流感病毒基因组注释帮助文档

输入文件：

Fasta 序列

```
>EXAMPLE
ATGGAAGACCTTGTGCGACAATGCTTCAATCCGATGATCGTCGAGCTTGCGGAAAA
GGCAATGAAAGAAT
ATGGGGAAGATCCGAAAATCGAAACAAACAAGTTCGCATCAATATGTACACATTTAG
AAGTCTGCTTCAT
GTATTCTGATTTCCACTTCATAGACGAACGAGGTGAATCAACTATTATTGAATCTGGC
GATCCAAATGTG
TTGTTGAAACATCGATTTGAAATAATCGAAGGGAGAGACCGAACAATGGCTTGGAC
AGTGGTGAATAGTA
TTTGCAACACCACGGGTGTCGAAAAACCTAAATTTCTCCCTGATCTGTATGACTACA
AGGAAAACCGATT
CATTGAAATTGGAGTGACAAGGAGGGAAGTCCACATATATTACCTAGAGAAAGCTAA
TAAAATAAAATCC
GAGAAAACACACATACACATTTTCTCATTCACTGGAGAAGAAATGGCCACCAAAGC
AGATTATACTCTTG
ATGAAGAAAGCAGGGCAAGAATCAAAACCAGGCTGTTCACCATAAGGCAGGAGAT
GGCTAGCAGGGGTCT
ATGGGATTCCTTTCGTCAGTCCGAAAGAGGCGAAGAAACAATTGAAGAAAGATTTG
AAATCACAGGAACC
ATGCGCAGGCTTGCCGACCAAAGTCTCCCACCGAACTTCTCCAGCCTTGAAAACTT
TAGAGCCTATGTGG
ATGGATTCGAACCGAACGGCTGCATTGAGGGCAAGCTTTCTCAAATGTCAAAAGAA
```

GTGAACGCCCGGAT

CGAGCCATTTCTAAAGACAACACCACGCCCGCTCAGATTGCCTAATGGGACTCCCTG

TTCTCAGCGGTCG

AAATTCTTGCTGATGGATGCTTTAAAATTAAGCATTGAAGACCCAAGCCACGAAGGG

GAGGGGATACCGC

TATATGATGCGATCAAATGCATGAAAACGTTCTTCGGGTGGAAAGAGCCCAACATTA

TCAAACCACATGA

GAAGGGAATAAACCCAAACTATCTCCTTACTTGGAAGCAGGTGCTGTCAGAACTTC

AGGACATTGAAAAT

GAAGAGAAGATCCCAAGGACAAAAAACATGAAGAAGACAAGCCAATTAAAGTGGG

CACTTGGTGAGAACA

TGGCACCGGAGAAGGTGGACTTTGAGGATTGTAAAGATGTCAACGACTTGAAACAG

TATGACAGTGAAGA

GCCGGAGCCCAGATCAATAGCATGTTGGATCCAAAATGAATTCAACAAGGCATGTGA

ATTGACCGACTCA

AGCTGGGTAGAACTTGATGAAATAGGGGAAGATGTTGCCCCAATCGAACACATTGC

AAGCATGAGAAGGA

ACTACTTTACAGCAGAGGTATCCCACTGCAGGGCTACTGAATACATAATGAAGGGAG

TGTACATAAATAC

AGCTTTGCTCAATGCATCTTGTGCAGCCATGGATGATTTTCAACTGATTCCAATGATA

AGTAAATGCAGA

ACCAAAGAGGGAAGACGTAAAACAAACCTATATGGATTCATTATAAAAGGAAGATC

CCATTTGAGGAATG

ATACCGATGTGGTGAACTTTGTAAGTATGGAGTTTTCCCTTACCGACCCAAGGTTGG

AACCACATAAATG

GGAAAAGTATTGTGTTCTTGAAATAGGGGATATGCTCCTGCGAACGGCAGTAGGCCA

AGTGTCAAGACCC

ATGTTTCTGTATGTGAGAACTAATGGGACCTCCAAGATCAAGATGAAATGGGGTATG

GAAATGAGACGTT

GCCTTCTCCAGTCTCTCCAACAGATTGAGAGTATGATTGAAGCTGAATCCTCCGTCA
AAGAGAAAGACCT
AACTAAAGAATTCTTTGAAAACAAATCAGAAACATGGCCAATTGGAGAATCACCTA
AAAGGGTGGAGGAA
GGTTCCATTGGGAAGGTGTGCAGAACCTTACTAGCAAAATCTGTATTCAACAGCTTA
TATGCATCTCCGC
AACTCGAGGGATTCTCAGCTGAATCAAGAAAACTGCTACTCATTGTCCAGGCGCTTA
GGGATAACCTGGA
ACCTGGAACCTTCGATCTGGAGGGGCTATATGAAGCAATCGAGGAGTGCCTGATTAA
TGATCCCTGGGTT
TTGCTTAATGCATCTTGGTTCAACTCCTTCCTCACACATGCACTAAGATAG

输入 Fasta 文件注意事项：

1. 文件支持.fa,.fsa,.fas,.fasta 格式。
2. 请以纯文本文件的形式上传 Fasta 文件。
3. 请使用 FASTA 格式，以定义行（Definition line）开始，然后是序列行。
4. 最简单的定义行需要“>”符号和一个序列标识符（Sequence ID）。
5. Sequence ID 命名要求：
    a) 以字母开头，建议用单位的缩写（比如 QHCDC），避免重复；
    b) 可以包含字母、数字、横线“-”和下划线“_”；
    c) Sequence ID 的长度需要小于 23 个字符；
6. 输入的 Fasta 序列中，不能存在“-”或者“*”等非法字符。
7. 序列长度须大于 50 碱基，且未知碱基 Ns <50%。
8. **输入序列两端如果包含未知碱基 Ns，输出序列中两端的 Ns 会被自动删除。**

9. Fasta 文件可包含一条或多条序列，且多条序列的 Sequence ID 必须具有唯一性。

输出文件：

文件类型	文件名后缀	描述
注释成功	*.vadr.pass.list	注释成功序列 ID 列表
	*.vadr.pass.fa	注释成功序列文件（.fasta 格式）
	*.vadr.pass.tbl	注释成功序列的注释结果（.tbl 格式，详见“注释结果 TBL 文件”部分）
注释失败	*.vadr.fail.list	注释失败序列 ID 列表
	*.vadr.fail.fa	注释失败序列文件（.fasta 格式）
	*.vadr.fail.tbl	注释失败序列的注释结果（.tbl 格式，详见“注释结果 TBL 文件”部分）和注释错误提示说明（详见“输出错误描述”部分）

注：输入序列两端如果包含未知碱基 Ns，输出序列文件中两端的 Ns 会被自动删除。

输出错误描述

在输出的*.vadr.fail.tbl 文件中，包含注释失败序列的注释结果（.tbl 格式，详见“注释结果 TBL 文件”部分）和错误信息提示说明。按照 43 种错误类型的属性，表 1 中列出了 43 种错误类型，并对其进行了简要说明。其中还需要特别注意：

1. 以下报错信息仅是告知用户有关序列的非常规的、意外的或其他显著特征。

2. 43 种错误类型中有 38 种是致命的，因为它们导致序列注释失败（fail），而另外 5 种报告

的错误类型不是致命的。当且仅当没有致命错误产生时，序列才会被标记为通过（pass）。

3. S/F 列指示错误适用于整个序列（S）或序列中的一个特征（F）。

表1.注释中43种错误类型属性说明

Alerts types	S/F	Error message	description
Fatal alerts detected in the classification stage	S	NO_ANNOTATION	no significant similarity detected
	S	REVCOMPLEM	sequence appears to be reverse complemented
	S	INCORRECT_SPECIFIED_SUBGROUP	score difference too large between best overall model and best specified subgroup model
	S	NCORRECT_SPECIFIED_GROUP	score difference too large between best overall model and best specified group model
Non-fatal alerts detected in the classification stage	S	QUESTIONABLE_SPECIFIED_SUBGROUP	best overall model is not from specified
	S	QUESTIONABLE_SPECIFIED_GROUP	best overall model is not from specified group
	S	INDEFINITE_CLASSIFICATION	low score difference between best overall model and second best model (not in best model's subgroup)
	S	LOW_SCORE	score to homology model below low threshold
Fatal alerts detected in the coverage stage	S	LOW_COVERAGE	low sequence fraction with significant similarity to homology model
	S	DUPLICATE_REGIONS	similarity to a model region occurs more than once
	S	DISCONTINUOUS_SIMILARITY	not all hits are in the same order in the sequence and the homology model
	S	INDEFINITE_STRAND	significant similarity detected on both strands
	S	LOW_SIMILARITY_START	significant similarity not detected at 5' end of the sequence
	S	LOW_SIMILARITY_END	significant similarity not detected at 3' end of the sequence
	S	LOW_SIMILARITY	internal region without significant similarity
Non-fatal alerts detected in the coverage stage	S	BIASED_SEQUENCE	high fraction of score attributed to biased sequence composition
Fatal alerts detected in the annotation stage	S	UNEXPECTED_DIVERGENCE	sequence is too divergent to confidently assign nucleotide-based annotation
	S	NO_FEATURES_ANNOTATED	sequence similarity to homology model

			does not overlap with any features
	F	MUTATION_AT_START	expected start codon could not be identified
	F	MUTATION_AT_END	expected stop codon could not be identified, predicted CDS stop by homology is invalid
	F	MUTATION_AT_END	expected stop codon could not be identified, no in-frame stop codon exists 3’ of predicted valid start codon
	F	MUTATION_AT_END	expected stop codon could not be identified, first in-frame stop codon exists 3’ of predicted stop position
	F	UNEXPECTED_LENGTH	length of complete coding (CDS or mat_peptide) feature is not a multiple of 3
	F	CDS_HAS_STOP_CODON	in-frame stop codon exists 5’ of stop position predicted by homology to reference
	F	PEPTIDE_TRANSLATION_PROBLEM	mat_peptide may not be translated because its parent CDS has a problem
	F	PEPTIDE_ADJACENCY_PROBLEM	predictions of two mat_peptides expected to be adjacent are not adjacent
	F	INDEFINITE_ANNOTATION	nucleotide-based search identifies CDS not identified in protein-based search
	F	INDEFINITE_ANNOTATION_START	alignment to homology model is a gap at 5’ boundary
	F	INDEFINITE_ANNOTATION_START	alignment to homology model has low confidence at 5’ boundary
	F	INDEFINITE_ANNOTATION_END	alignment to homology model is a gap at 3’ boundary
	F	INDEFINITE_ANNOTATION_END	alignment to homology model has low confidence at 3’ boundary similarity
	F	LOW_FEATURE_SIMILARITY_START	region within annotated feature at 5’ end of sequence lacks significant
	F	LOW_FEATURE_SIMILARITY_END	region within annotated feature at 3’ end of sequence lacks significant similarity
	F	LOW_FEATURE_SIMILARITY	region within annotated feature lacks significant similarity
Fatal alerts detected in the protein validation stage	F	CDS_HAS_STOP_CODON	stop codon in protein-based alignment
	F	INDEFINITE_ANNOTATION	protein-based search identifies CDS not identified in nucleotide-based search
	F	INDEFINITE_ANNOTATION_START	protein-based alignment extends past

			nucleotide-based alignment at 5’ end
	F	INDEFINITE_ANNOTATION_START	protein-based alignment does not extend close enough to nucleotide-based alignment 5’ endpoint
	F	INDEFINITE_ANNOTATION_END	protein-based alignment extends past nucleotide-based alignment at 3’ end
	F	INDEFINITE_ANNOTATION_END	protein-based alignment does not extend close enough to nucleotide -based alignment 3’ endpoint
	F	INDEFINITE_STRAND	strand mismatch between protein-based and nucleotide-based predictions
	F	INSERTION_OF_NT	too large of an insertion in protein-based alignment
	F	DELETION_OF_NT	too large of a deletion in protein-based alignment

注释结果 TBL 文件：

TBL 格式文件包含以制表符分隔的五列特征表，允许对不同类型的特征（例如 gene, mRNA, coding region, tRNA）和限定符（例如/product, /note）进行标注。有效的特征和限定符仅限于国际核苷酸序列数据库合作组织（International Nucleotide Sequence Database Collaboration，INSDC）批准的特征和限定符。以制表符分隔的五列特征表规定了每个特征的位置和类型。

文件包含两大部分的信息

（1）注释的基因组序列 SeqID 信息，序列标识符（SeqID）必须与序列文件中使用的标识符保持一致，以“>Feature ”开头，具体表征如下：

>Feature SeqId

（2）紧随其后的行列出了注释特征，每个特征都在单独的一行上，描述该特性的限定符在下一行，每列由制表符分隔。

Column 1: Start location of feature Column 2: Stop location of feature

Column 3: Feature key

Line2:

Column 4: Qualifier key

Column 5: Qualifier value

· **TBL 格式文件中的一些特殊情况说明：**

（1） 部分/不完整特征的位置在核苷酸位置前面用“>”或“<”表示。“<”符号总是出现在第 1 列中，而“>”总是出现在第 2 列中，无论特征是否存在。例如，gene、CDS 和 mRNA 都从核苷酸序列的上游开始，并在核苷酸序列末端的下游结束。其中“<”符号表示它们是 5'端部分/不完整特征，“>”符号表示它们是 3'端部分/不完整特征。

```
>Feature Sc_16
1     7000      REFERENCE
                      PubMed   8849441
<1    >1050     gene
                      gene ATH1
<1    >1009     CDS
                      product    acid trehalase
                      product    Ath1p
                      codon_start     2
<1    >1050     mRNA
                      product    acid trehalase
```

（2） 如果一个特征包含多个间隔，则每个间隔在随后的限定符之前按其起始和终止位置列在单独的行上。

（3）当 TBL 文件中第一列上的特征起始位置值大于第二列特征终止位置值时，这种情况属于注释特征存在于互补链上（即 strand=`-`）的特征坐标表示方式。

（4）示例中的 CDS 特征 protein_id 被标记为 EXAMPLE_1 和 EXAMPLE_2 分别表示同一基因转录生成的不同转录本对应的翻译区，以_1 和_2 做区分。

· **输出 TBL 格式文件示例：**

```
>Feature EXAMPLE
1     2151       gene
                 gene PA
1     760  gene
                 gene PA-X
1     2151       CDS
                 product    polymerase PA
                 protein_id       EXAMPLE_1
1     570  CDS
572  760
                 product    PA-X protein
                 exception ribosomal slippage
                 protein_id       EXAMPLE_2
```