"mat_peptide" coords:"2420..3475:+" parent_idx_str:"3" product:"nonstructural protein NS1"
FEATURE NC_001477 type:"mat_peptide" coords:"3476..4129:+" parent_idx_str:"3" product:"nonstructural protein NS2A"
FEATURE NC_001477 type:"mat_peptide" coords:"4130..4519:+" parent_idx_str:"3" product:"nonstructural protein NS2B"
FEATURE NC_001477 type:"mat_peptide" coords:"4520..6376:+" parent_idx_str:"3" product:"nonstructural protein NS3"
FEATURE NC_001477 type:"mat_peptide" coords:"6377..6757:+" parent_idx_str:"3" product:"nonstructural protein NS4A"
FEATURE NC_001477 type:"mat_peptide" coords:"6758..6826:+" parent_idx_str:"3" product:"protein 2K"
FEATURE NC_001477 type:"mat_peptide" coords:"6827..7573:+" parent_idx_str:"3" product:"nonstructural protein NS4B"
FEATURE NC_001477 type:"mat_peptide" coords:"7574..10270:+" parent_idx_str:"3" product:"RNA-dependent RNA polymerase NS5"
```

Build a model



Validate and annotate input sequences



VADR 1.0 model library

- 38 *Caliciviridae* models:
 - 9 Norovirus models
 - 7 Sapovirus models
 - 4 Vesivirus models
- 156 *Flaviviridae* models:
 - 10 Pegivirus models
 - 8 HCV models
 - 7 Pestivirus models
 - 4 Dengue virus models
 - 2 West Nile virus models
 - 2 Zika virus models

`v-annotate.pl` annotates each sequence using its best-matching model

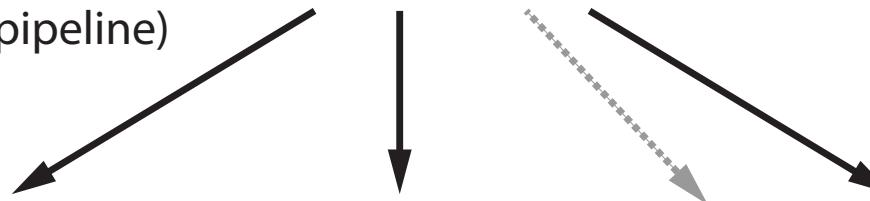
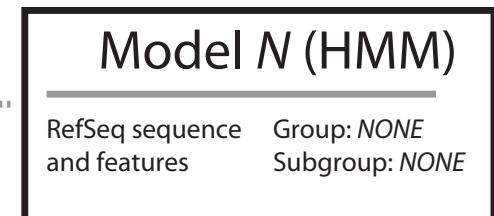
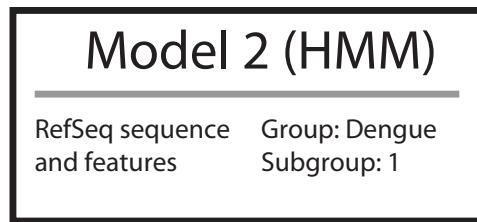
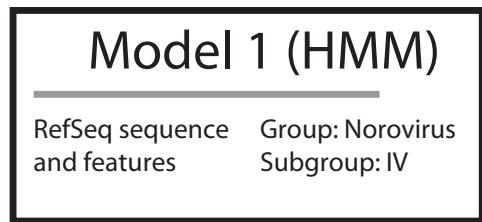
- For each sequence S :
 1. **Classification:** compare S to all models to find best matching model M
 2. **Coverage determination:** search M against S to find 'hits'
 3. **Alignment:** align S to M and map features from M to S
 4. **Protein validation:** compare predicted CDS in S to proteins from M using BLASTX

Different types of alerts are identified and reported at each stage

Stage 1: Classification

Score each sequence
with all models
(HMMER3 shortened pipeline)

input sequences:



low HMM score

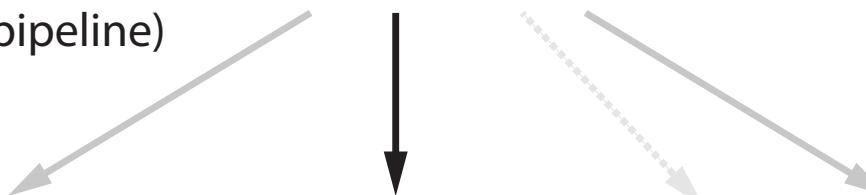
highest HMM score

low HMM score

Stage 1: Classification

Score each sequence
with all models
(HMMER3 shortened pipeline)

input sequences:



Model 1 (HMM)

RefSeq sequence
and features Group: Norovirus
Subgroup: IV

Model 2 (HMM)

RefSeq sequence
and features Group: Dengue
Subgroup: 1

Model N (HMM)

RefSeq sequence
and features Group: NONE
Subgroup: NONE

low HMM score

highest HMM score

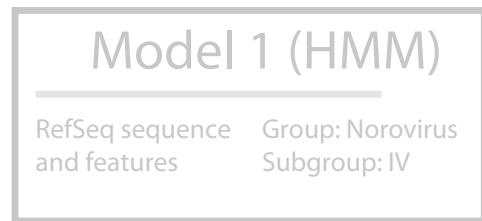
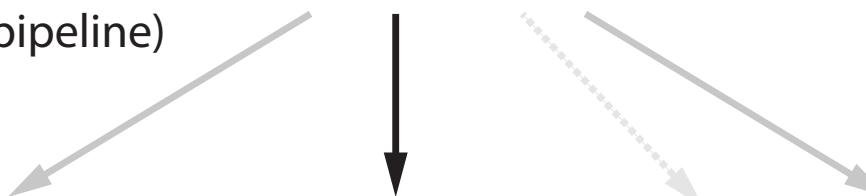
low HMM score

***best-matching model
used in remaining stages***

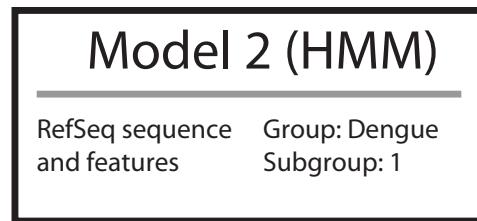
Stage 1: Classification

Score each sequence
with all models
(HMMER3 shortened pipeline)

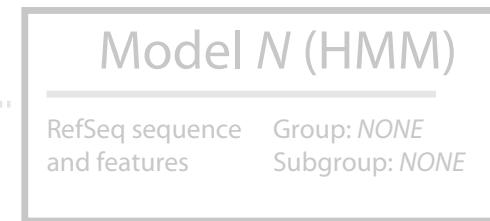
input sequences:



low HMM score



highest HMM score



low HMM score

***best-matching model
used in remaining stages***

| code | S/F | error message | description |
|--|-----|---------------------------------|--|
| Fatal alerts detected in the classification stage | | | |
| noannotn* | S | NO_ANNOTATION | no significant similarity detected |
| revcompl* | S | REVCOMPLEM | sequence appears to be reverse complemented |
| incsbgrp | S | INCORRECT_SPECIFIED_SUBGROUP | score difference too large between best overall model and best specified subgroup model |
| incgroup | S | INCORRECT_SPECIFIED_GROUP | score difference too large between best overall model and best specified group model |
| Non-fatal alerts detected in the classification stage | | | |
| qstsbgp | S | QUESTIONABLE_SPECIFIED_SUBGROUP | best overall model is not from specified subgroup |
| qstgroup | S | QUESTIONABLE_SPECIFIED_GROUP | best overall model is not from specified group |
| indfclas | S | INDEFINITE_CLASSIFICATION | low score difference between best overall model and second best model (not in best model's subgroup) |
| lowscore | S | LOW_SCORE | score to homology model below low threshold |

Stage 2: Coverage determination

Search each sequence with best-matching model (HMMER3 full pipeline)

input sequences that match best to NC_001477:

S1 _____
S2 _____
S3 _____
S4 _____



NC_001477 MODEL



Group: Dengue; Subgroup: 1



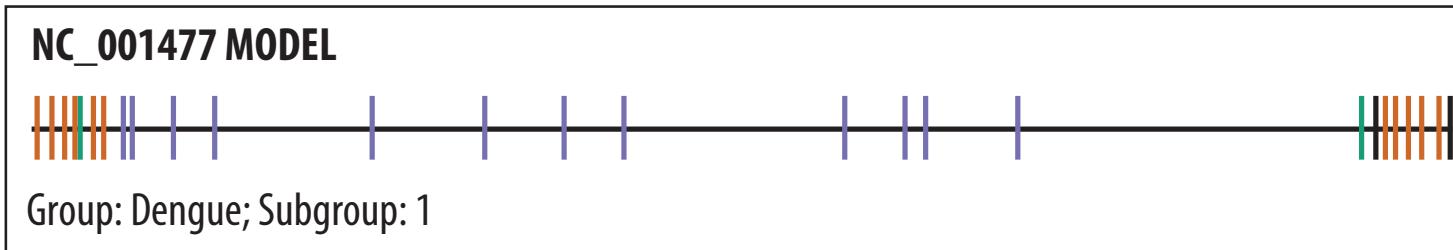
NC_001477 full length sequence
S1 (expected)
NC_001477 partial or truncated sequence
S2 (expected)

Stage 2: Coverage determination

Search each sequence with best-matching model (HMMER3 full pipeline)

input sequences that match best to NC_001477:

S1 _____
S2 _____
S3 _____
S4 _____

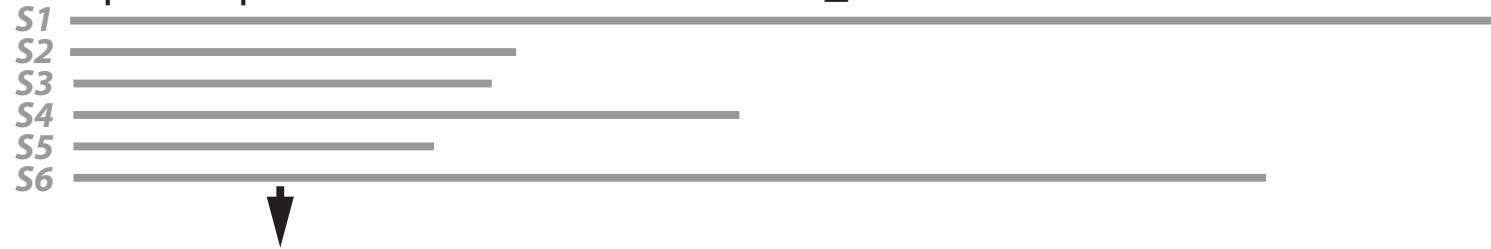


| code | S/F | error message | description |
|--|-----|--------------------------|---|
| Fatal alerts detected in the coverage stage | | | |
| lowcovrg | S | LOW_COVERAGE | low sequence fraction with significant similarity to homology model |
| dupregin | S | DUPLICATE_REGIONS | similarity to a model region occurs more than once |
| discontn | S | DISCONTINUOUS_SIMILARITY | not all hits are in the same order in the sequence and the homology model |
| indfstrn | S | INDEFINITE_STRAND | significant similarity detected on both strands |
| lowsim5s | S | LOW_SIMILARITY_START | significant similarity not detected at 5' end of the sequence |
| lowsim3s | S | LOW_SIMILARITY_END | significant similarity not detected at 3' end of the sequence |
| lowsimis | S | LOW_SIMILARITY | internal region without significant similarity |
| Non-fatal alerts detected in the coverage stage | | | |
| biasdseq | S | BIASED_SEQUENCE | high fraction of score attributed to biased sequence composition |

Stage 3: Alignment and feature mapping

Align each sequence to its best-matching model (Infernal's cmalign)

input sequences that match best to NC_001477:



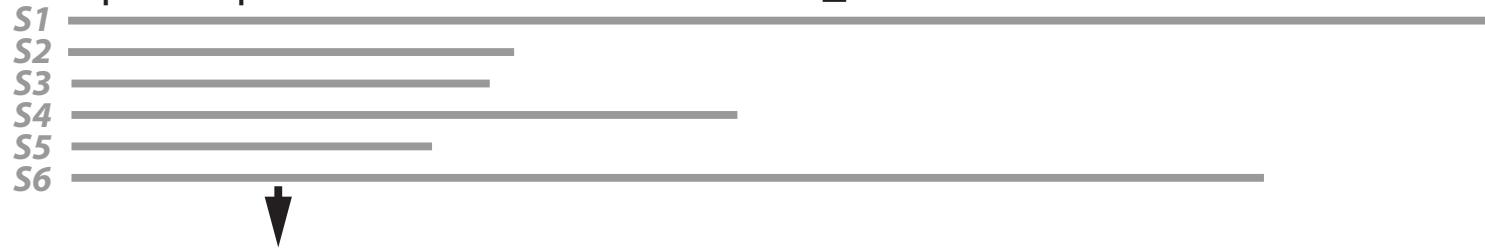
NC_001477 MODEL



Stage 3: Alignment and feature mapping

Align each sequence to its best-matching model (Infernal's cmalign)

input sequences that match best to NC_001477:



NC_001477 MODEL



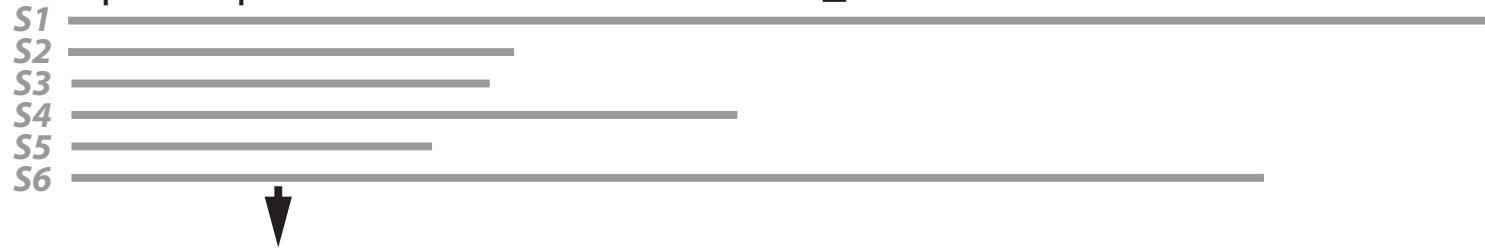
Group: Dengue; Subgroup: 1



Stage 3: Alignment and feature mapping

Align each sequence to its best-matching model (Infernal's cmalign)

input sequences that match best to NC_001477:



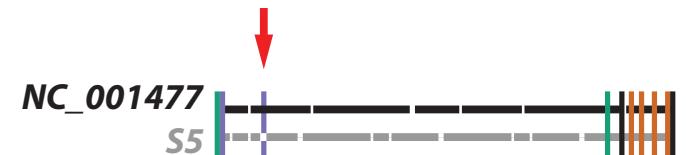
NC_001477 MODEL



Group: Dengue; Subgroup: 1



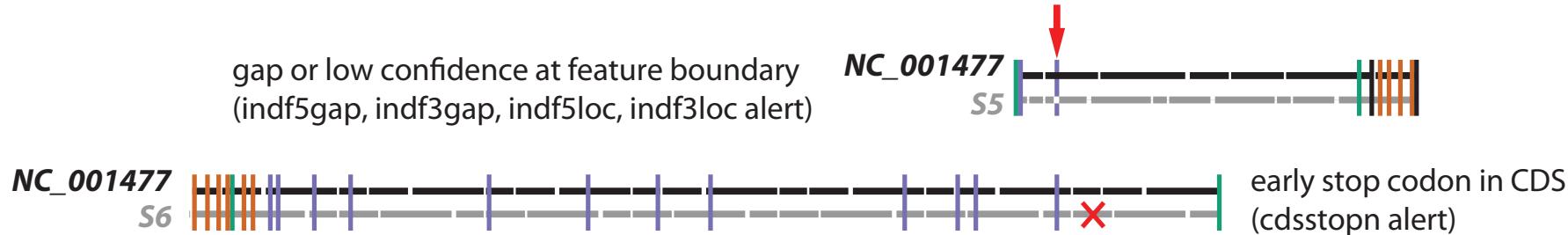
gap or low confidence at feature boundary
(indf5gap, indf3gap, indf5loc, indf3loc alert)



early stop codon in CDS
(cdsstopn alert)

Stage 3: Alignment and feature mapping

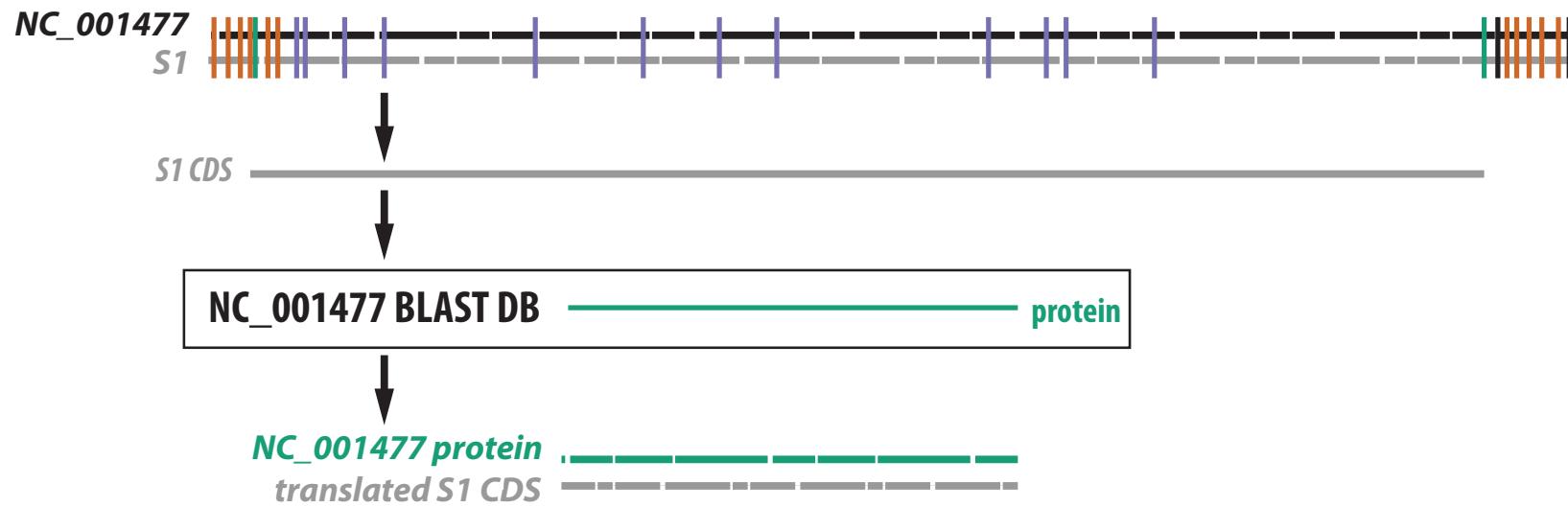
Align each sequence to its best-matching model (Infernal's cmalign)



| code | S/F | error message | description |
|--|-----|------------------------------|--|
| Fatal alerts detected in the annotation stage | | | |
| unexdivg* | S | UNEXPECTED_DIVERGENCE | sequence is too divergent to confidently assign nucleotide-based annotation |
| noftrann* | S | NO_FEATURES_ANNOTATED | sequence similarity to homology model does not overlap with any features |
| mutstart | F | MUTATION_AT_START | expected start codon could not be identified |
| mutendcd | F | MUTATION_AT_END | expected stop codon could not be identified, predicted CDS stop by homology is invalid |
| mutendns | F | MUTATION_AT_END | expected stop codon could not be identified, no in-frame stop codon exists 3' of predicted valid start codon |
| mutendex | F | MUTATION_AT_END | expected stop codon could not be identified, first in-frame stop codon exists 3' of predicted stop position |
| unexleng | F | UNEXPECTED_LENGTH | length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 |
| cdsstoppn | F | CDS_HAS_STOP_CODON | in-frame stop codon exists 5' of stop position predicted by homology to reference |
| peptrans | F | PEPTIDE_TRANSLATION_PROBLEM | mat_peptide may not be translated because its parent CDS has a problem |
| pepadjcy | F | PEPTIDE_ADJACENCY_PROBLEM | predictions of two mat_peptides expected to be adjacent are not adjacent |
| indfantn | F | INDEFINITE_ANNOTATION | nucleotide-based search identifies CDS not identified in protein-based search |
| indf5gap | F | INDEFINITE_ANNOTATION_START | alignment to homology model is a gap at 5' boundary |
| indf5loc | F | INDEFINITE_ANNOTATION_START | alignment to homology model has low confidence at 5' boundary |
| indf3gap | F | INDEFINITE_ANNOTATION_END | alignment to homology model is a gap at 3' boundary |
| indf3loc | F | INDEFINITE_ANNOTATION_END | alignment to homology model has low confidence at 3' boundary |
| lowsim5f | F | LOW FEATURE SIMILARITY_START | region within annotated feature at 5' end of sequence lacks significant similarity |
| lowsim3f | F | LOW FEATURE SIMILARITY_END | region within annotated feature at 3' end of sequence lacks significant similarity |
| lowsimif | F | LOW FEATURE SIMILARITY | region within annotated feature lacks significant similarity |

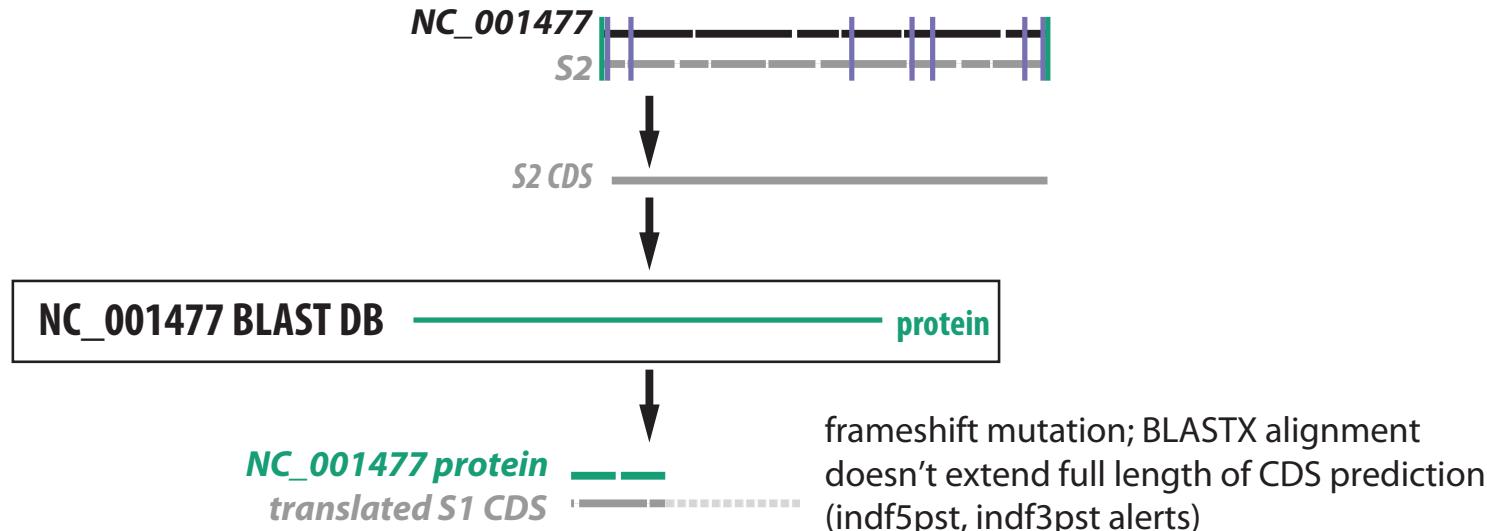
Stage 4: Protein validation

Compare each predicted CDS to model (RefSeq) proteins with BLASTX



Stage 4: Protein validation

Compare each predicted CDS to model (RefSeq) proteins with BLASTX



| code | S/F | error message | description |
|--|-----|-----------------------------|--|
| Fatal alerts detected in the protein validation stage | | | |
| cdsstopp | F | CDS_HAS_STOP_CODON | stop codon in protein-based alignment |
| indfantp | F | INDEFINITE_ANNOTATION | protein-based search identifies CDS not identified in nucleotide-based search |
| indf5plg | F | INDEFINITE_ANNOTATION_START | protein-based alignment extends past nucleotide-based alignment at 5' end |
| indf5pst | F | INDEFINITE_ANNOTATION_START | protein-based alignment does not extend close enough to nucleotide-based alignment 5' endpoint |
| indf3plg | F | INDEFINITE_ANNOTATION_END | protein-based alignment extends past nucleotide-based alignment at 3' end |
| indf3pst | F | INDEFINITE_ANNOTATION_END | protein-based alignment does not extend close enough to nucleotide-based alignment 3' endpoint |
| indfstrp | F | INDEFINITE_STRAND | strand mismatch between protein-based and nucleotide-based predictions |
| insertnp | F | INSERTION_OF_NT | too large of an insertion in protein-based alignment |
| deletinp | F | DELETION_OF_NT | too large of a deletion in protein-based alignment |

VADR results on all Norovirus and Dengue sequences

| dataset | # seqs | min length | max length | # pass | # fail | fraction pass |
|-------------------------|--------|------------|------------|--------|--------|---------------|
| Norovirus complete (NC) | 1,384 | 7380 | 7839 | 1,157 | 227 | 0.836 |
| Dengue complete (DC) | 4,580 | 10372 | 16254 | 4,171 | 409 | 0.911 |
| Norovirus partial (NP) | 32,190 | 50 | 7376 | 29,488 | 2,702 | 0.916 |
| Dengue partial (DP) | 20,973 | 50 | 10370 | 17,276 | 3,697 | 0.824 |

| alert code | error message | NC 1384 seqs ct(seqs) | NP 32190 seqs ct(seqs) | DC 4580 seqs ct(seqs) | DP 20973 seqs ct(seqs) | total 59127 seqs ct(seqs) |
|------------|------------------------------|-----------------------------|------------------------------|-----------------------------|------------------------------|---------------------------------|
| peptrans | PEPTIDE_TRANSLATION_PROBLEM | 516(86) | 716(535) | 1330(95) | 4051(1065) | 6613(1781) |
| noannotn | NO_ANNOTATION | - | 512(512) | 5(5) | 2236(2236) | 2753(2753) |
| indf3pst | INDEFINITE_ANNOTATION_END | 82(70) | 1059(1029) | 56(56) | 600(593) | 1797(1748) |
| indf5pst | INDEFINITE_ANNOTATION_START | 59(57) | 940(876) | 16(16) | 660(574) | 1675(1523) |
| indf3loc | INDEFINITE_ANNOTATION_END | 85(48) | 185(90) | 206(98) | 293(136) | 769(372) |
| incgroup | INCORRECT_SPECIFIED_GROUP | 19(19) | 302(302) | 30(30) | 286(286) | 637(637) |
| indf5loc | INDEFINITE_ANNOTATION_START | 19(15) | 66(35) | 222(135) | 286(144) | 593(329) |
| lowcovrg | LOW_COVERAGE | 3(3) | 217(217) | 60(60) | 279(279) | 559(559) |
| unexleng | UNEXPECTED_LENGTH | 42(34) | 66(55) | 105(49) | 318(182) | 531(320) |
| indf5gap | INDEFINITE_ANNOTATION_START | 6(3) | 23(12) | 117(100) | 220(127) | 366(242) |
| indf3gap | INDEFINITE_ANNOTATION_END | 4(2) | 83(71) | 15(14) | 237(133) | 339(220) |
| lowsim3f | LOW FEATURE SIMILARITY_END | - | - | 272(88) | 20(9) | 292(97) |
| cdsstopp | CDS_HAS_STOP_CODON | 7(5) | 112(111) | 15(15) | 153(153) | 287(284) |
| revcompl | REVCOMPLEM | 3(3) | 85(85) | 35(35) | 120(120) | 243(243) |
| cdsstoppn | CDS_HAS_STOP_CODON | 96(93) | 72(71) | 58(58) | 5(4) | 231(226) |
| insertnp | INSERTION_OF_NT | 50(43) | 151(138) | - | 2(2) | 203(183) |
| lowsim5f | LOW_FEATURE_SIMILARITY_START | - | - | 101(101) | 79(39) | 180(140) |
| lowsim3s | LOW_SIMILARITY_END | 61(61) | 80(80) | 2(2) | 5(5) | 148(148) |
| mutstart | MUTATION_AT_START | 13(11) | 58(58) | 8(8) | 35(27) | 114(104) |
| mutendcd | MUTATION_AT_END | 52(50) | 47(46) | 6(6) | 5(4) | 110(106) |
| discontn | DISCONTINUOUS_SIMILARITY | - | 8(8) | 25(25) | 35(35) | 68(68) |
| dupregin | DUPLICATE_REGIONS | - | 6(6) | 33(33) | 25(25) | 64(64) |
| indfstrn | INDEFINITE_STRAND | 1(1) | 4(4) | 56(56) | 2(2) | 63(63) |
| deletinp | DELETION_OF_NT | 22(20) | 26(25) | - | 12(6) | 60(51) |
| lowsimif | LOW_FEATURE_SIMILARITY | - | - | 29(14) | 18(9) | 47(23) |
| indf3plg | INDEFINITE_ANNOTATION_END | 1(1) | 40(40) | - | 2(2) | 43(43) |
| indfantn | INDEFINITE_ANNOTATION | 1(1) | 23(23) | - | 18(17) | 42(41) |
| lowsim5s | LOW_SIMILARITY_START | 12(12) | - | 6(6) | 20(20) | 38(38) |
| noftrann | NO_FEATURES_ANNOTATED | - | 1(1) | - | 26(26) | 27(27) |
| indf5plg | INDEFINITE_ANNOTATION_START | - | 10(10) | - | - | 10(10) |
| indfantp | INDEFINITE_ANNOTATION | - | 3(3) | - | 6(6) | 9(9) |
| pepadjcy | PEPTIDE_ADJACENCY_PROBLEM | - | 3(3) | - | 6(6) | 9(9) |
| mutendex | MUTATION_AT_END | 2(2) | 5(5) | 1(1) | - | 8(8) |
| mutendns | MUTATION_AT_END | 1(1) | 5(5) | - | - | 6(6) |

SOFTWARE

Open Access



CrossMark

VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank

Ryan C. Shean^{1,2}, Negar Makhsoos^{1,2}, Graham D. Stoddard³, Michelle J. Lin^{1,2} and Alexander L. Greninger^{1,2*} 

- large reference database of nearly all complete genomes in GenBank
- annotates each sequence based on best-match in reference database using blastn
- designed for complete viral genomes
- identifies early and absent stop codons and frameshift indels
- simplifies submission by adding metadata expected by GenBank

SOFTWARE

Open Access

VIGOR, an annotation program for small viral genomes

Shiliang Wang^{1*}, Jaideep P Sundaram^{1,2}, David Spiro^{1*}

- determines most appropriate database (e.g. Norovirus) and annotates based on comparison of all proteins and mature peptides in that database
- no Dengue database
- identifies early and absent stop codons and frameshift indels

Summary of VAPiD, VIGOR and VADR on 200 randomly chosen seqs

| dataset | VADR pass/fail | VAPiD pass/fail | VIGOR pass/fail |
|---------|-------------------|--------------------|--------------------|
| NC | 167/33 | 161/39 | 198/2 |
| DC | 189/11 | 196/4 | - |
| NP | 191/9 | - | 195/5 |
| DP | 163/37 | - | - |

Comparison of VAPiD and VADR

| dataset | Both pass | Both fail | VADR-pass VAPiD-fail | VADR-fail VAPiD-pass |
|---------|-----------|-----------|-------------------------|-------------------------|
| NC | 137 | 9 | 30 | 24 |
| DC | 188 | 3 | 1 | 8 |

Comparison of VAPiD and VADR

| dataset | Both pass | Both fail | VADR-pass VAPiD-fail | VADR-fail VAPiD-pass |
|---------|-----------|-----------|-------------------------|-------------------------|
| NC | 137 | 9 | 30 | 24 |
| DC | 188 | 3 | 1 | 8 |

- 12 that fail both:
 - 9 have an internal stop or another CDS translation problem
 - one has a start codon problem
 - one is reverse complemented
 - one fails because no reference is found
- 31 that fail only VAPiD are:
 - 5' and/or 3' truncated in the first or final CDS
 - otherwise valid according to the VADR and VIGOR results
- Conclusion: VADR finds all problems that VAPiD finds that should be caught.

Comparison of VIGOR and VADR

| dataset | Both pass | Both fail | VADR-pass VIGOR-fail | VADR-fail VIGOR-pass |
|---------|-----------|-----------|-------------------------|-------------------------|
| NC | 167 | 2 | 0 | 31 |
| NP | 191 | 5 | 0 | 4 |

Comparison of VIGOR and VADR

| dataset | Both pass | Both fail | VADR-pass VIGOR-fail | VADR-fail VIGOR-pass |
|---------|-----------|-----------|-------------------------|-------------------------|
| NC | 167 | 2 | 0 | 31 |
| NP | 191 | 5 | 0 | 4 |

- 7 that fail both:
 - 3 have premature stops
 - 2 are reverse complemented
 - one has a frameshift
 - one fails because no similar reference is found
- Conclusion: VADR finds all problems that VIGOR finds that should be caught.

35 sequences in NC, DC, and NP fail only VADR

- These all have issues indexers want to manually review:
 - 16 sequences with early stops compared to closest RefSeqs
 - 12 sequences for which BLASTX alignment does not extend close enough to alignment-based prediction
 - 10 sequences with low similarity to RefSeq at 5' or 3' end of sequence or a feature
 - 7 sequences where 5' or 3' boundary is a gap or not aligned with sufficient confidence (high enough posterior probability)
 - 5 sequences that are expected to be Norovirus but are really Sapovirus
 - 5 sequences with too large of an insertion/deletion in the BLASTX alignment
 - 1 sequence not recognized by any of the *Caliciviridae* or *Flaviviridae* models (a Salivirus from *Picornaviridae* family)

VADR is general

- Now testing on COX1 (cytochrome C oxidase subunit I) sequences
- Models:
 - Started with 9000 vetted COX1 protein sequences
 - Split based on taxonomy and aligned with MUSCLE
 - Derived 43 *alignment-based* models (e.g. porifera, amphibia) covering 5 genetic codes
 - Classification stage compares against 43 models
 - Protein validation stage compares against 9000 proteins
- Goal: extend to other marker genes; building reference sets and alignments is limiting

Limitations

- nucleotide space, not protein space
- RefSeq or alignment must be 'representative'
 - divergent sequences, regions, introns, gene order are problematic

Limitations

- nucleotide space, not protein space
- RefSeq or alignment must be 'representative'
 - divergent sequences, regions, introns, gene order are problematic
- current length limit of model is about 20Kb due to CM alignment memory requirements
- slow
 - Norovirus complete: 30 seconds/sequence (8X/6X slower than VAPiD/VIGOR)
 - Dengue complete: 90 seconds/sequence (20X slower than VAPiD)
 - Norovirus partial: 1.5 seconds/sequence (1.5X slower than VIGOR)
 - Dengue partial: 9 seconds/sequence

Limitations

- nucleotide space, not protein space
- RefSeq or alignment must be 'representative',
 - divergent sequences, regions, introns, gene order are problematic
- current length limit of model is about 20Kb due to CM alignment memory requirements
- slow
 - Norovirus complete: 30 seconds/sequence (8X/6X slower than VAPiD/VIGOR)
 - Dengue complete: 90 seconds/sequence (20X slower than VAPiD)
 - Norovirus partial: 1.5 seconds/sequence (1.5X slower than VIGOR)
 - Dengue partial: 9 seconds/sequence
- 190 CPU hours for all Norovirus and Dengue seqs (1 day on 8 core machine)

Future directions

- alignment-based models for viruses (testing with HCV)
- profile-based protein validation to replace BLASTX
- allow scanning of large sequences (e.g. genomes)
- extend to more genes, including ribosomal RNAs and possible ITS sequences

Acknowledgements

NCBI - viral annotation

Alejandro Schäffer
Rodney Brister
Ilene Mizrachi
Eneida Hatcher
Linda Yankie
Alex Kotliarov
Susan Schafer
Lara Shonkwiler
Sophia Hu

NCBI - leadership

David Landsman
Jim Ostell
David Lipman

HMMER

Sean Eddy
Travis Wheeler

