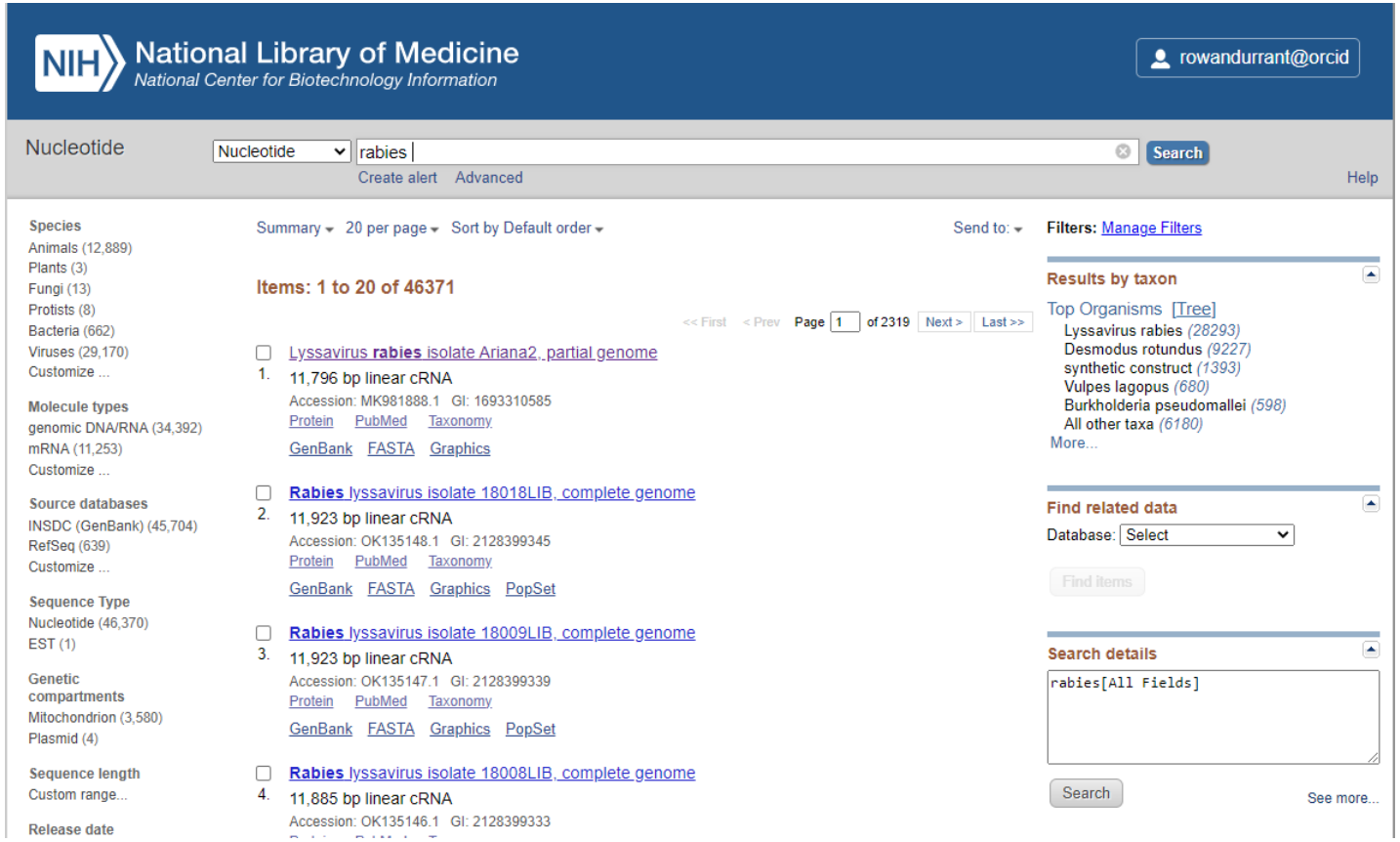
# How to submit sequences to GenBank

using the BankIt website (other methods are available)

# What is GenBank?



# Why should I submit my sequences?

Get accession numbers

Most journals require it

Allows other researchers to use them in their work

JOURNAL ARTICLE

# NCBI-compliant genome submissions: tips and tricks to save time and money

Walter Pirovano, Marten Boetzer, Martijn F L Derks, Sandra Smit

*Briefings in Bioinformatics*, Volume 18, Issue 2, March 2017, Pages 179–182,
https://doi.org/10.1093/bib/bbv104

**Published:** 10 December 2015 **Article history**

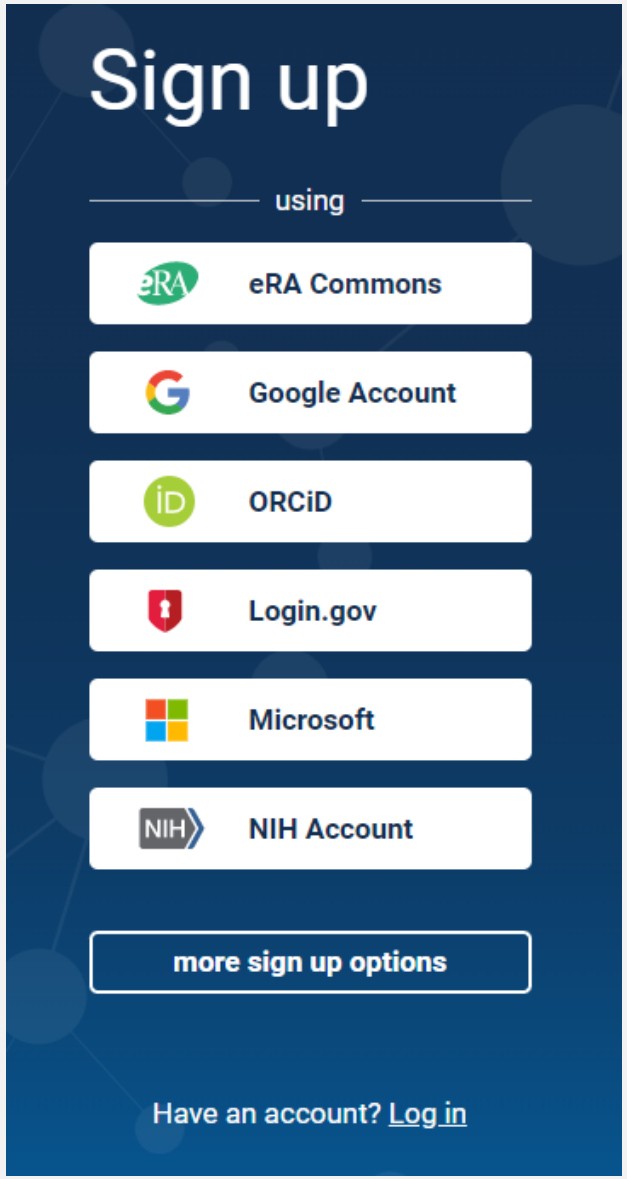
PDF Split View Cite Permissions Share

## Abstract

Genome sequences nowadays play a central role in molecular biology and bioinformatics. These sequences are shared with the scientific community through sequence databases. The sequence repositories of the International Nucleotide Sequence Database Collaboration (INSDC, comprising GenBank, ENA and DDBJ) are the largest in the world. Preparing an annotated sequence in such a way that it will be accepted by the database is challenging because many validation criteria apply. In our opinion, it is an undesirable situation that researchers who want to submit their sequence need either a lot of experience or help from partners to get the job done. To save valuable time and money, we list a number of recommendations for people who want to submit an annotated genome to a sequence database, as well as for tool developers, who could help to ease the process.

# How do I submit my sequences?

- Online is easiest: [www.ncbi.nlm.nih.gov/WebSub/](www.ncbi.nlm.nih.gov/WebSub/)
- You will need to make an account
- You can log in through ORCiD, a Google account, Microsoft etc



Step 1: What kind of sequence do you have?

# Submit new sequences to GenBank

## What type of sequence data do you have?

- ○ SARS-CoV-2
- ○ Ribosomal RNA (rRNA) or rRNA-ITS
- ○ Metazoan (multicellular animal) COX1
- ○ Eukaryotic nuclear mRNA **NEW**
- ○ Influenza virus
- ○ Norovirus
- ○ Dengue virus
- ○ Eukaryotic and Prokaryotic Genomes (WGS or Complete)
- ○ Transcriptome Shotgun Assembly (TSA)
- ○ Unassembled sequence reads (SRA)

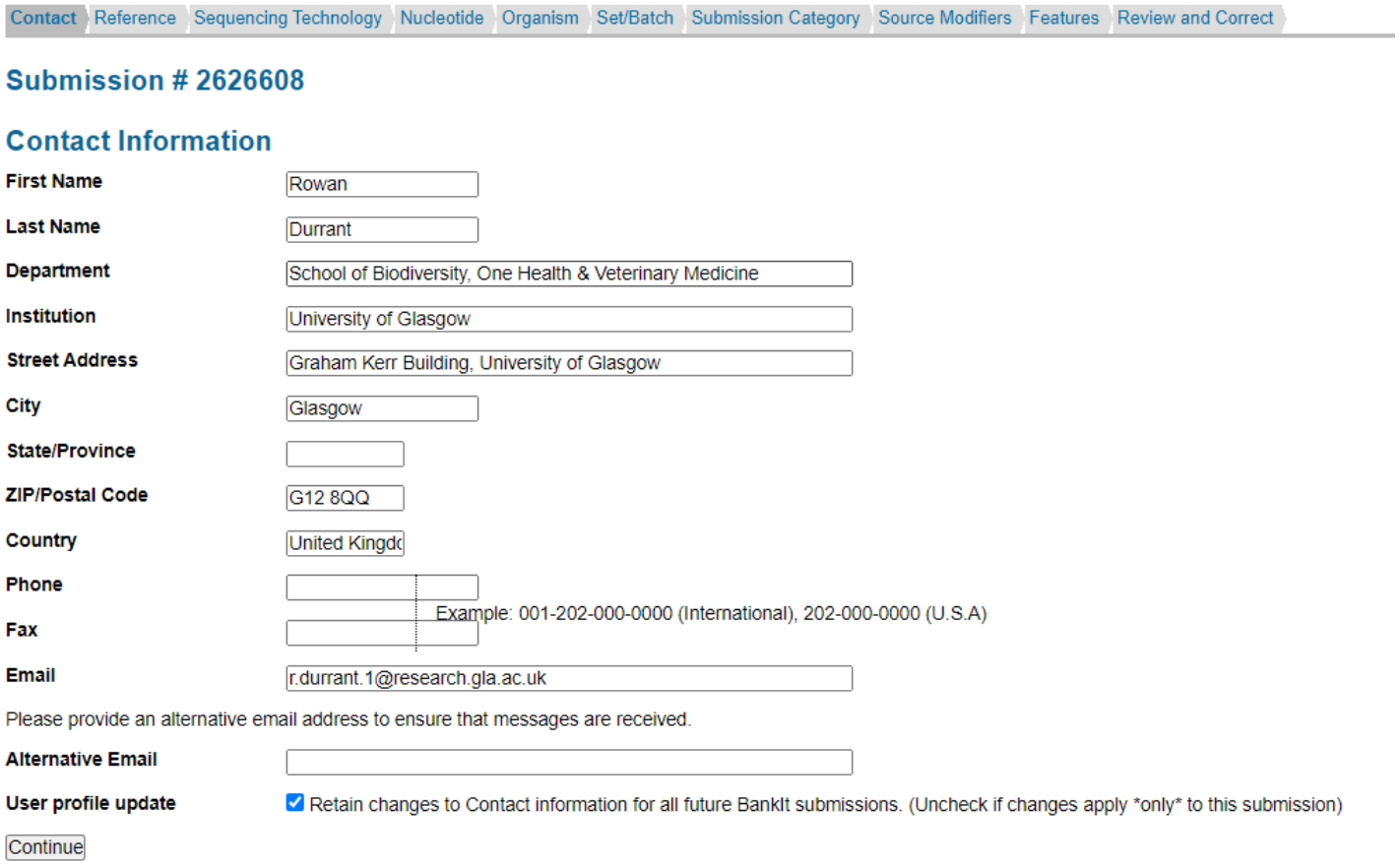
- ◉ Sequence data not listed above (through BankIt): genomic DNA, organelle, ncRNA, plasmids, other viruses, phages, other mRNA, synthetic constructs

Submission requirements:

- Sequence data in FASTA or alignment format
- Name(s) of the organism(s) where sequence data were isolated and any other descriptive data
- Sequence features (Examples: CDS, gene, tRNA, with nucleotide intervals and product names)

Start

# Step 2: Contact details



# Step 3: Authors & Reference

## Submission # 2626608

### Sequence Authors

| First Name | Middle Initial(s) | Last Name | Suffix | Remove |
|---|---|---|---|---|
| Gurdeep | | Jaswant | | X |
| Kirstyn | | Brunker | | X |
| Katie | | Hampson | | X |
| S | M | Thumbi | | X |
| Julius | O | Oyugi | | X |
| Charles | | Kayuki | | X |
| Davis | | Kuchaka | | X |
| Marco | | van Zwetselaar | | X |
| Kathryn | | Campbell | | X |
| Rowan | | Durrant | | X |
| Marieke | | Dekker | | X |

Add more sequence authors.

### Reference Information #1

Please provide the title and relevant publication details (volume, issue, etc.) of a paper that discusses this submission.

**PUBLICATION STATUS**

(•) Unpublished ( ) In-Press ( ) Published

**Reference Title** Molecular characterization of human rabies in East Africa - Human series report

**REFERENCE AUTHORS**

(•) Same As Sequence Authors

( ) Specify New Authors

# Step 4: Sequencing Technology

## Sequencing Technology

This information is required if you are submitting over 500 sequences or if your sequences were generated using next-generation sequencing technology.

What methods were used to obtain these sequences?

- [ ] **Sanger dideoxy sequencing**
- [ ] **454**
- [ ] **Helicos**
- [ ] **Illumina**
- [ ] **IonTorrent**
- [ ] **PacBio**
- [ ] **SOLiD**
- [x] **Other**

Oxford Nanopore MinION

Are these sequence(s):

- ( ) **unassembled sequence reads**
- (•) **assembled sequences (consisting of two or more sequence reads)**

What program(s) did you use to assemble these sequences?

| Assembly Program | Version or Date | Remove |
|---|---|---|
| Guppy | | X |
| Minimap2 | | X |

Add additional assembly programs.

**Assembly Name:**

# Step 5: Nucleotide

## Submission Release Date

**When may we release your sequence record?**

- (•) Immediately After Processing
- ( ) Release Date: [ ]

Date format is 'DD-Mon-YYYY' (example: 20-Feb-2004).
Release date must be after 6 months from today and before 10 years from today.

## Sequence(s) and Definition Line(s)

**Molecule Type:** [ ] Select the type of molecule that was isolated and sequenced.

**Topology:** [Linear]

**Are you submitting the complete sequence of an organelle genome, virus, viral segment, viroid, plasmid, or cloning vector?** (•) Yes ( ) No

### Nucleotide Sequence Format

Sequence data format:

- (•) **FASTA sequences (not an alignment, most common data format, FASTA help)**

  For example:

```
>Seq1 [organism=genus species] Definition Line for Seq1
aaccgatatagagagagga....
>Seq2 [organism=genus species] Definition Line for Seq2
atctgaatagagattattt....
```

  Definition Lines which are used to describe each sequence, should be included with your sequence data.

- ( ) **Alignment (FASTA+GAP, NEXUS, PHYLIP, CLUSTAL(W), alignment help)**

### Nucleotide Sequence(s)

**Paste sequence data**

[ ]

**(OR)**

**Upload file** [Choose file] No file chosen

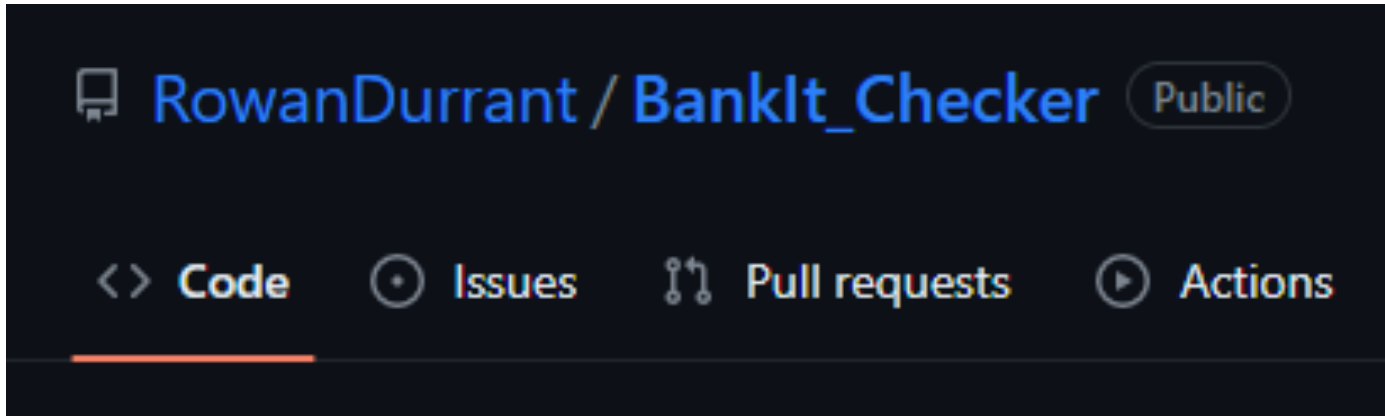
**Warning:** Terminal ends of the following sequence(s) are low quality (too many ambiguous bases) and have been trimmed. These are your options:

- Press Continue below to continue with the trimmed sequence(s)
- or, trim the sequence(s) yourself and upload the new corrected sequence(s)
- or, click here to undo **all** trimming, and then click Continue to submit the original untrimmed sequence(s)

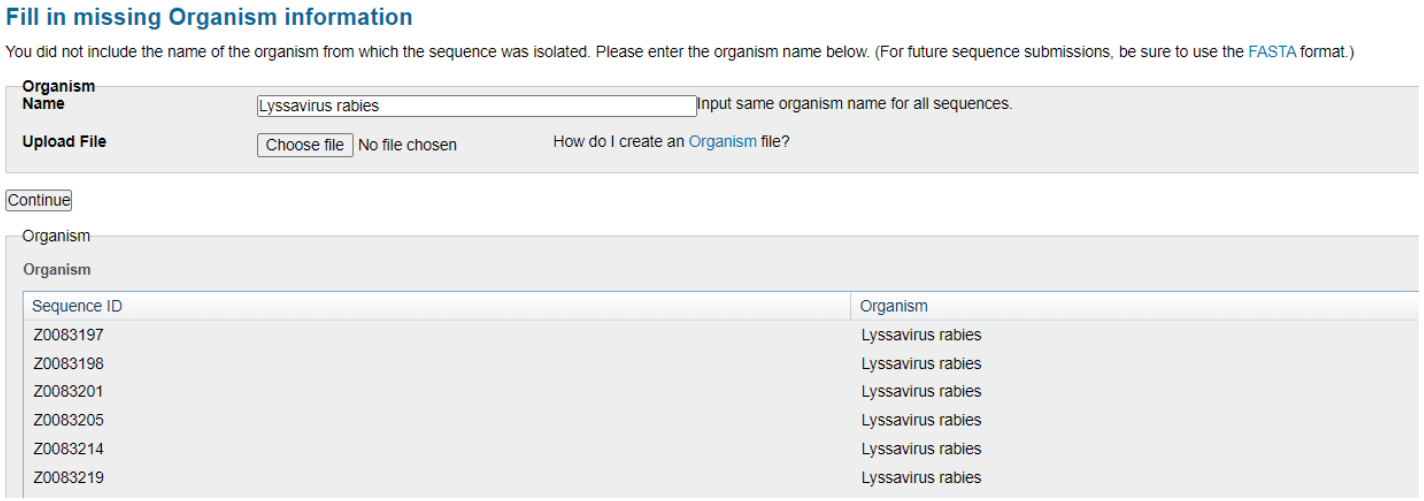
| Sequence ID | Original Length | New Trimmed Length |
|---|---|---|
| 2021153276 | 11926 | 11698 |
| HB002 | 11923 | 11695 |
| MR001 | 11923 | 11695 |
| MR005 | 11923 | 11695 |
| MR006 | 11926 | 11698 |
| MR008 | 11923 | 11695 |
| MR009 | 11923 | 11695 |
| RB001 | 11923 | 11695 |
| RB002 | 11923 | 11695 |
| RB003 | 11923 | 11695 |
| RB004 | 11923 | 11695 |
| SD703 | 11923 | 11620 |
| SD738 | 11923 | 11695 |
| SD741 | 11923 | 11695 |
| SD743 | 11923 | 11694 |

**Warning:** The sequences listed below contain more than 50% N's (unknown nucleotides) and are of low quality, which we cannot accept. Please trim or resequence these sequences to provide better quality data and upload the corrected data.

| Sequence ID |
|---|
| Z0826751 |



# Step 6: Organism



## Fill in missing Organism information

You did not include the name of the organism from which the sequence was isolated. Please enter the organism name below. (For future sequence submissions, be sure to use the FASTA format.)

**Organism Name** Lyssavirus rabies Input same organism name for all sequences.

**Upload File** Choose file No file chosen How do I create an Organism file?

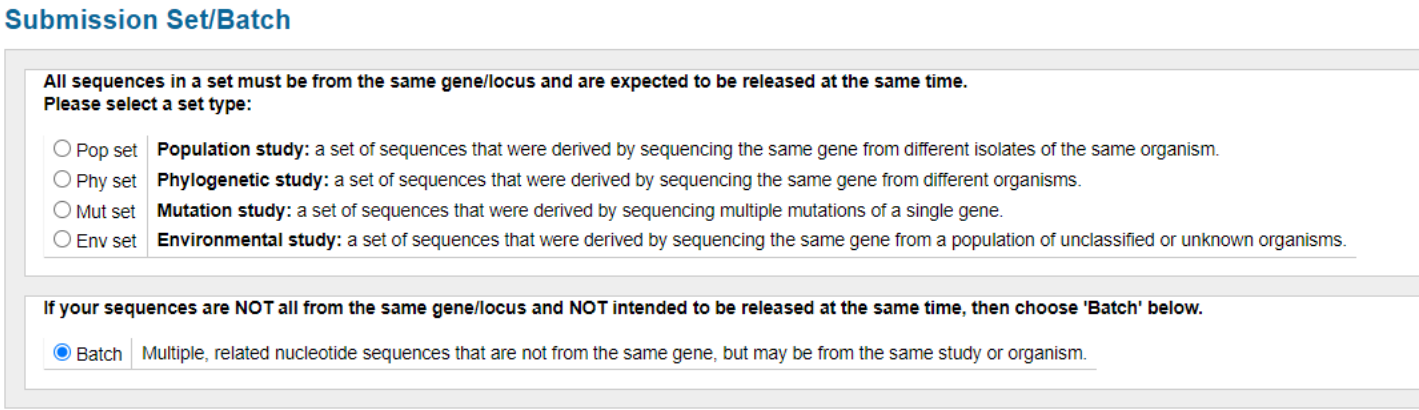
Continue

Organism

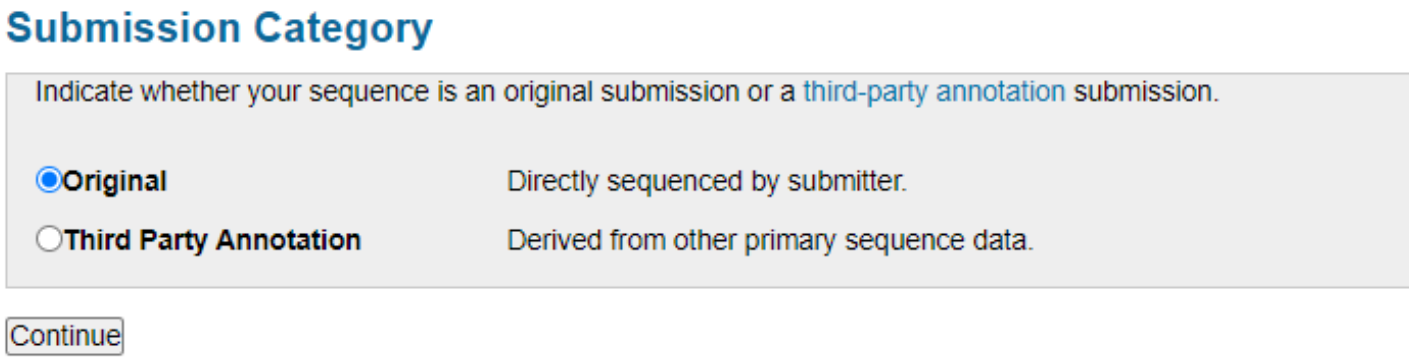
Organism

| Sequence ID | Organism |
|---|---|
| Z0083197 | Lyssavirus rabies |
| Z0083198 | Lyssavirus rabies |
| Z0083201 | Lyssavirus rabies |
| Z0083205 | Lyssavirus rabies |
| Z0083214 | Lyssavirus rabies |
| Z0083219 | Lyssavirus rabies |

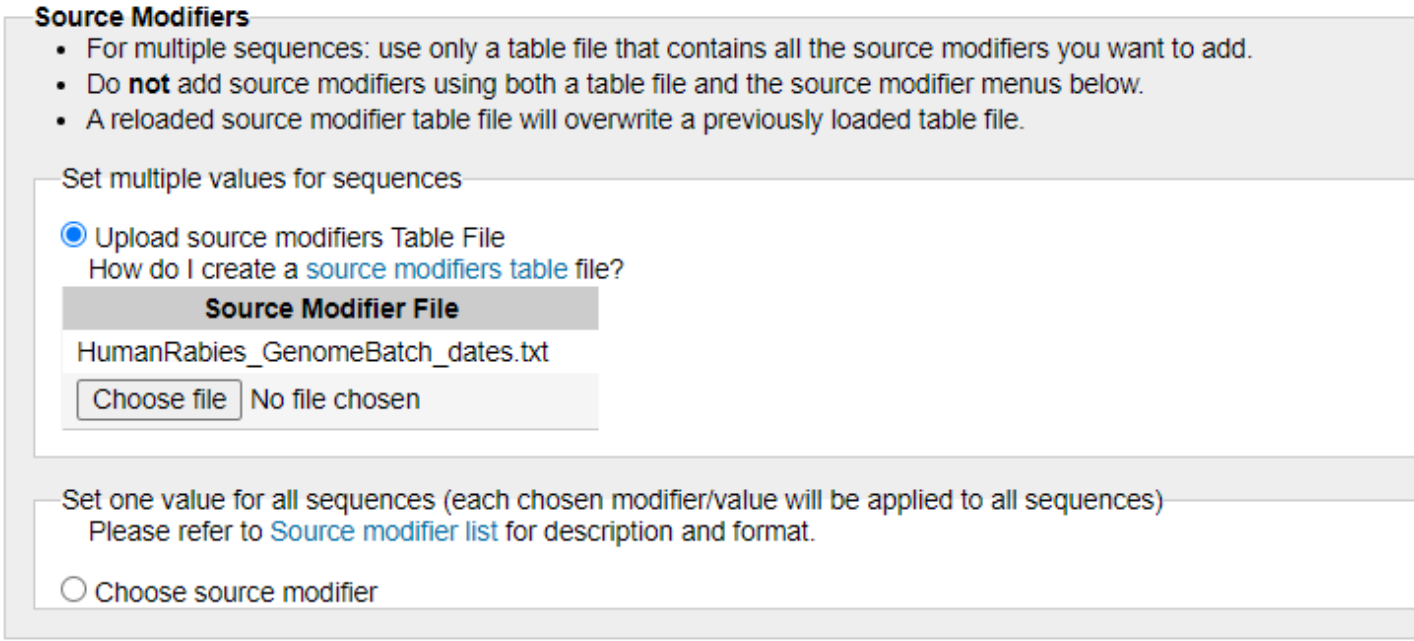
# Step 7: Set/ Batch



# Step 8: Submission Category



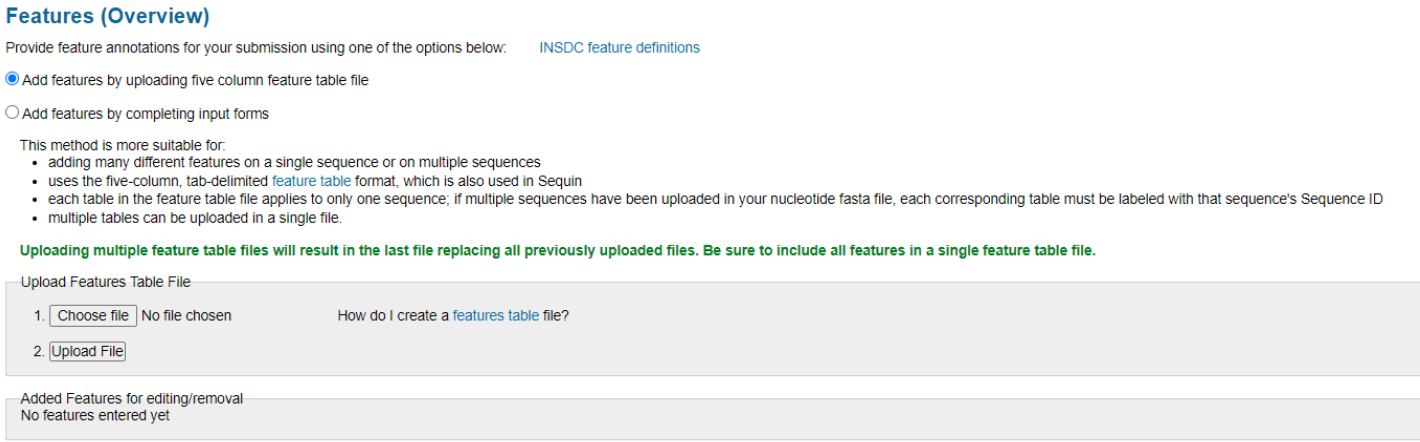
# Step 9: Source Modifiers



List of source modifiers and formats can be found here: https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Sequence_ID | Isolation_ | Country | Host | Collection_date |
| 2 | 02196160414A | | Kenya | | 2018 |
| 3 | 2010818843 | Brain | Kenya: Nairobi | Human | 12-Mar-13 |
| 4 | 2011100075 | Brain | Kenya: Nthangu | Bovine | 13-Jul-18 |
| 5 | 2011100176 | Brain | Kenya: Kwa DC | Bovine | 11-Jul-18 |
| 6 | 2011100186 | Brain | Kenya | Domestic dog | 12-Sep-17 |
| 7 | AR02 | | Tanzania: Arusha | Domestic dog | 01-Mar-22 |
| 8 | AR03 | | Tanzania: Arusha | Domestic dog | 01-Mar-22 |

# Step 10: Feature table



The word “Feature” + the name of the sequence as it appears in the FASTA file

```
>Feature Seq3
1080    1210    CDS
1275    1315
                        product         actin
                        note            alternatively spliced
1055    1210    mRNA
1275    1340
                        product         actin
1055    1340    gene
                        gene            ACT
1055    1079    5'UTR
1316    1340    3'UTR
```

The word “Feature” + the name of the sequence as it appears in the FASTA file

```
>Feature Seq3
1080    1210    CDS
1275    1315
                        product         actin
                        note            alternatively spliced
1055    1210    mRNA
1275    1340
                        product         actin
1055    1340    gene
                        gene            ACT
1055    1079    5'UTR
1316    1340    3'UTR
```

Columns 1&2: Start and stop locations of the feature

The word “Feature” + the name of the sequence as it appears in the FASTA file

```
>Feature Seq3
1080   1210   CDS
1275   1315
                              product         actin
                              note            alternatively spliced
1055   1210   mRNA
1275   1340
                              product         actin
1055   1340   gene
                              gene            ACT
1055   1079   5'UTR
1316   1340   3'UTR
```

Columns 1&2: Start and stop locations of the feature

Column 3: Feature type

The word “Feature” + the name of the sequence as it appears in the FASTA file

```
>Feature Seq3
1080    1210    CDS
1275    1315
                                product         actin
                                note            alternatively spliced
1055    1210    mRNA
1275    1340
                                product         actin
1055    1340    gene
                                gene            ACT
1055    1079    5'UTR
1316    1340    3'UTR
```

Columns 1&2: Start and stop locations of the feature

Column 3: Feature type

On a new line,
Column 4: Qualifier name

The word “Feature” + the name of the sequence as it appears in the FASTA file

```
>Feature Seq3
1080   1210   CDS
1275   1315
                                product         actin
                                note            alternatively spliced
1055   1210   mRNA
1275   1340
                                product         actin
1055   1340   gene
                                gene            ACT
1055   1079   5'UTR
1316   1340   3'UTR
```

Columns 1&2: Start and stop locations of the feature

Column 3: Feature type

On a new line,

Column 4: Qualifier name

Column 5: Qualifier value

# This is the hard bit, especially if you have lots of sequences!

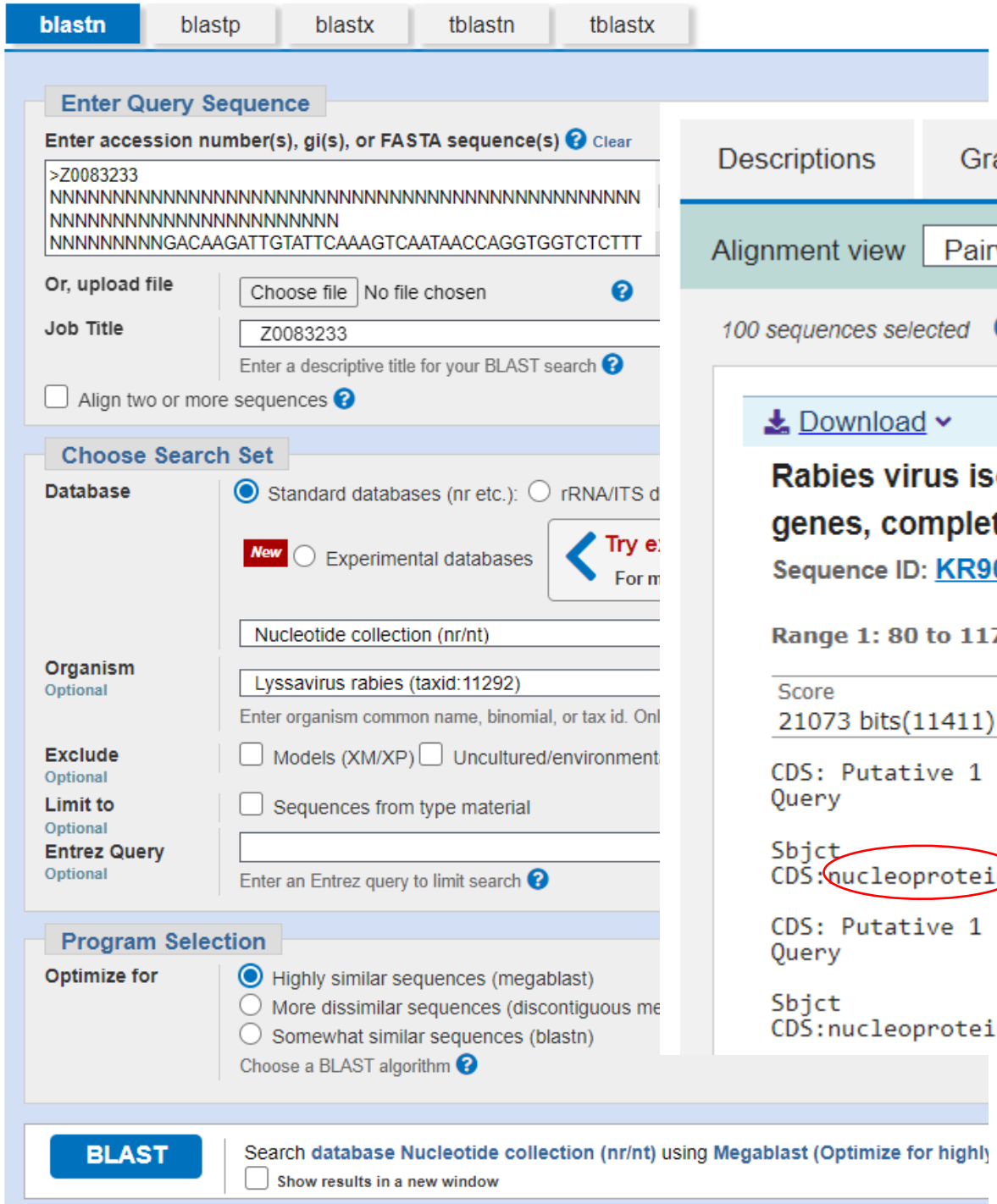
- GenBank seems to want you to do it by hand? Which takes ages
- More details on the ncbi site: https://www.ncbi.nlm.nih.gov/WebSub/html/help/feature-table.html
- BLAST might help
- I think there are some programs available to do this but I don’t know how reliable they are

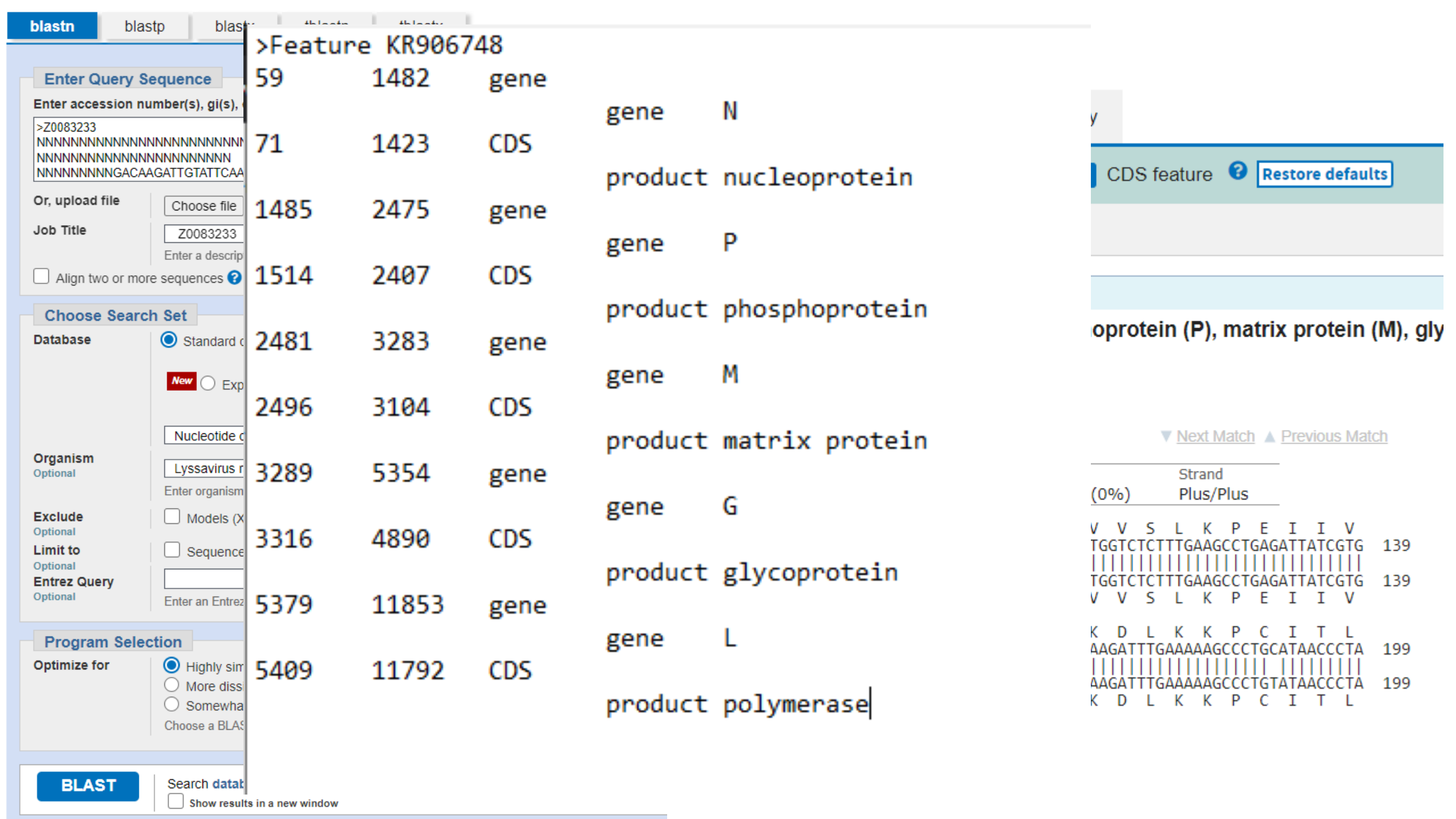
blastn | blastp | blastx | tblastn | tblastx

## Enter Query Sequence

**Enter accession number(s), gi(s), or FASTA sequence(s)** Clear

```
>Z0083233
NNNN
```


blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
>Z0083233
NNNN


blastn
blastp
Enter Query Sequence
Enter accession number(s), gi(s),
>Z0083233
NNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNGACAAGATTGTATTCAA
Or, upload file
Choose file
Job Title
Z0083233
Enter a descrip
Align two or more sequences
Choose Search Set
Database
Standard
New
Exp
Nucleotide
Organism
Optional
Lyssavirus r
Enter organism
Exclude
Optional
Models (X
Limit to
Optional
Sequence
Entrez Query
Optional
Enter an Entrez
Program Selection
Optimize for
Highly sim
More diss
Somewha
Choose a BLAS
BLAST
Search datab
Show results in a new window
>Feature KR906748
59 1482 gene
gene N
71 1423 CDS
product nucleoprotein
1485 2475 gene
gene P
1514 2407 CDS
product phosphoprotein
2481 3283 gene
gene M
2496 3104 CDS
product matrix protein
3289 5354 gene
gene G
3316 4890 CDS
product glycoprotein
5379 11853 gene
gene L
5409 11792 CDS
product polymerase
CDS feature
Restore defaults
oprotein (P), matrix protein (M), gly
Next Match
Previous Match
Strand
(0%)
Plus/Plus
V V S L K P E I I V
TGGTCTCTTTGAAGCCTGAGATTATCGTG 139
TGGTCTCTTTGAAGCCTGAGATTATCGTG 139
V V S L K P E I I V
K D L K K P C I T L
AAGATTTGAAAAAGCCCTGCATAACCCTA 199
AAGATTTGAAAAAGCCCTGTATAACCCTA 199
K D L K K P C I T L

# Step 11: Review & Submit!

## Review Submission

1. Additional Email Addresses?
   Correspondence regarding this submission will be sent to the following email address:
   r.durrant.1@research.gla.ac.uk
   Separate multiple email addresses with commas.

2. Resubmission?
   If you were asked by GenBank staff to resubmit your sequence data, check here: ☐

3. Submission Title (Optional)
   If you want to title your submission for your own record-keeping, check here: ☐

4. Additional Information

   If you have additional or corrected source modifier or feature files, or other plain text description for your sequence data submission, check here: ☐

5. Updates
   You may update or revise your submissions at any time by sending new or corrected information in an email to update@ncbi.nlm.nih.gov. You may also contact us at this address with any questions.

## Review Records of Your Set

Below please find your **first 5** genbank submission record(s) of your set for your review.

You can download the complete set as a compressed ZIP file.

Finish Submission

```
LOCUS       Z0083197              11695 bp    cRNA    linear   VRL 24-SEP-2022
DEFINITION  Lyssavirus rabies.
ACCESSION   Z0083197
VERSION
```

# Step 11: Review & Submit!

## Review Submission

1. Additional Email Addresses?
   Correspondence regarding this submission will be sent to the following email address:
   r.durrant.1@research.gla.ac.uk
   Separate multiple email addresses with commas.

2. Resubmission?
   If you were asked by GenBank staff to resubmit your sequence data, check here: ☐

3. Submission Title (Optional)
   If you want to title your submission for your own record-keeping, check here: ☐

4. Additional Information

   If you have additional or corrected source modifier or feature files, or other plain text description for your sequence data submission, check here: ☐

5. Updates
   You may update or revise your submissions at any time by sending new or corrected information in an email to update@ncbi.nlm.nih.gov. You may also contact us at this address with any questions.

## Review Records of Your Set

Below please find your **first 5** genbank submission record(s) of your set for your review.

You can download the complete set as a compressed ZIP file.

Finish Submission

```
LOCUS       Z0083197              11695 bp    cRNA    linear   VRL 24-SEP-2022
DEFINITION  Lyssavirus rabies.
ACCESSION   Z0083197
VERSION
```