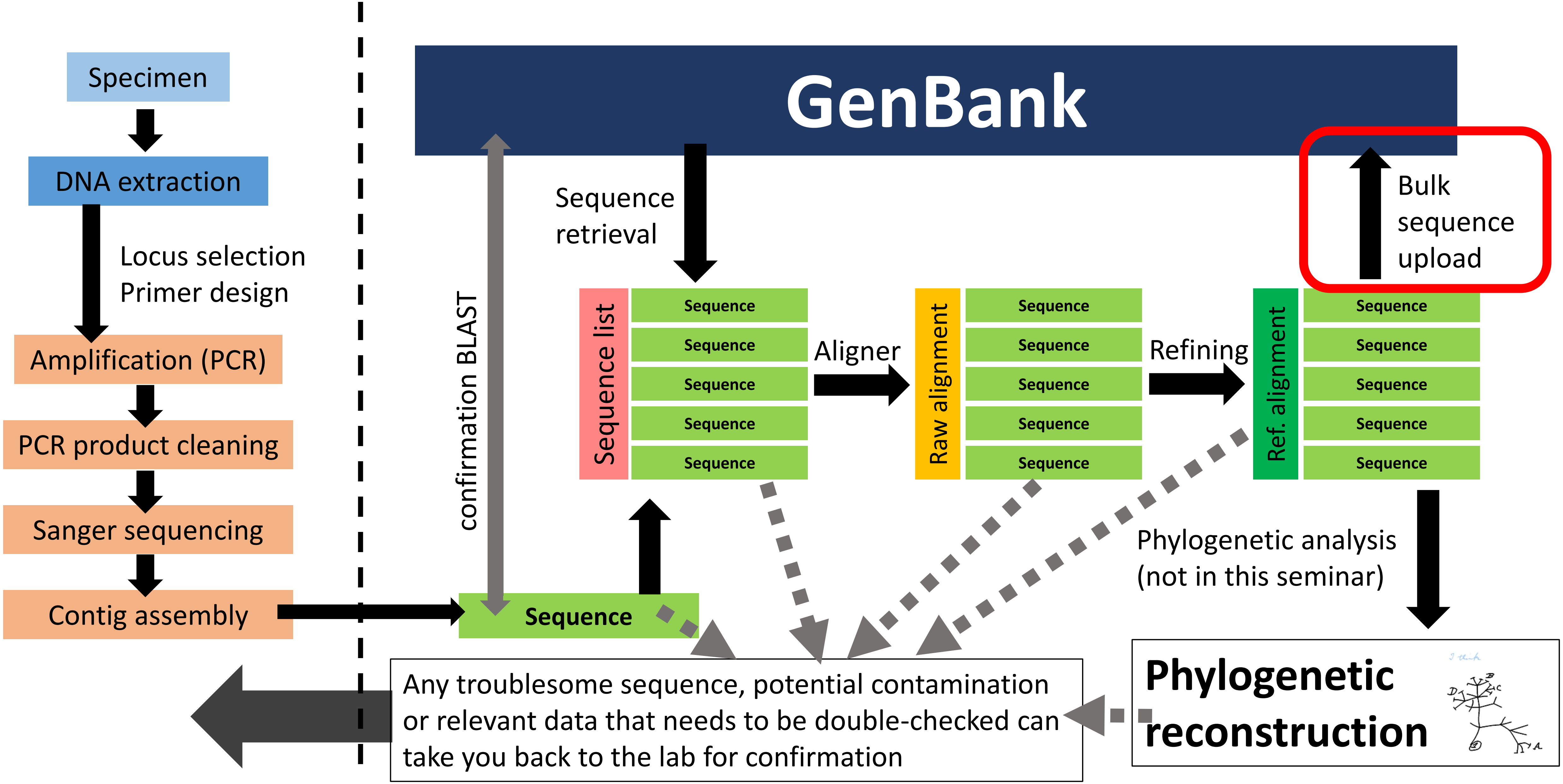
**A primer for practical phylogenetic data gathering. Uconn EEB3899-007. Spring 2015**

Session 5

# Uploading sequences to GenBank

**Rafael Medina (rafael.medina.bry@gmail.com)**
**Yang Liu (yang.liu@uconn.edu)**

# A road map to phylogenetic data gathering



Specimen

DNA extraction

Locus selection
Primer design

Amplification (PCR)

PCR product cleaning

Sanger sequencing

Contig assembly

Sequence

confirmation BLAST

GenBank

Sequence retrieval

Sequence list

Sequence
Sequence
Sequence
Sequence
Sequence

Aligner

Raw alignment

Sequence
Sequence
Sequence
Sequence
Sequence

Refining

Ref. alignment

Sequence
Sequence
Sequence
Sequence
Sequence

Bulk sequence upload

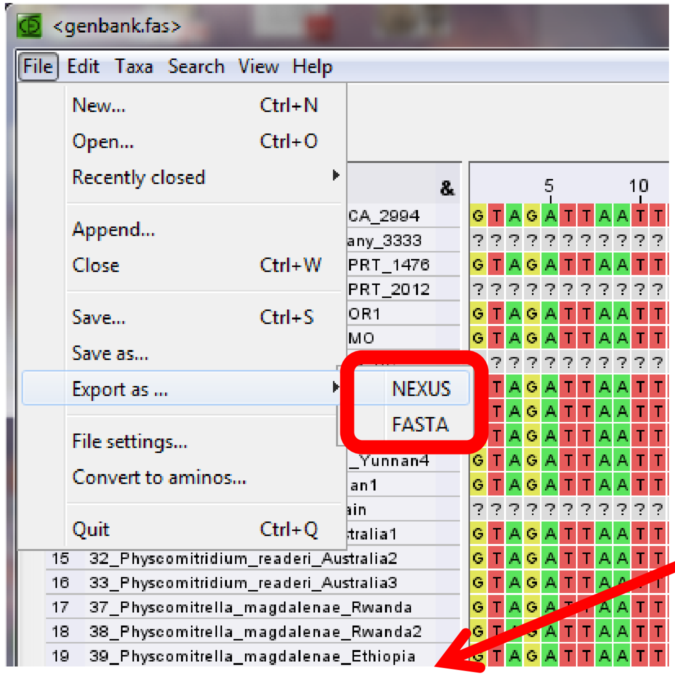
Phylogenetic analysis (not in this seminar)

**Phylogenetic reconstruction**

Any troublesome sequence, potential contamination or relevant data that needs to be double-checked can take you back to the lab for confirmation

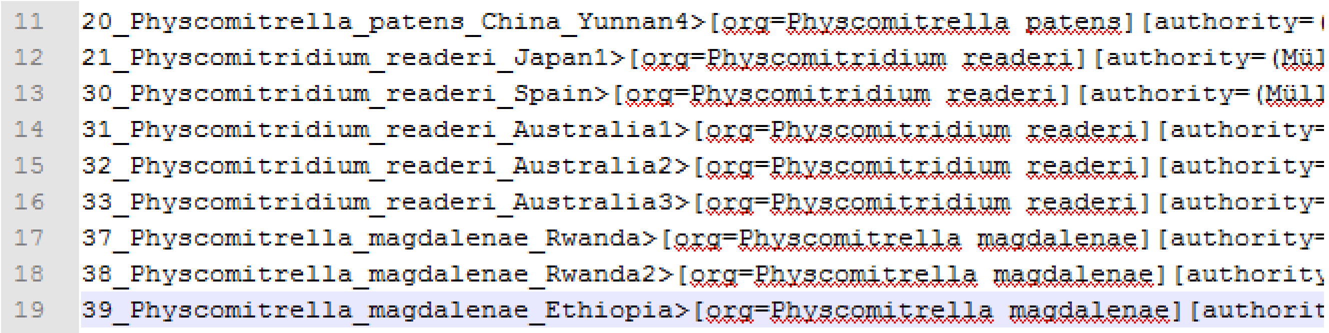
# Preparation of the NEXUS file

- Use the full alignment (without exclusions) for Genbank submission
- Export as a NEXUS file



```
CA_2994                                    GTAGATTAATT
any_3333                                   ???????????
PRT_1476                                   GTAGATTAATT
PRT_2012                                   ???????????
OR1                                        GTAGATTAATT
MO                                         GTAGATTAATT
                                           ???????????
                                           TAGATTAATT
                                           TAGATTAATT
                                           TAGATTAATT
_Yunnan4                                   GTAGATTAATT
an1                                        GTAGATTAATT
ain                                        ???????????
tralia1                                    GTAGATTAATT
15  32_Physcomitridium_readeri_Australia2  GTAGATTAATT
16  33_Physcomitridium_readeri_Australia3  GTAGATTAATT
17  37_Physcomitrella_magdalenae_Rwanda    GTAGATTAATT
18  38_Physcomitrella_magdalenae_Rwanda2   GTAGATTAATT
19  39_Physcomitrella_magdalenae_Ethiopia  GTAGATTAATT
```

Create a text file with voucher specifications (this will save you time later)



```
20_Physcomitrella_patens_China_Yunnan4>[org=Physcomitrella patens][authority=(
21_Physcomitridium_readeri_Japan1>[org=Physcomitridium readeri][authority=(Mül
30_Physcomitridium_readeri_Spain>[org=Physcomitridium readeri][authority=(Mül
31_Physcomitridium_readeri_Australia1>[org=Physcomitridium readeri][authority=
32_Physcomitridium_readeri_Australia2>[org=Physcomitridium readeri][authority=
33_Physcomitridium_readeri_Australia3>[org=Physcomitridium readeri][authority=
37_Physcomitrella_magdalenae_Rwanda>[org=Physcomitrella magdalenae][authority=
38_Physcomitrella_magdalenae_Rwanda2>[org=Physcomitrella magdalenae][authority
39_Physcomitrella_magdalenae_Ethiopia>[org=Physcomitrella magdalenae][authorit
```

Use the same sequence name and order as in the alignment

Use “>” as separator

List of attributes:
[org=] (organism)
[authority=]
[molecule=]
[location=]
[specimen-voucher=]
[note=]

Once it is ready, erase everything before the “>”. Do not change the order afterwards

# Preparation of the NEXUS file

Add a NCBI block in the nexus file

It should start with “BEGIN NCBI” and should be closed with “END;”

The voucher specifications are added as a matrix with the label “SEQUIN”

End the matrix with “;”

“DATA” block, generated by the editor

NCBI block

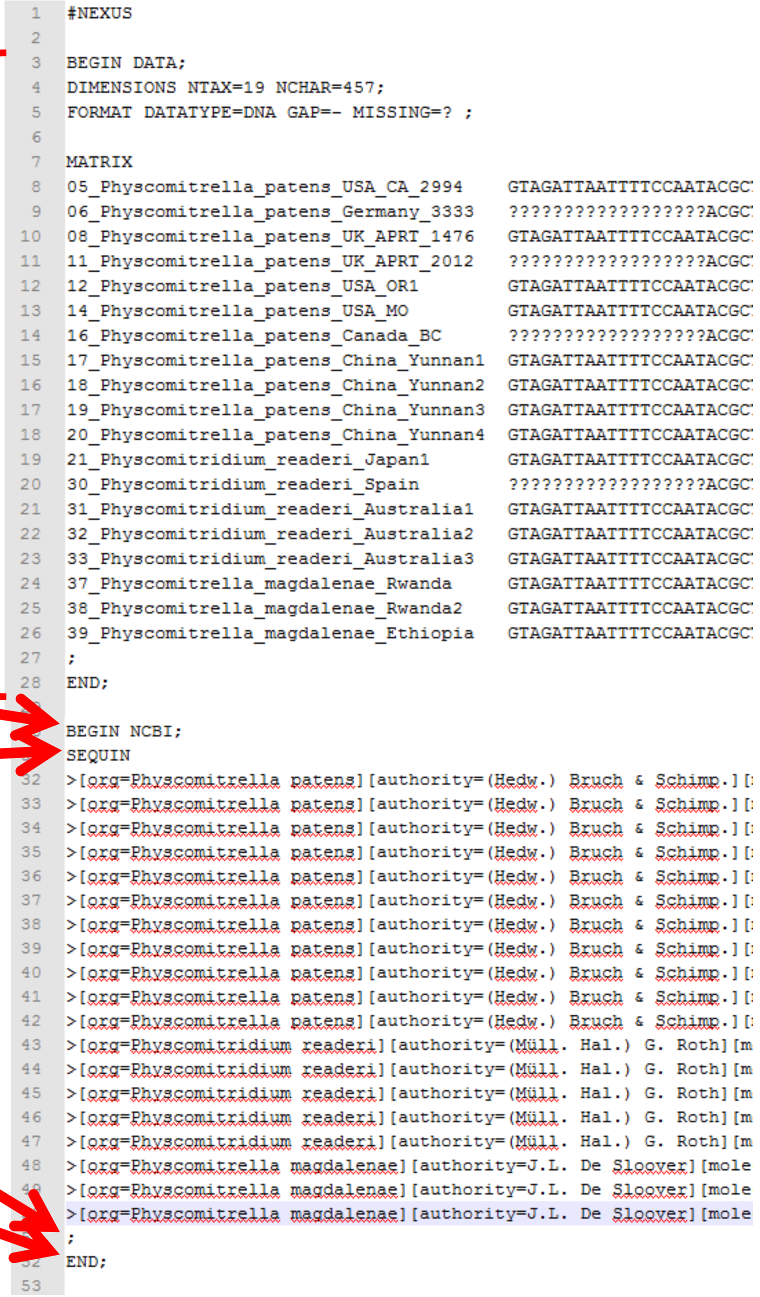
Beginning of the block

“Sequin” label

Specifications, notes etc (pasted as in the previous slide)

This ends the sequin matrix

End of the NCBI block

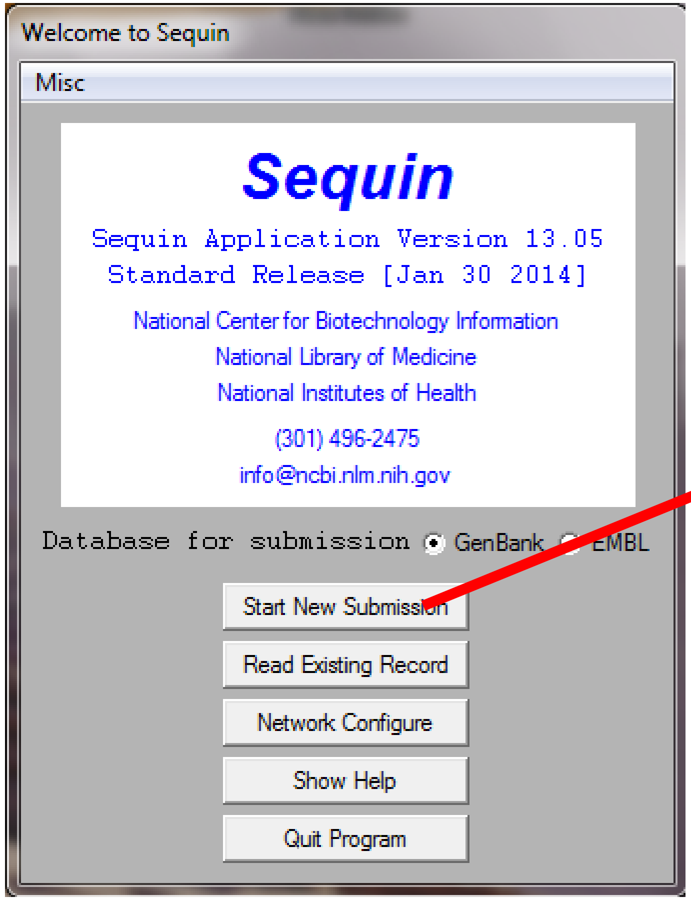


```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=19 NCHAR=457;
FORMAT DATATYPE=DNA GAP=- MISSING=? ;

MATRIX
05_Physcomitrella_patens_USA_CA_2994    GTAGATTAATTTTCCAATACGC
06_Physcomitrella_patens_Germany_3333   ??????????????????ACGC
08_Physcomitrella_patens_UK_APRT_1476   GTAGATTAATTTTCCAATACGC
11_Physcomitrella_patens_UK_APRT_2012   ??????????????????ACGC
12_Physcomitrella_patens_USA_OR1        GTAGATTAATTTTCCAATACGC
14_Physcomitrella_patens_USA_MO         GTAGATTAATTTTCCAATACGC
16_Physcomitrella_patens_Canada_BC      ??????????????????ACGC
17_Physcomitrella_patens_China_Yunnan1  GTAGATTAATTTTCCAATACGC
18_Physcomitrella_patens_China_Yunnan2  GTAGATTAATTTTCCAATACGC
19_Physcomitrella_patens_China_Yunnan3  GTAGATTAATTTTCCAATACGC
20_Physcomitrella_patens_China_Yunnan4  GTAGATTAATTTTCCAATACGC
21_Physcomitridium_readeri_Japan1       GTAGATTAATTTTCCAATACGC
30_Physcomitridium_readeri_Spain        ??????????????????ACGC
31_Physcomitridium_readeri_Australia1   GTAGATTAATTTTCCAATACGC
32_Physcomitridium_readeri_Australia2   GTAGATTAATTTTCCAATACGC
33_Physcomitridium_readeri_Australia3   GTAGATTAATTTTCCAATACGC
37_Physcomitrella_magdalenae_Rwanda     GTAGATTAATTTTCCAATACGC
38_Physcomitrella_magdalenae_Rwanda2    GTAGATTAATTTTCCAATACGC
39_Physcomitrella_magdalenae_Ethiopia   GTAGATTAATTTTCCAATACGC
;
END;

BEGIN NCBI;
SEQUIN
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitrella patens][authority=(Hedw.) Bruch & Schimp.][
>[org=Physcomitridium readeri][authority=(Müll. Hal.) G. Roth][m
>[org=Physcomitridium readeri][authority=(Müll. Hal.) G. Roth][m
>[org=Physcomitridium readeri][authority=(Müll. Hal.) G. Roth][m
>[org=Physcomitridium readeri][authority=(Müll. Hal.) G. Roth][m
>[org=Physcomitridium readeri][authority=(Müll. Hal.) G. Roth][m
>[org=Physcomitrella magdalenae][authority=J.L. De Sloover][mole
>[org=Physcomitrella magdalenae][authority=J.L. De Sloover][mole
>[org=Physcomitrella magdalenae][authority=J.L. De Sloover][mole
;
END;

```

# SEQUIN

Sequin is the NCBI software developed to upload sequences to GenBank

Download it from here


Welcome to Sequin

**Sequin**

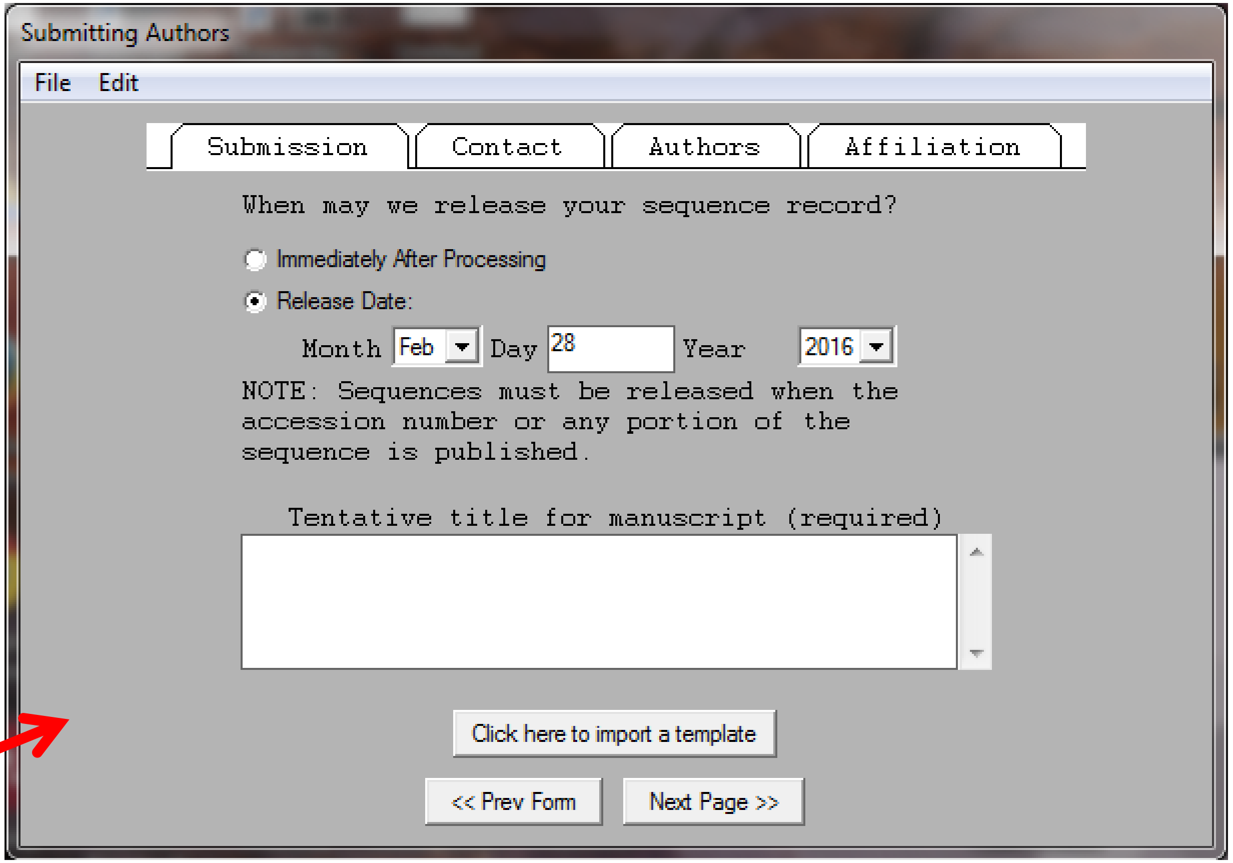
Sequin Application Version 13.05
Standard Release [Jan 30 2014]

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

(301) 496-2475
info@ncbi.nlm.nih.gov

Database for submission ● GenBank ○ EMBL

New submission


Submitting Authors

When may we release your sequence record?

○ Immediately After Processing

● Release Date:

Month Feb Day 28 Year 2016

NOTE: Sequences must be released when the accession number or any portion of the sequence is published.

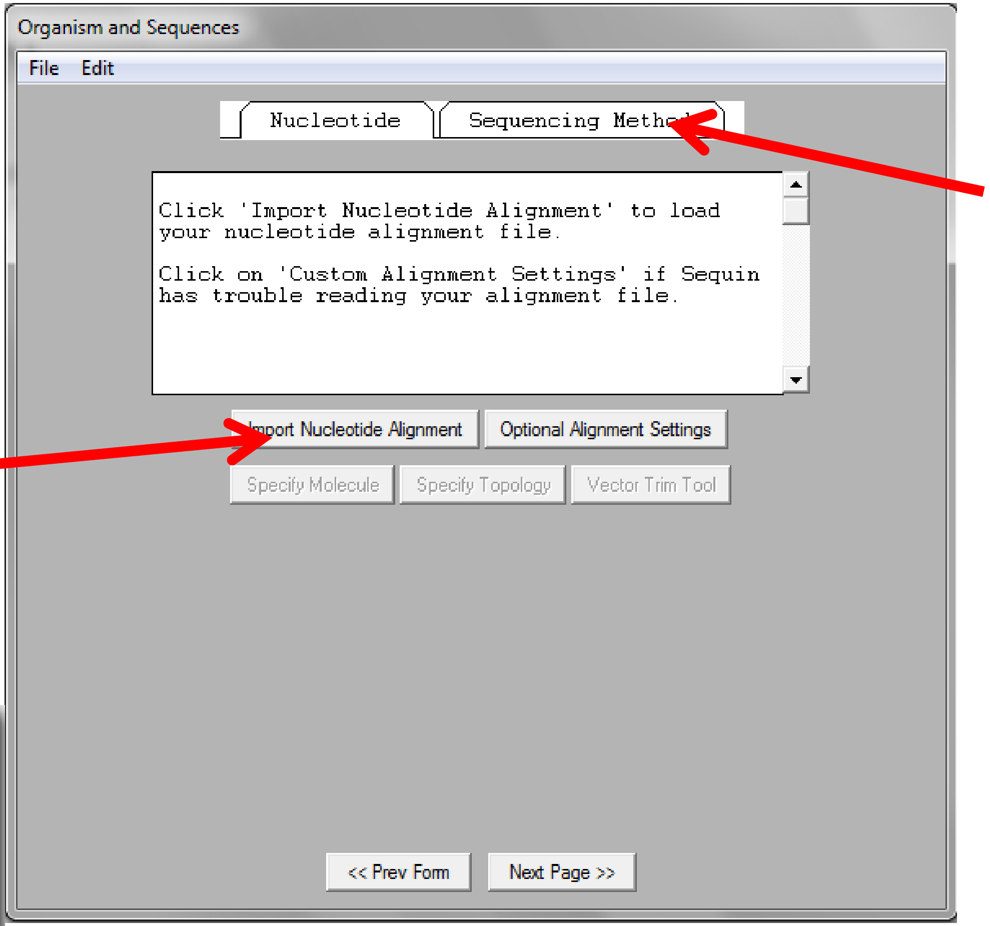
Tentative title for manuscript (required)

Form 1: authors
Fill the details of the publication, desired release date, authors, affiliation, etc.

Save this information in a template before proceeding to the next form, it will save you time if the program crashes (surprisingly likely scenario)

# SEQUIN

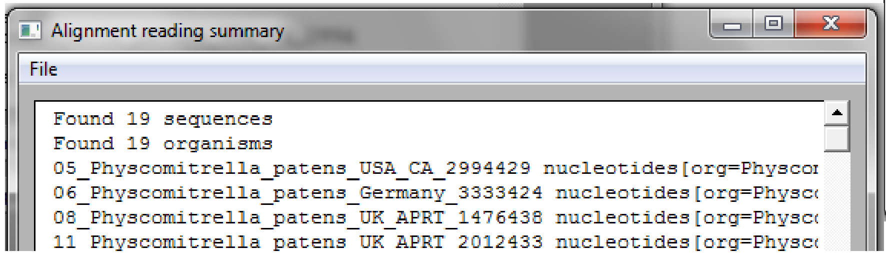
Form 2: alignment
In this example we are submitting an alignment through the “normal submission dialog”, it is a phylogenetic study and we will upload a nexus alignment



```
Click 'Import Nucleotide Alignment' to load
your nucleotide alignment file.

Click on 'Custom Alignment Settings' if Sequin
has trouble reading your alignment file.
```

You will be then required to input how the sequences were obtained

Import the alignment. If you have done everything correctly, the sequences AND the attributes of the nexus file will be incorporated

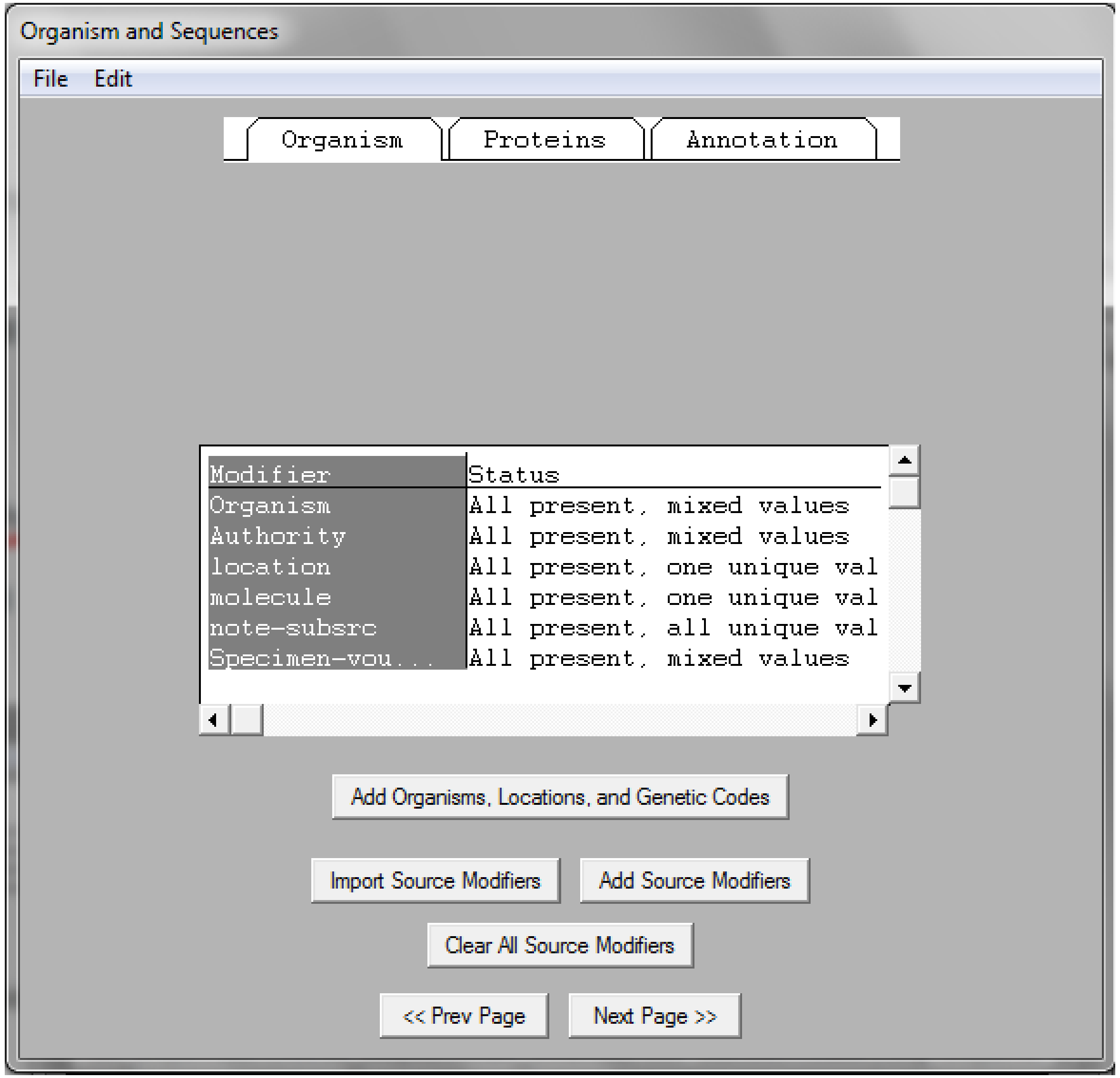


```
Found 19 sequences
Found 19 organisms
05_Physcomitrella_patens_USA_CA_2994429 nucleotides[org=Physcor
06_Physcomitrella_patens_Germany_3333424 nucleotides[org=Physc
08_Physcomitrella_patens_UK_APRT_1476438 nucleotides[org=Physc
11 Physcomitrella patens UK APRT 2012433 nucleotides[org=Physc
```

Specify the type of molecule and topology (lineal DNA in this case)

# SEQUIN

Form 3: general characteristics



| Modifier | Status |
|---|---|
| Organism | All present, mixed values |
| Authority | All present, mixed values |
| location | All present, one unique val |
| molecule | All present, one unique val |
| note-subsrc | All present, all unique val |
| Specimen-vou... | All present, mixed values |

**Organism**

The organism names, details, genetic code, etc, should have been incorporated automatically by now. Check it before proceeding

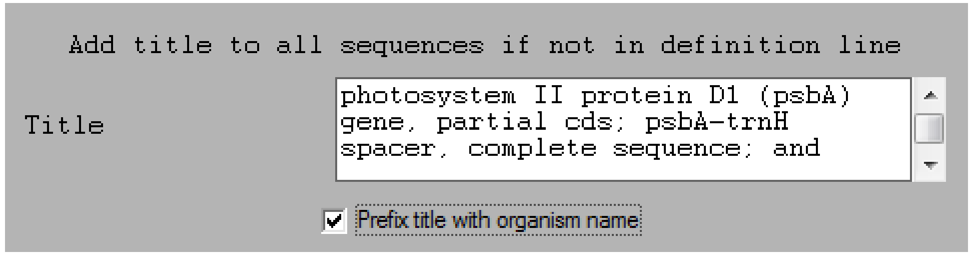
**Proteins**

Unless you are submitting a single protein gene, skip this tab

**Annotation**

Unless your sequence includes only one feature (region, gene), just tick “none”

Then…



Add title to all sequences if not in definition line

Title: photosystem II protein D1 (psbA) gene, partial cds; psbA-trnH spacer, complete sequence; and

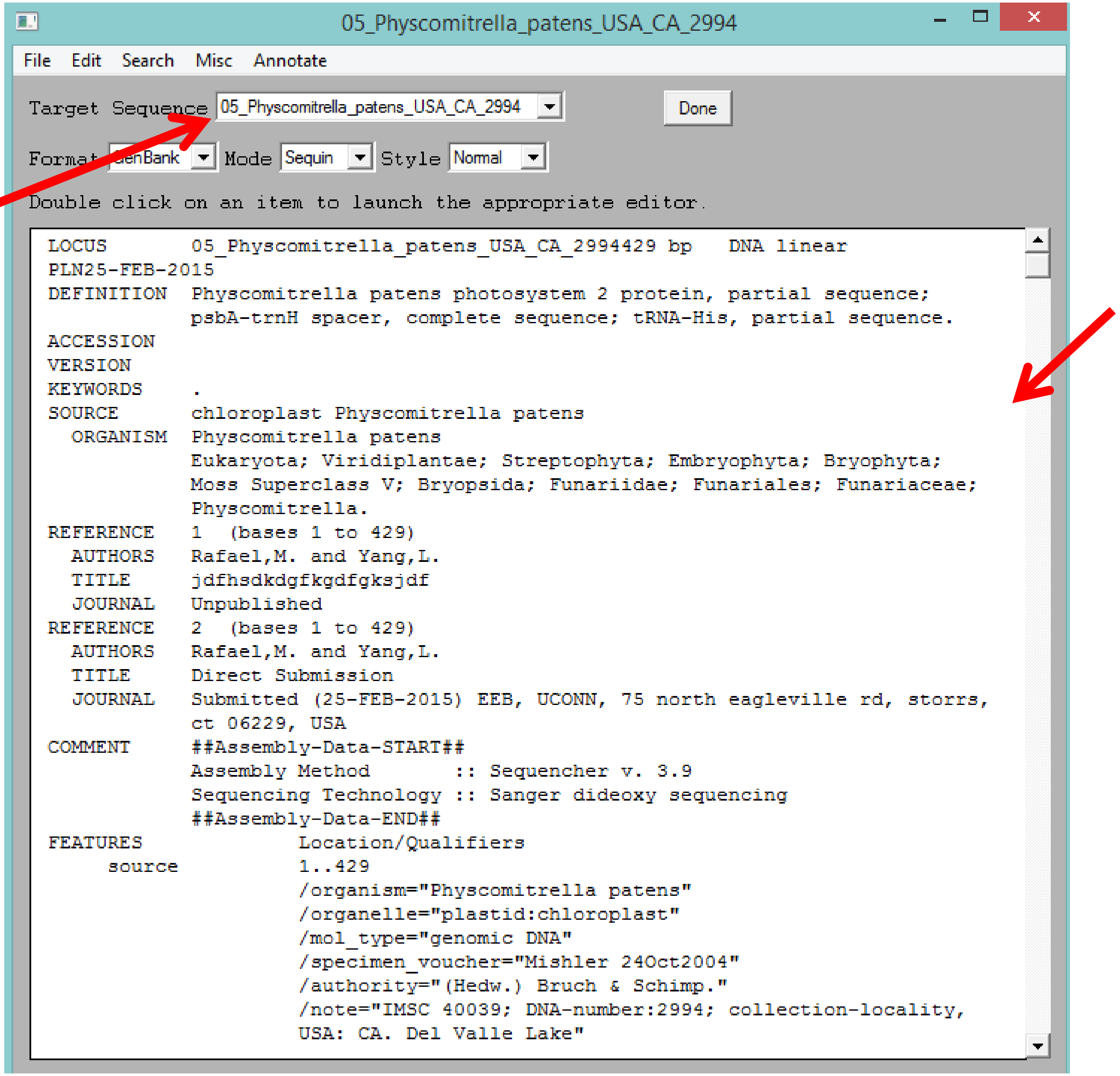
☑ Prefix title with organism name

Define the title of your sequences. Include a detailed description of the region and choose to prefix the title with the organism name

# SEQUIN

Form 4: annotation

Pick a sample with a good, long sequence to use as a template for annotations



Target Sequence 05_Physcomitrella_patens_USA_CA_2994

Format GenBank Mode Sequin Style Normal

Double click on an item to launch the appropriate editor.

```
LOCUS       05_Physcomitrella_patens_USA_CA_2994429 bp    DNA linear
PLN25-FEB-2015
DEFINITION  Physcomitrella patens photosystem 2 protein, partial sequence;
            psbA-trnH spacer, complete sequence; tRNA-His, partial sequence.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      chloroplast Physcomitrella patens
  ORGANISM  Physcomitrella patens
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Bryophyta;
            Moss Superclass V; Bryopsida; Funariidae; Funariales; Funariaceae;
            Physcomitrella.
REFERENCE   1  (bases 1 to 429)
  AUTHORS   Rafael,M. and Yang,L.
  TITLE     jdfhsdkdgfkgdfgksjdf
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 429)
  AUTHORS   Rafael,M. and Yang,L.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-FEB-2015) EEB, UCONN, 75 north eagleville rd, storrs,
            ct 06229, USA
COMMENT     ##Assembly-Data-START##
            Assembly Method       :: Sequencher v. 3.9
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..429
                     /organism="Physcomitrella patens"
                     /organelle="plastid:chloroplast"
                     /mol_type="genomic DNA"
                     /specimen_voucher="Mishler 24Oct2004"
                     /authority="(Hedw.) Bruch & Schimp."
                     /note="IMSC 40039; DNA-number:2994; collection-locality,
                     USA: CA. Del Valle Lake"
```

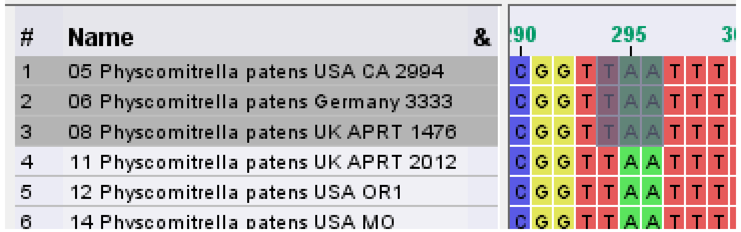
In this area you can see how the sequence will look in Genbank. So far all the relevant information is included except the annotations, the “indices” that will tell the reader where each region starts and ends

At this stage, go back to your lab book and summarize the limits of each feature of the sequence that you want to annotate

# SEQUIN

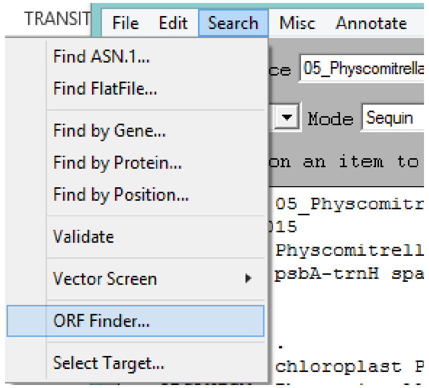
Example 1: annotation of a CDS (Coding DNA Sequence)

The first part of this sequence is the partial psbA protein. The stop codon (TAA) ends in position 296

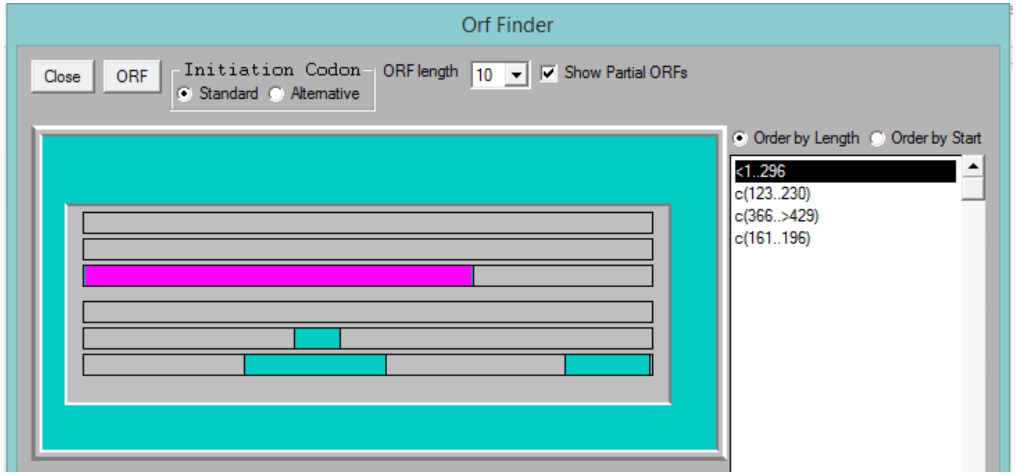


| # | Name | & | 290 |  |  |  |  | 295 |  |  |  | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 05 Physcomitrella patens USA CA 2994 |  | C | G | G | T | T | A | A | T | T | T |
| 2 | 06 Physcomitrella patens Germany 3333 |  | C | G | G | T | T | A | A | T | T | T |
| 3 | 08 Physcomitrella patens UK APRT 1476 |  | C | G | G | T | T | A | A | T | T | T |
| 4 | 11 Physcomitrella patens UK APRT 2012 |  | C | G | G | T | T | A | A | T | T | T |
| 5 | 12 Physcomitrella patens USA OR1 |  | C | G | G | T | T | A | A | T | T | T |
| 6 | 14 Physcomitrella patens USA MO |  | C | G | G | T | T | A | A | T | T | T |

Use the ORF (Open Reading Frame) finder to pick this CDS



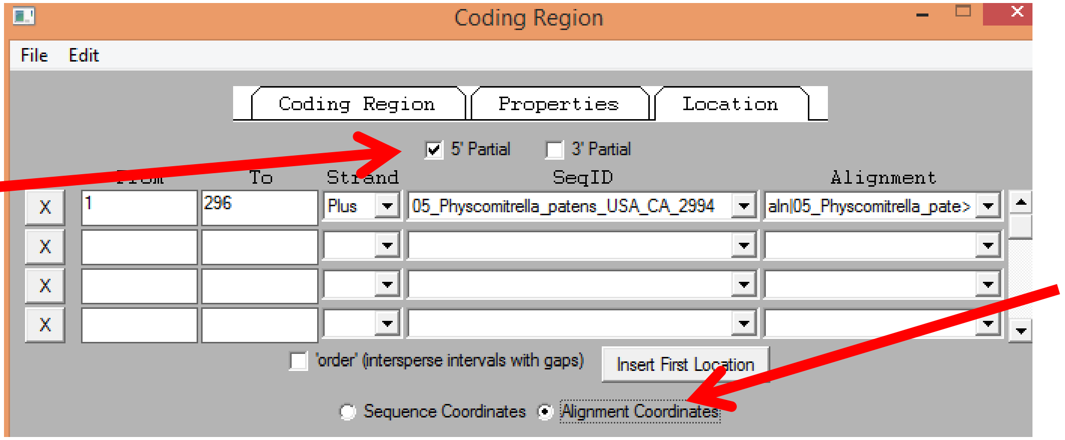
You can easily explore the six possible frames and select the one that corresponds to the actual protein



<1..296
c(123..230)
c(366..>429)
c(161..196)

Double click the CDS and complete the relevant information, at least the name and abbreviation of the protein, and a comment clarifying that it is partial

5' partial



☑ 5' Partial ☐ 3' Partial

| From | To | Strand | SeqID | Alignment |
|---|---|---|---|---|
| 1 | 296 | Plus | 05_Physcomitrella_patens_USA_CA_2994 | aln\|05_Physcomitrella_pate> |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

☐ 'order' (intersperse intervals with gaps)

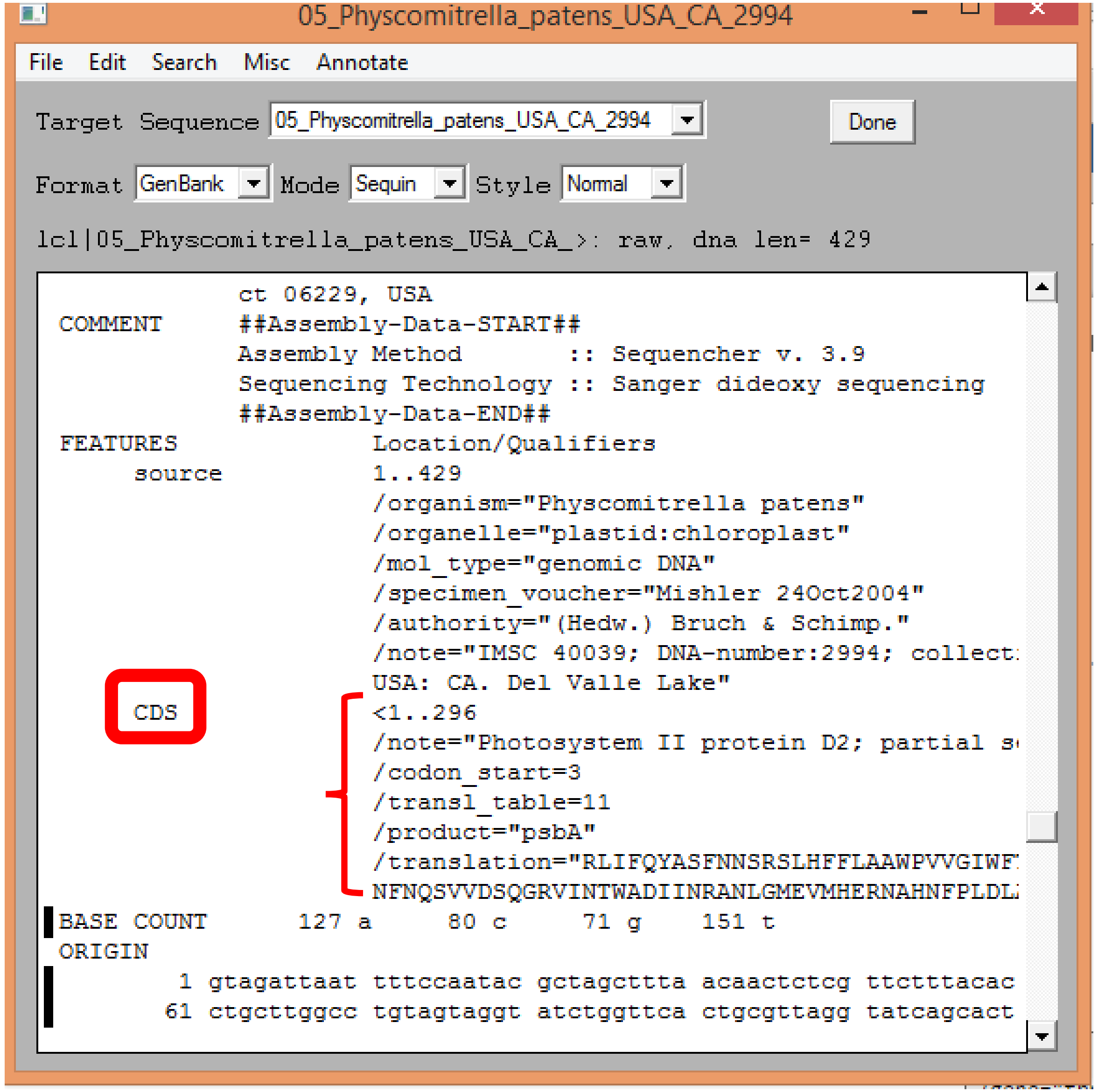
○ Sequence Coordinates ● Alignment Coordinates

Use better alignment coordinates if you have indels

Finally, check the location details and accept

# SEQUIN

After the annotation is done, it should appear correctly in the visualization window



Target Sequence 05_Physcomitrella_patens_USA_CA_2994

Format GenBank Mode Sequin Style Normal

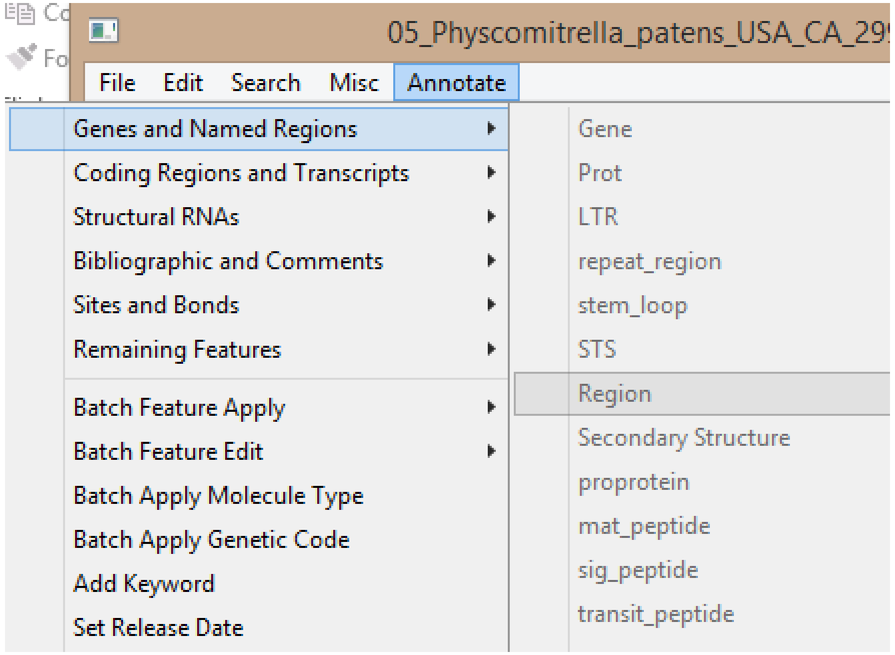
lcl|05_Physcomitrella_patens_USA_CA_>: raw, dna len= 429

```
            ct 06229, USA
COMMENT     ##Assembly-Data-START##
            Assembly Method       :: Sequencher v. 3.9
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..429
                     /organism="Physcomitrella patens"
                     /organelle="plastid:chloroplast"
                     /mol_type="genomic DNA"
                     /specimen_voucher="Mishler 24Oct2004"
                     /authority="(Hedw.) Bruch & Schimp."
                     /note="IMSC 40039; DNA-number:2994; collect
                     USA: CA. Del Valle Lake"
     CDS             <1..296
                     /note="Photosystem II protein D2; partial s
                     /codon_start=3
                     /transl_table=11
                     /product="psbA"
                     /translation="RLIFQYASFNNSRSLHFFLAAWPVVGIWF
                     NFNQSVVDSQGRVINTWADIINRANLGMEVMHERNAHNFPLDL
BASE COUNT      127 a     80 c     71 g    151 t
ORIGIN
        1 gtagattaat tttccaatac gctagcttta acaactctcg ttctttacac
       61 ctgcttggcc tgtagtaggt atctggttca ctgcgttagg tatcagcact
```

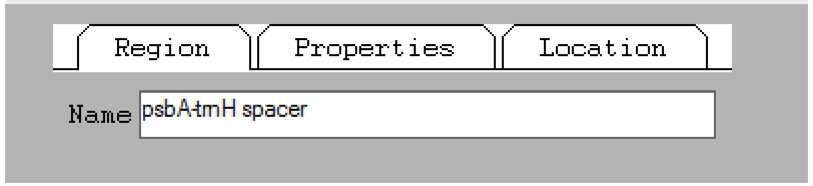
# SEQUIN

## Example 2: annotation of a spacer

Assuming you want to annotate a well-known, non-coding spacer (as psbA/trnH), you can consider it a “named region”

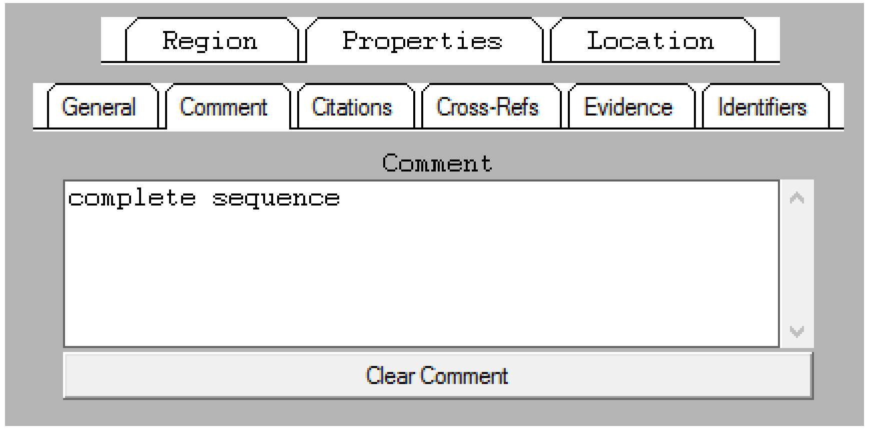


Name the region



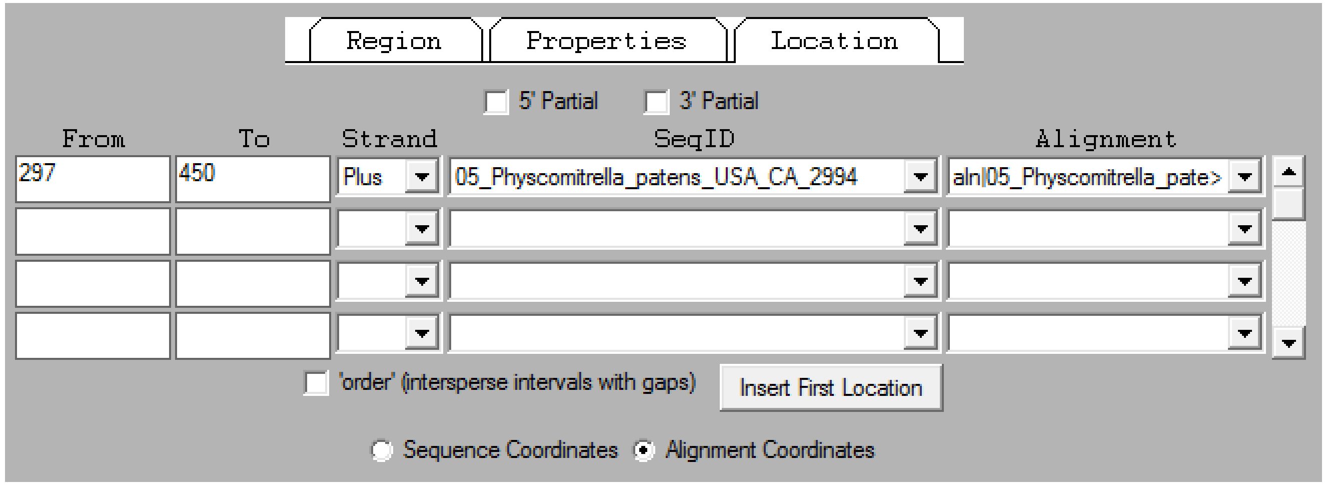
Name: psbA-trnH spacer

Comment if it is complete or partial



Comment: complete sequence

Specify the interval in the alignment



☐ 5' Partial ☐ 3' Partial

| From | To | Strand | SeqID | Alignment |
|---|---|---|---|---|
| 297 | 450 | Plus | 05_Physcomitrella_patens_USA_CA_2994 | aln\|05_Physcomitrella_pate> |
| | | | | |
| | | | | |
| | | | | |

☐ 'order' (intersperse intervals with gaps)

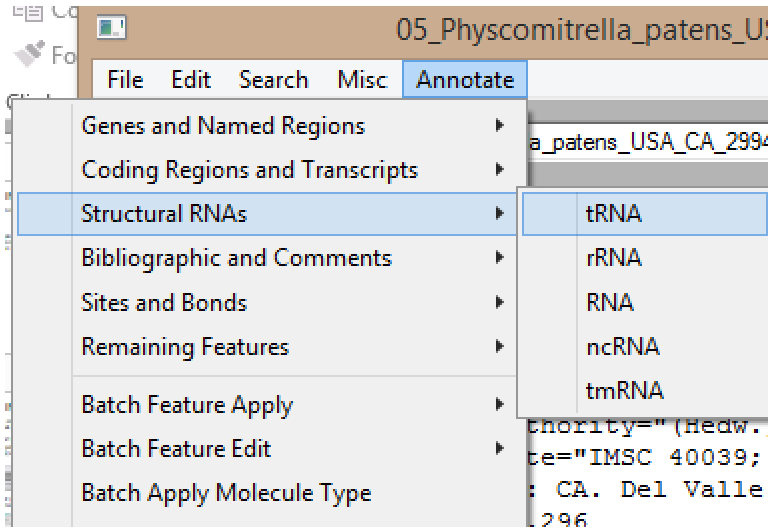
○ Sequence Coordinates ● Alignment Coordinates

# SEQUIN

Example 3: annotation of a tRNA

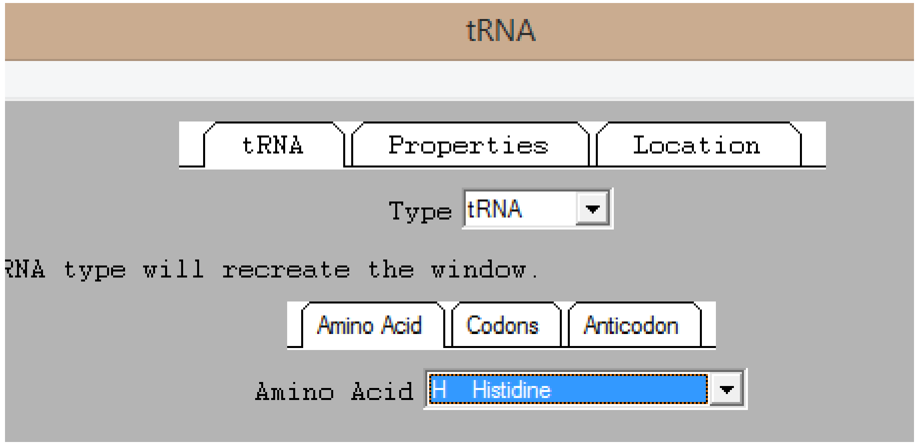
The last feature of our example is a fragment of the tRNA-His gene

tRNAs are considered in Sequin “structural RNAs”



```
a_patens_USA_CA_2994
thority="(Hedw.
te="IMSC 40039;
: CA. Del Valle
296
```

Specify the aminoacid and, if possible, the codon

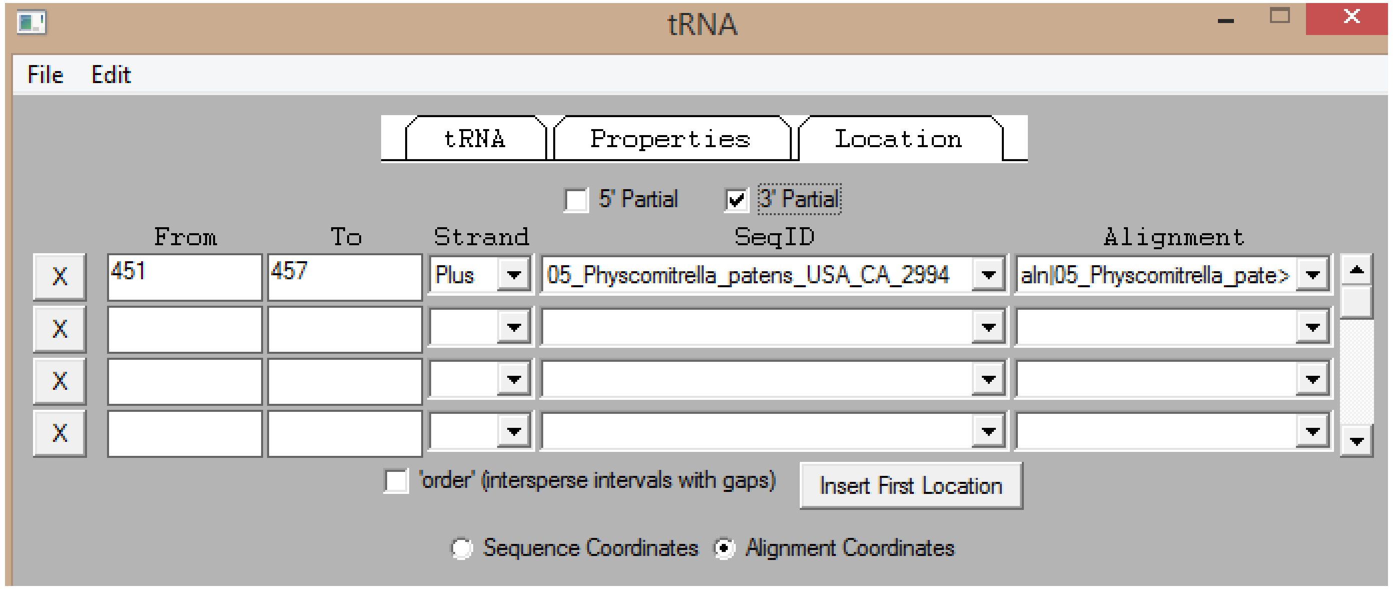


Type: tRNA

RNA type will recreate the window.

Amino Acid: H Histidine

Then, business as usual: specify limits, completeness, etc



□ 5' Partial ☑ 3' Partial

|  | From | To | Strand | SeqID | Alignment |
|---|---|---|---|---|---|
| X | 451 | 457 | Plus | 05_Physcomitrella_patens_USA_CA_2994 | aln\|05_Physcomitrella_pate> |
| X |  |  |  |  |  |
| X |  |  |  |  |  |
| X |  |  |  |  |  |

□ 'order' (intersperse intervals with gaps)

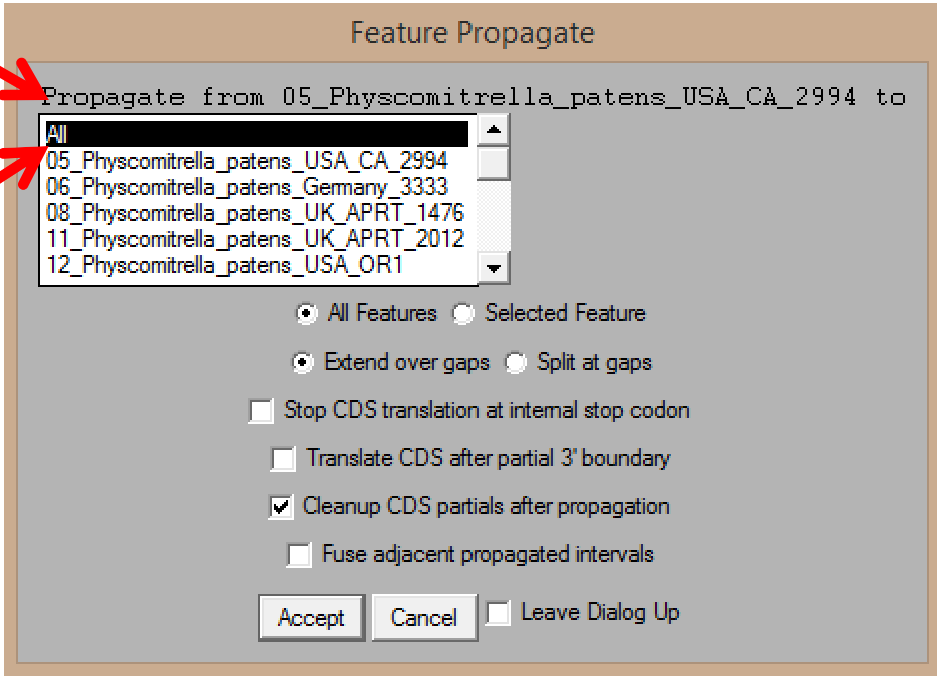
○ Sequence Coordinates ● Alignment Coordinates

# SEQUIN

Once you have your chosen sequence completely annotated, propagate those annotations across the whole alignment (Edit > Feature Propagate)

Your annotated sequence

To all the alignment



Propagate from 05_Physcomitrella_patens_USA_CA_2994 to

- All
- 05_Physcomitrella_patens_USA_CA_2994
- 06_Physcomitrella_patens_Germany_3333
- 08_Physcomitrella_patens_UK_APRT_1476
- 11_Physcomitrella_patens_UK_APRT_2012
- 12_Physcomitrella_patens_USA_OR1

● All Features ○ Selected Feature

● Extend over gaps ○ Split at gaps

□ Stop CDS translation at internal stop codon

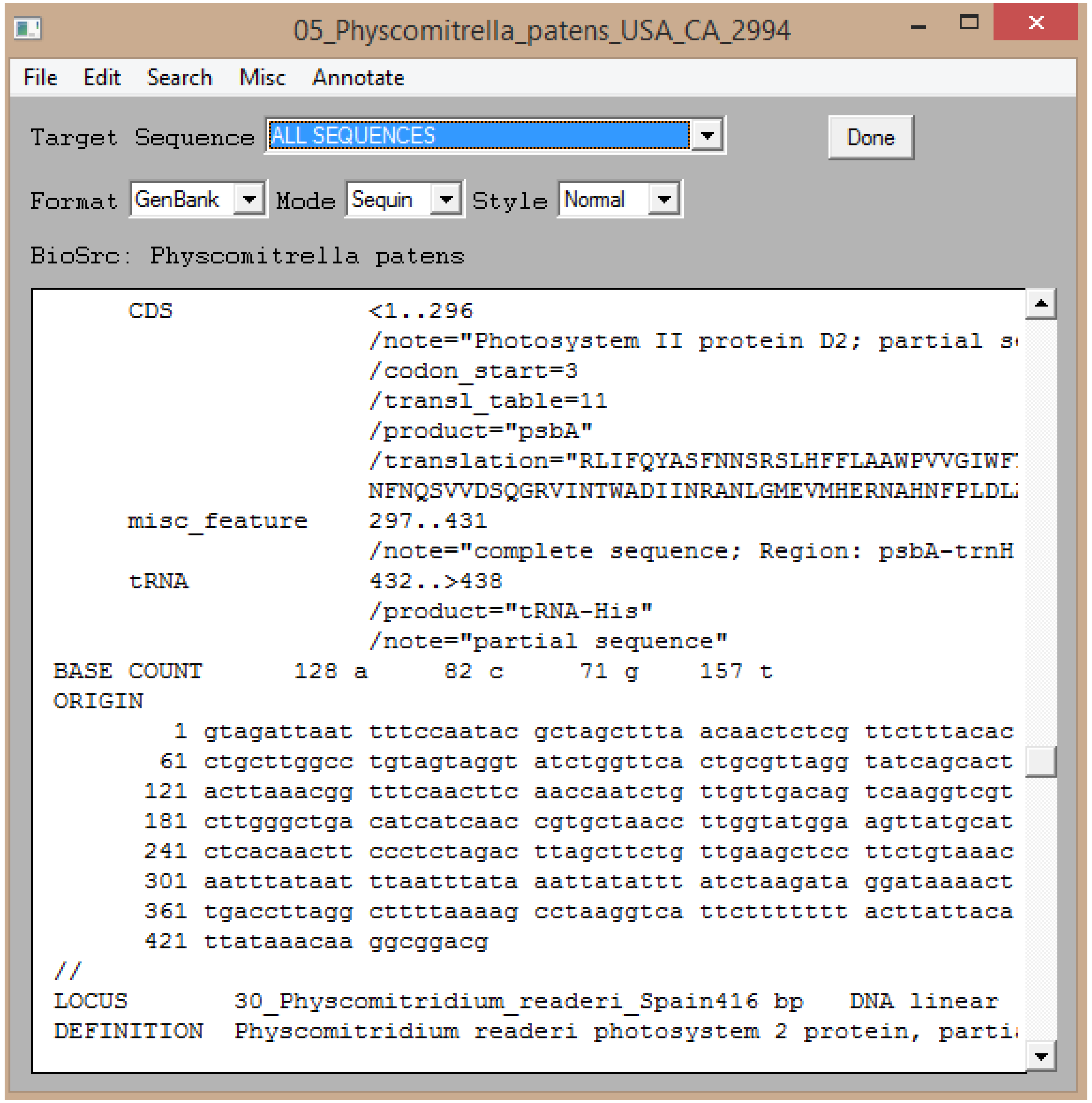
□ Translate CDS after partial 3' boundary

☑ Cleanup CDS partials after propagation

□ Fuse adjacent propagated intervals

□ Leave Dialog Up

Then, check that the other sequences seem to have acquired the annotations correctly



Target Sequence: ALL SEQUENCES

Format: GenBank Mode: Sequin Style: Normal

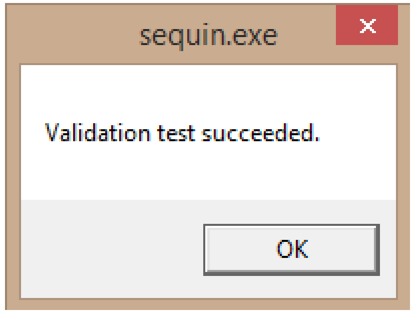
BioSrc: Physcomitrella patens

```
     CDS             <1..296
                     /note="Photosystem II protein D2; partial s
                     /codon_start=3
                     /transl_table=11
                     /product="psbA"
                     /translation="RLIFQYASFNNSRSLHFFLAAWPVVGIWF
                     NFNQSVVDSQGRVINTWADIINRANLGMEVMHERNAHNFPLDL
     misc_feature    297..431
                     /note="complete sequence; Region: psbA-trnH
     tRNA            432..>438
                     /product="tRNA-His"
                     /note="partial sequence"
BASE COUNT      128 a     82 c     71 g    157 t
ORIGIN
        1 gtagattaat tttccaatac gctagcttta acaactctcg ttctttacac
       61 ctgcttggcc tgtagtaggt atctggttca ctgcgttagg tatcagcact
      121 acttaaacgg tttcaacttc aaccaatctg ttgttgacag tcaaggtcgt
      181 cttgggctga catcatcaac cgtgctaacc ttggtatgga agttatgcat
      241 ctcacaactt ccctctagac ttagcttctg ttgaagctcc ttctgtaaac
      301 aatttataat ttaatttata aattatattt atctaagata ggataaaact
      361 tgaccttagg cttttaaaag cctaaggtca ttcttttttt acttattaca
      421 ttataaacaa ggcggacg
//
LOCUS       30_Physcomitridium_readeri_Spain416 bp    DNA linear
DEFINITION  Physcomitridium readeri photosystem 2 protein, parti
```

# SEQUIN

You are almost ready for submission. Cross your fingers and validate the file (Search>Validate)

If everything is ok, you will see something like this:

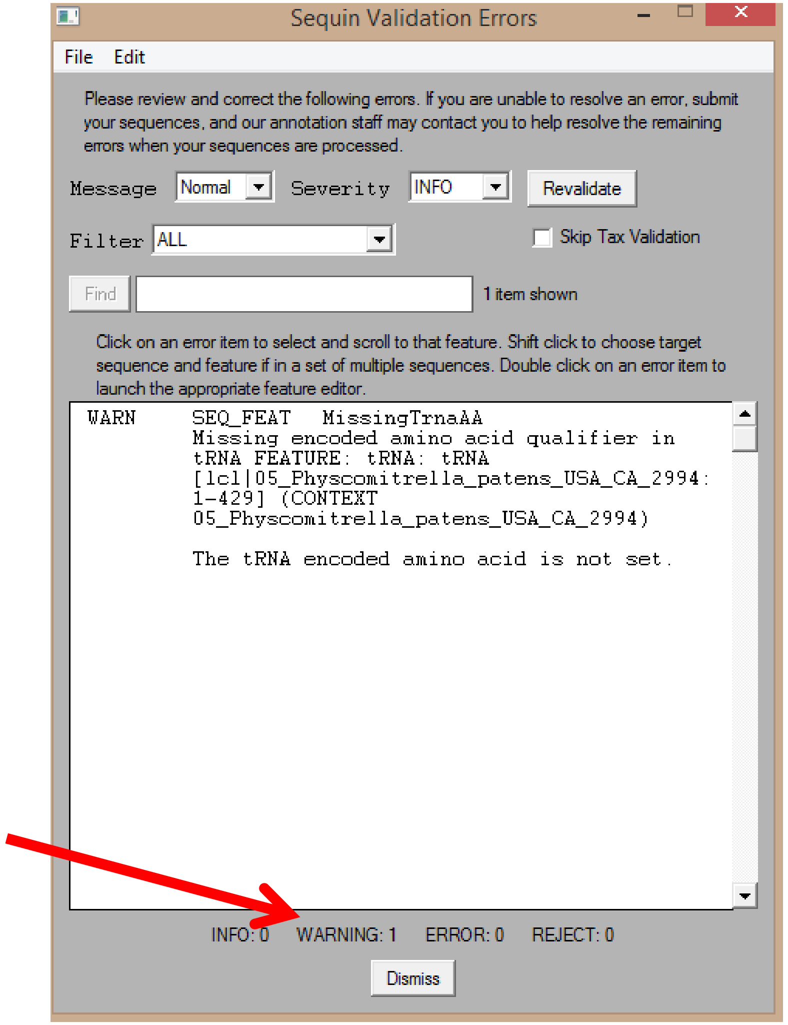

sequin.exe

Validation test succeeded.

Then you just need to export the sequin file: choose “Prepare submission” in the File menu and save your record. Send it by email to

**gb-sub@ncbi.nlm.nih.gov**

Alternatively you will receive some error messages with different degrees of “severity”


Sequin Validation Errors

Please review and correct the following errors. If you are unable to resolve an error, submit your sequences, and our annotation staff may contact you to help resolve the remaining errors when your sequences are processed.

Message Normal Severity INFO

Filter ALL

□ Skip Tax Validation

1 item shown

Click on an error item to select and scroll to that feature. Shift click to choose target sequence and feature if in a set of multiple sequences. Double click on an error item to launch the appropriate feature editor.

```
WARN     SEQ_FEAT   MissingTrnaAA
         Missing encoded amino acid qualifier in
         tRNA FEATURE: tRNA: tRNA
         [lcl|05_Physcomitrella_patens_USA_CA_2994:
         1-429] (CONTEXT
         05_Physcomitrella_patens_USA_CA_2994)

         The tRNA encoded amino acid is not set.
```

INFO: 0 WARNING: 1 ERROR: 0 REJECT: 0

Even the warnings or lesser errors can give you trouble after your submission. It is better to fix them all and revalidate the submission. Often this means to start the whole process again, so be patient and meticulous!