

Submitting RCC sequences to Genbank with Geneious

Daniel Vaultot & Adriana Lopes dos Santos

Version 2.0 - 23 08 2018

Contents


1	Aim of document	1
2	Assemble and clean sequences	2
3	Add informations to sequences	9
3.1	Taxonomy	9
3.2	Gene annotation	15
3.3	Metadata	17
3.4	Primers information	19
4	GenBank submission - General case (not for 16S, 18S or ITS, see next part)	21
5	GenBank submission - 16S, 18S or ITS	27
5.1	Prepare files	28
5.2	Submit to NCBI web portal	36
6	Appendixes	49
6.1	Retrieve sequences from Genbank using Geneious	49

1 Aim of document

This document explains how to use Geneious to :

- assemble and clean final sequences from several traces using different internal primers
- annotate the sequences
- submit to Genbank using Bankit through the Geneious plug-in
- submit to Genbank for 18S, ITS and 16S that cannot be any more be submitted using Bankit

Notes

- Look at legends **below** screen captures for directions.
- Changes from previous versions have been labelled with 

2 Assemble and clean sequences

- Import the ab1 trace
 - Drag and Drop



Figure 1: Import trace sequences

- Trim the sequences
 - Annotate & Predict -> Trim Ends
 - Use error probability limit from 0.01 to 0.02 (increase to 0.05 if cannot assemble correctly, the trimming will be less drastic). For single reads (e.g. 528F) use a maximum of 0.02.

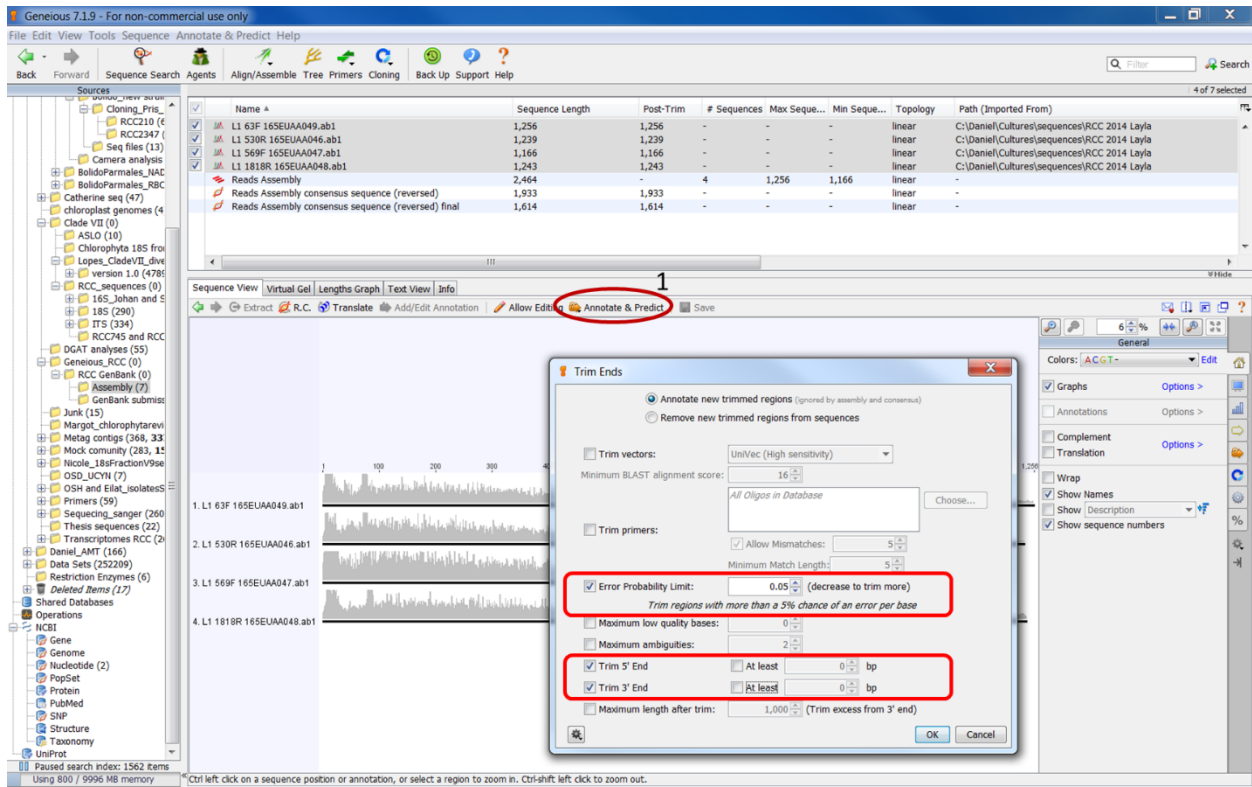


Figure 2: Trim sequences

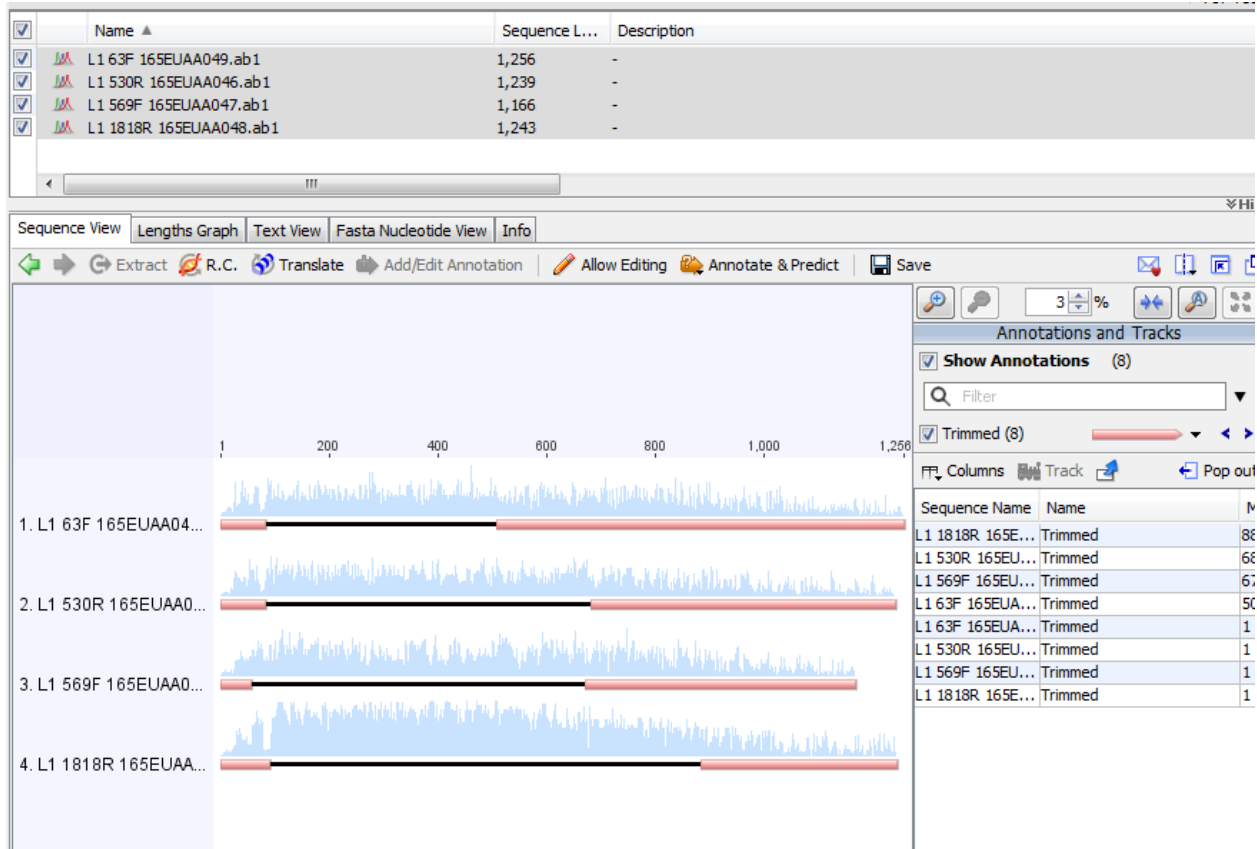


Figure 3: Visualize the trimmed sequences

- Assemble if several primers have been used
 - **Align/Assemble/De Novo Assemble**
 - Use for assembly name: `RCC####_gene-name_your-initials_date`
 - * e.g. `RCC2497_18S_PG_2018_02_15`
 - ⚠ The name should not contain any space
 - Select “save the consensus”
 - Select “save contigs”.
 - You may have to change the trimming level (increase probability level - see above) if traces cannot be assembled

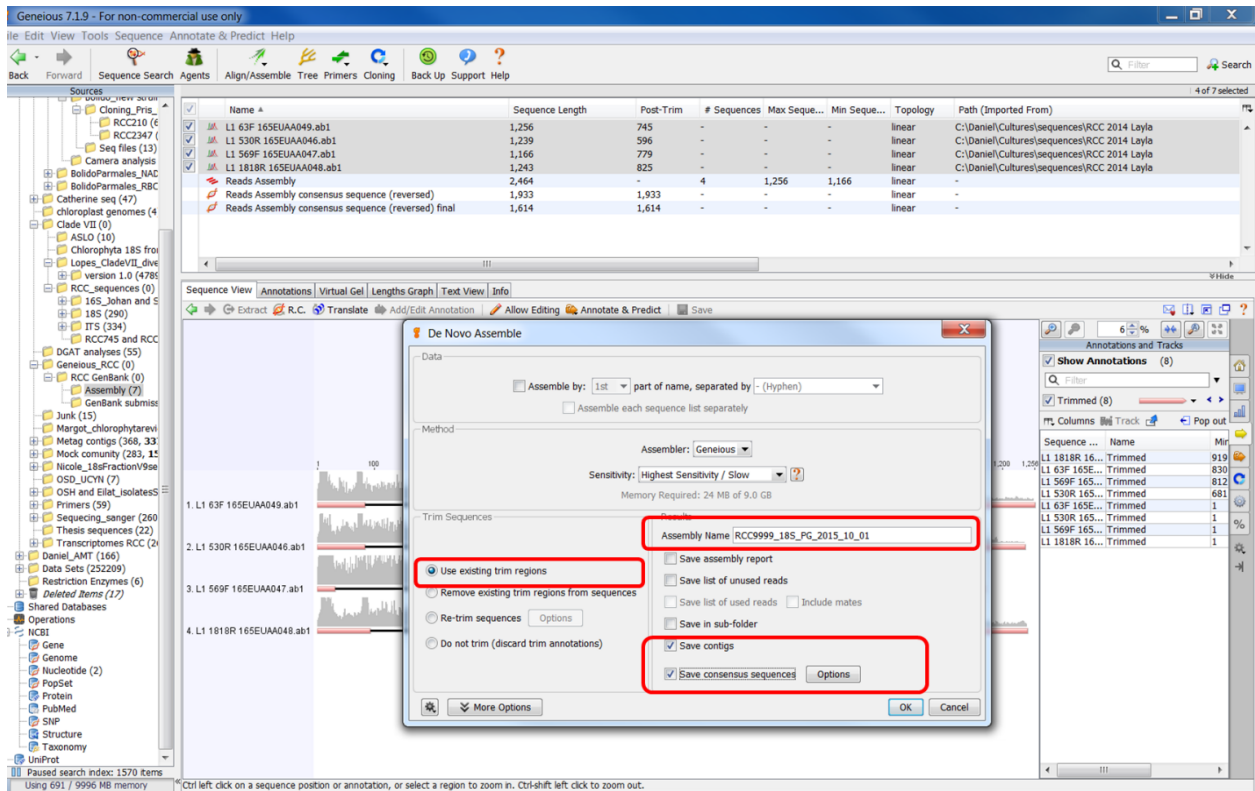


Figure 4: Assemble sequences

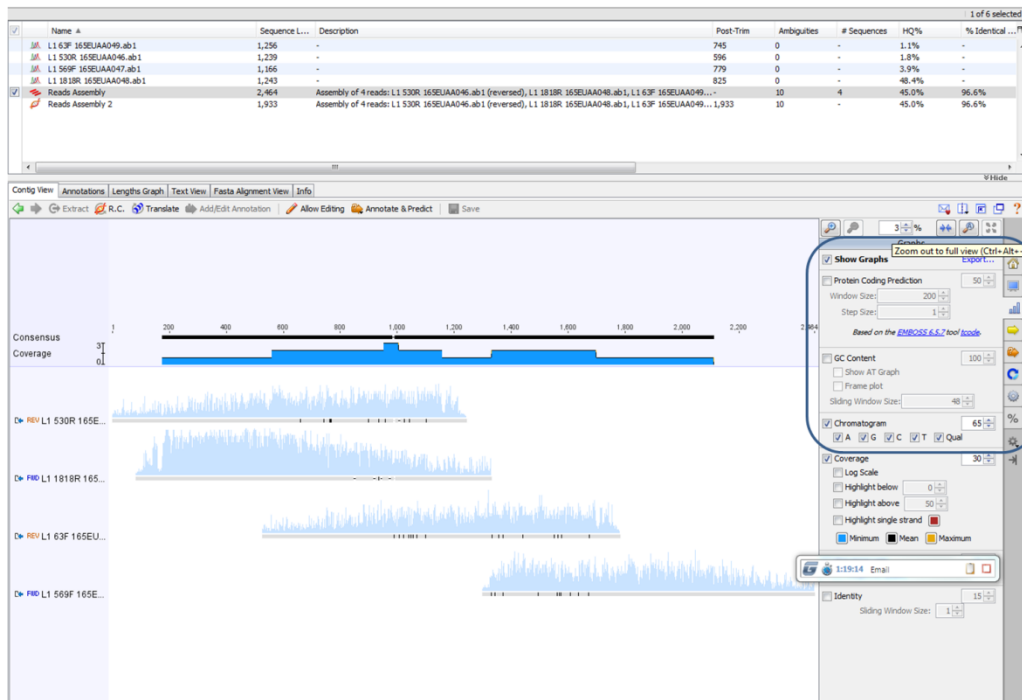


Figure 5: Visualize the assembled sequences

- Check the assembly and edit the consensus if necessary.
 - ⚠ This is very important to make sure that your sequence is clean.
 - Allow editing
 - Edit bases that maybe wrongly assigned in one the trace.

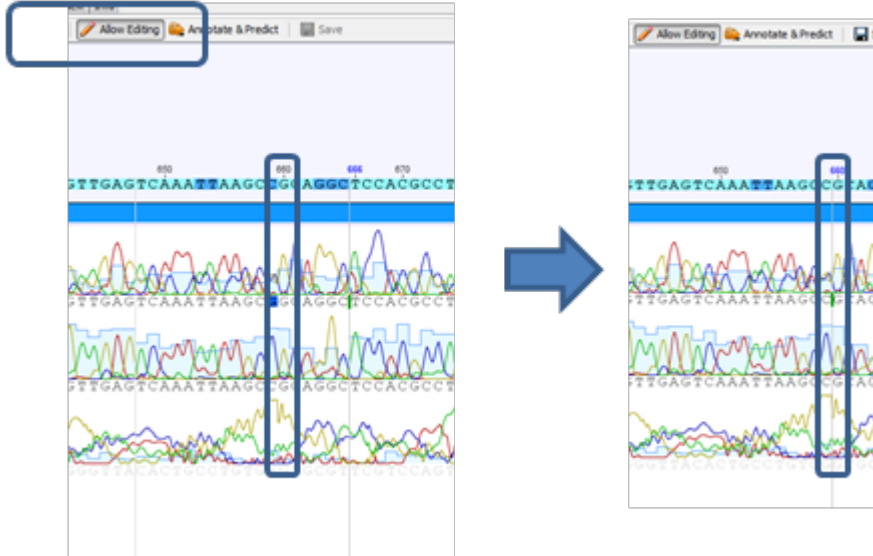


Figure 6: Check and correct assembly

- Select and extract consensus

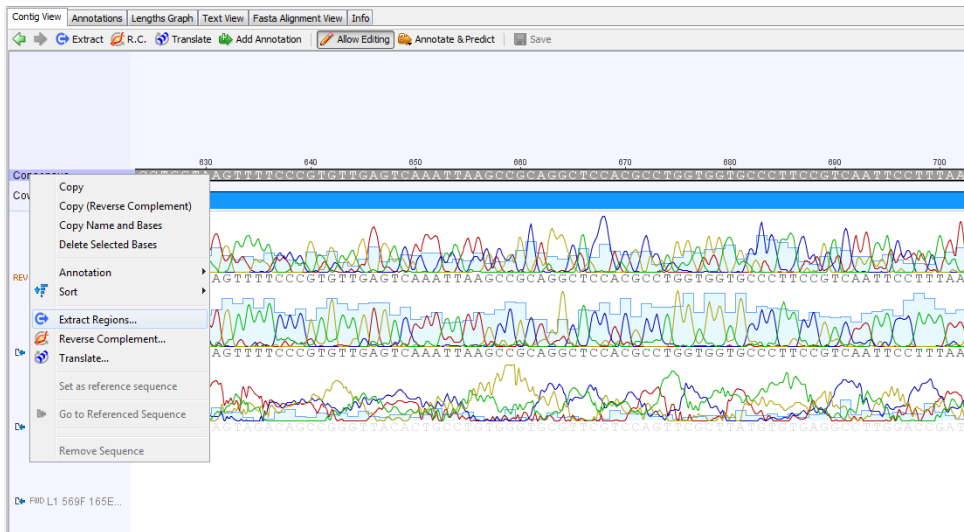


Figure 7: Extract consensus

- Reverse complement if necessary (if the sequence was assembled the other way around).
- Locate primers, test forward and reverse separately.
 - Tools > Primers > Test with saved primers

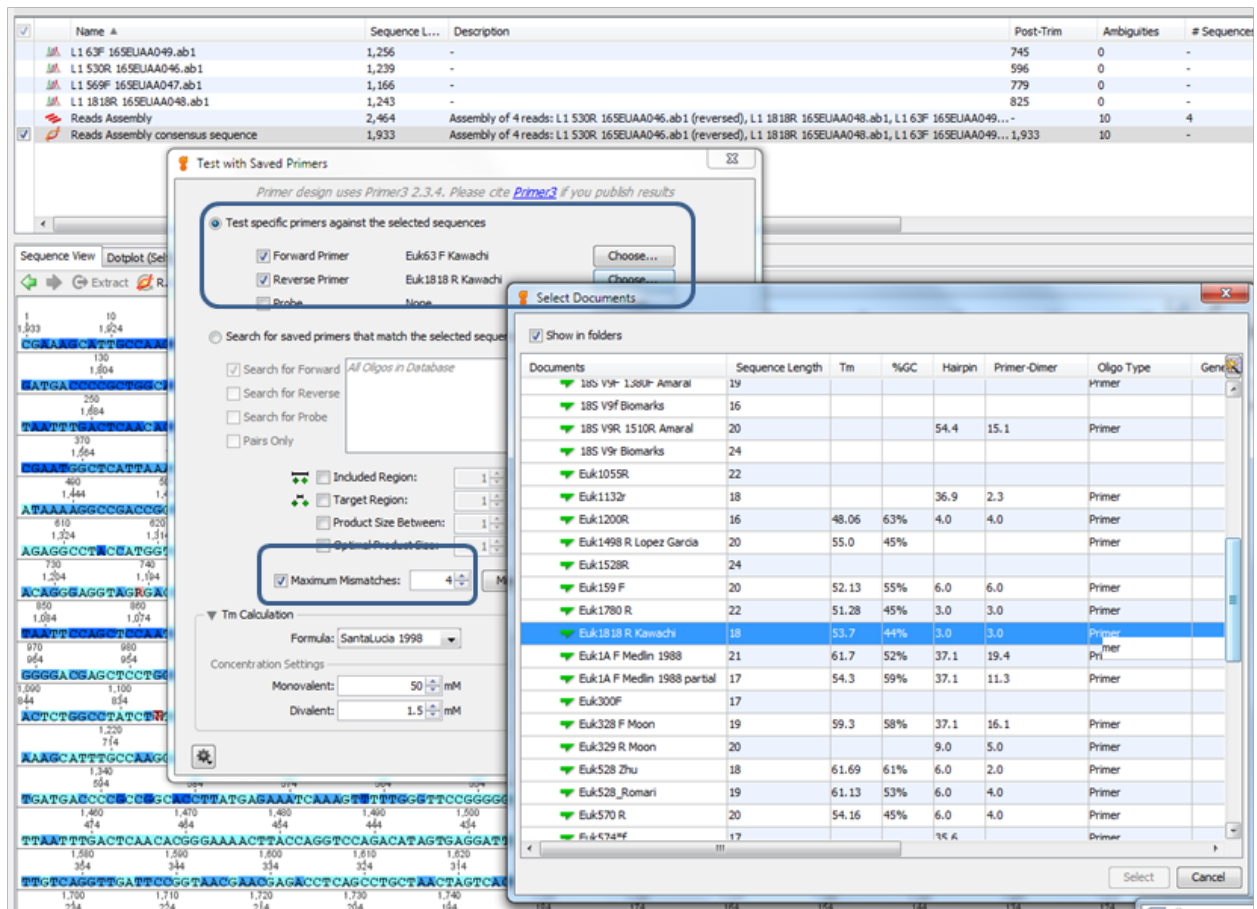


Figure 8: Test with saved primers

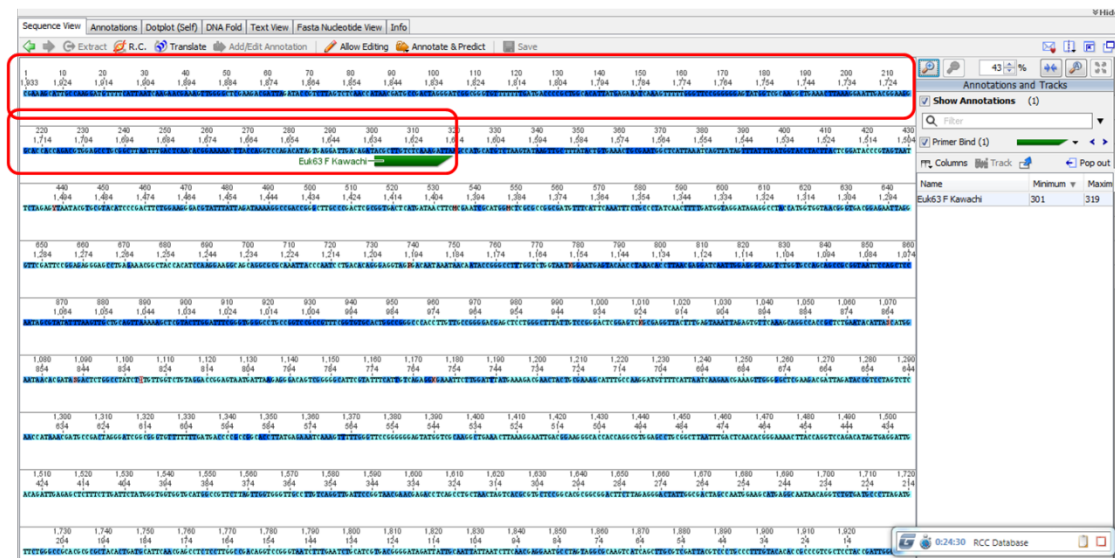


Figure 9: Locate primers

- Remove everything which is outside of primers **including the primers**.
 - allow editing
 - pressing the left button of the mouse, mark the region to be delete, It will show in green
 - press delete



Figure 10: Remove everything outside primers

- Et voilà , you have a clean sequence
 - The coloring corresponds to sequence quality based on the traces and assembly.

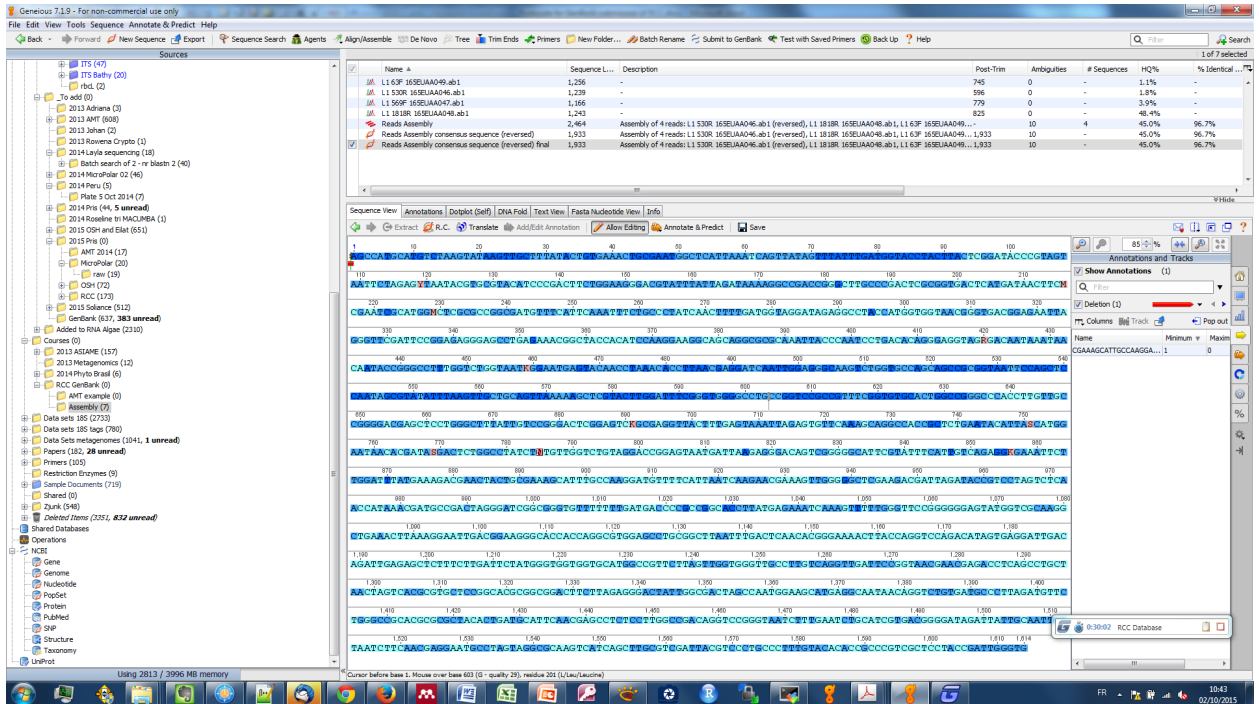


Figure 11: Final sequence

3 Add informations to sequences

3.1 Taxonomy

- Do a batch BLAST search
 - Select the files > sequence search or Blast search
 - Sometimes it does not work so you can do with the NCBI BLAST server
 - Pay attention to the following parameters:
 - * **database** : nr genbank
 - * **program** : blastn (for protein coding gene like rbcL the blastX can be also use to confirmed)
 - * **results** : hit table
 - * **maximum hits** : 25 at least
 - Parameters can be saved, recalled and deleted by clicking at the bottom of dialog box
 - * 'save current settings > name > save.
 - You can request from NCBI an API key which increase the number of request you can do. The process is explained on the NCBI web site

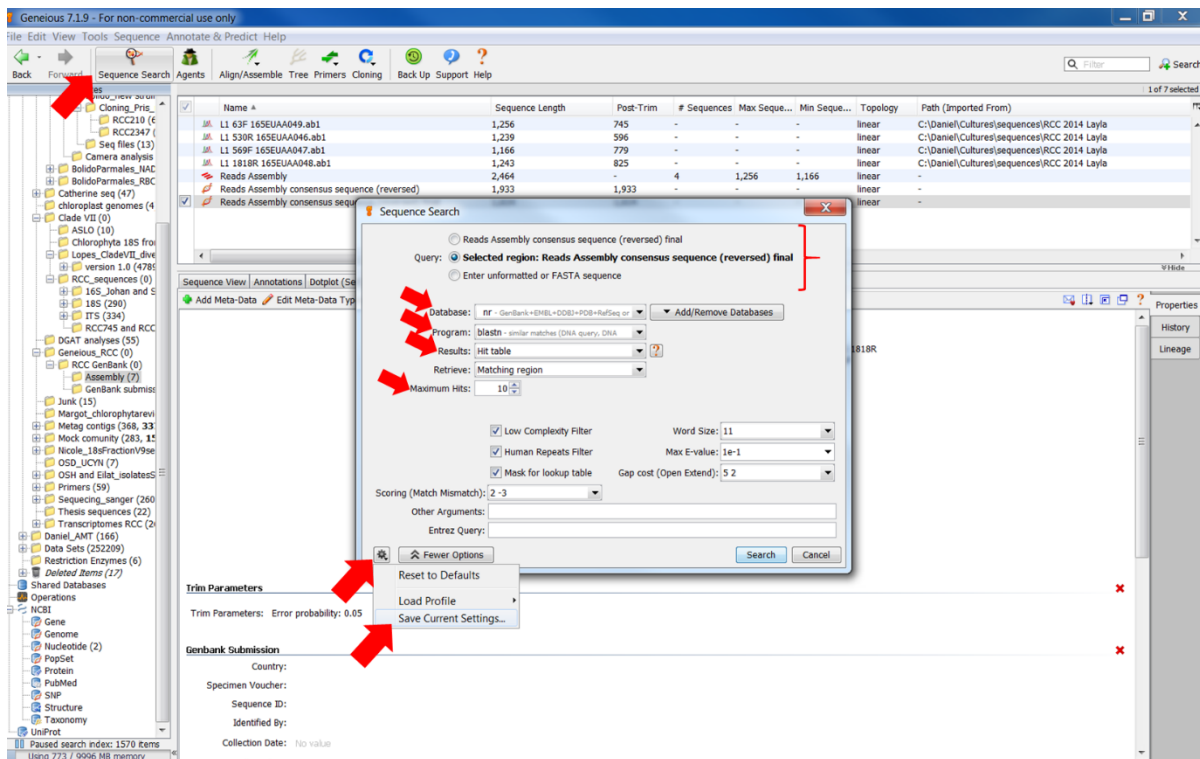


Figure 12: BLASTN

NCBI Resources How To vaulet My NCBI Sign Out

My NCBI » Settings [help](#)

✔ Your API Key has been created successfully.

NCBI Account Settings

Email

vaulet@sb-roscoff.fr (confirmed) Change

This email is used for delivery of saved searches and recovery of password for your native NCBI account.

Native NCBI Account *The following username and password is maintained by NCBI.*

Username:	vaulet	Change
Password:	*****	
Security Question:	What is your contact email address?	Change

Linked accounts *You can sign in via these 3rd-parties. Contact the 3rd party for sign-in related issues.*

None Change

Delegates

You can add delegates to help you manage your bibliography and/or SciENcv profiles.
[Add a Delegate](#)

API Key Management

API Key	Replace	Delete
87ce6407215b5d1b6f5ee3ce8a6703793608		

Click here

Generate API

Figure 13: Request a NCBI key

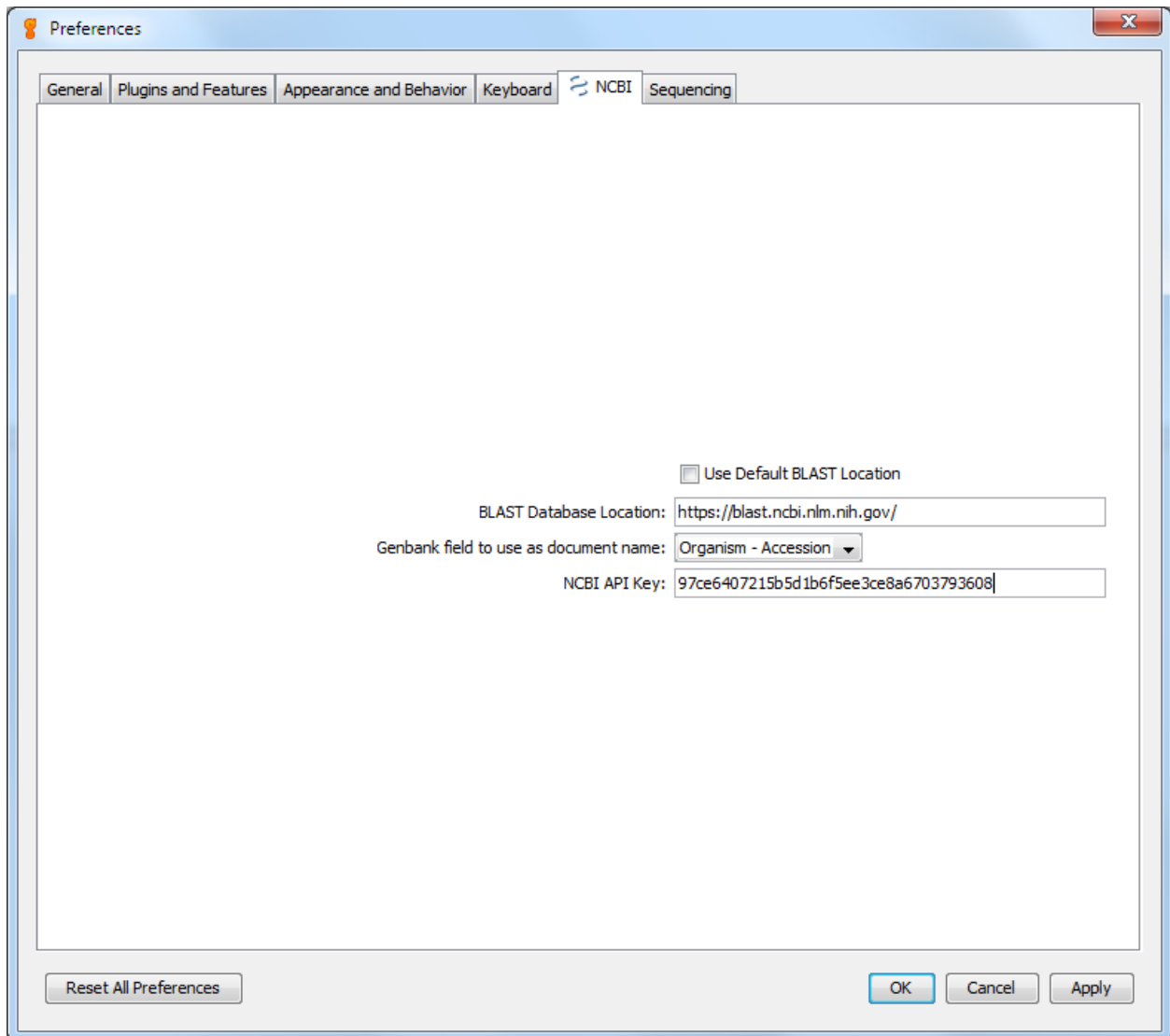


Figure 14: Enter they key in Geneious preferences

- Retrieve the closest sequence from GenBank (*Optional*)
 - From Geneious folder with the Blast results, select the closest result, drag the file into your folder in your local database if you wish to retain the file and/or modify it.
 - From Genbank: Copy the accession number > go to NCBI > nucleotide > paste the accession number (look the figure bellow). You drag the file into your folder in your local database if you wish to retain the file and/or modify it.

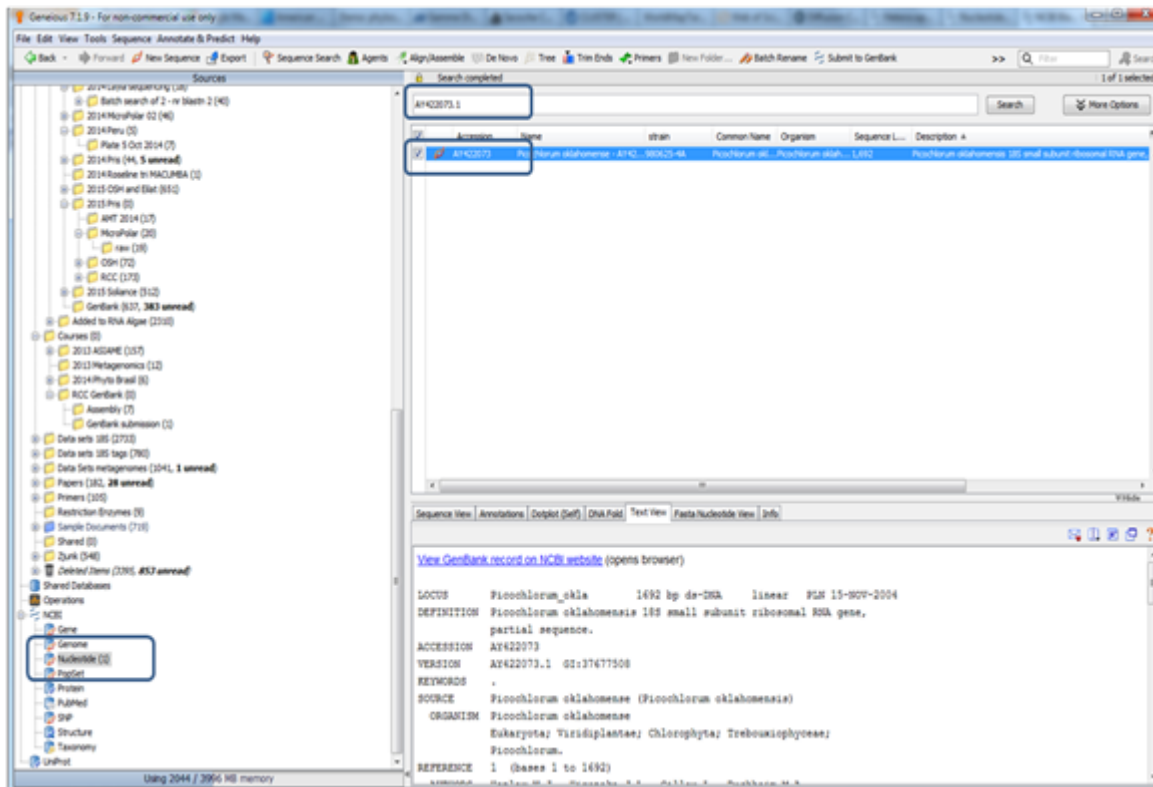


Figure 15: Retrieve closely related sequences from Genbank

- Do a manual alignment (*Optional*)
 - This is very useful to detect introns, for ITS sequences, combine gene sequencing partial 18S + 28S for example.
 - Align/Assemble > Pairwise Align MAFFT using the default parameters

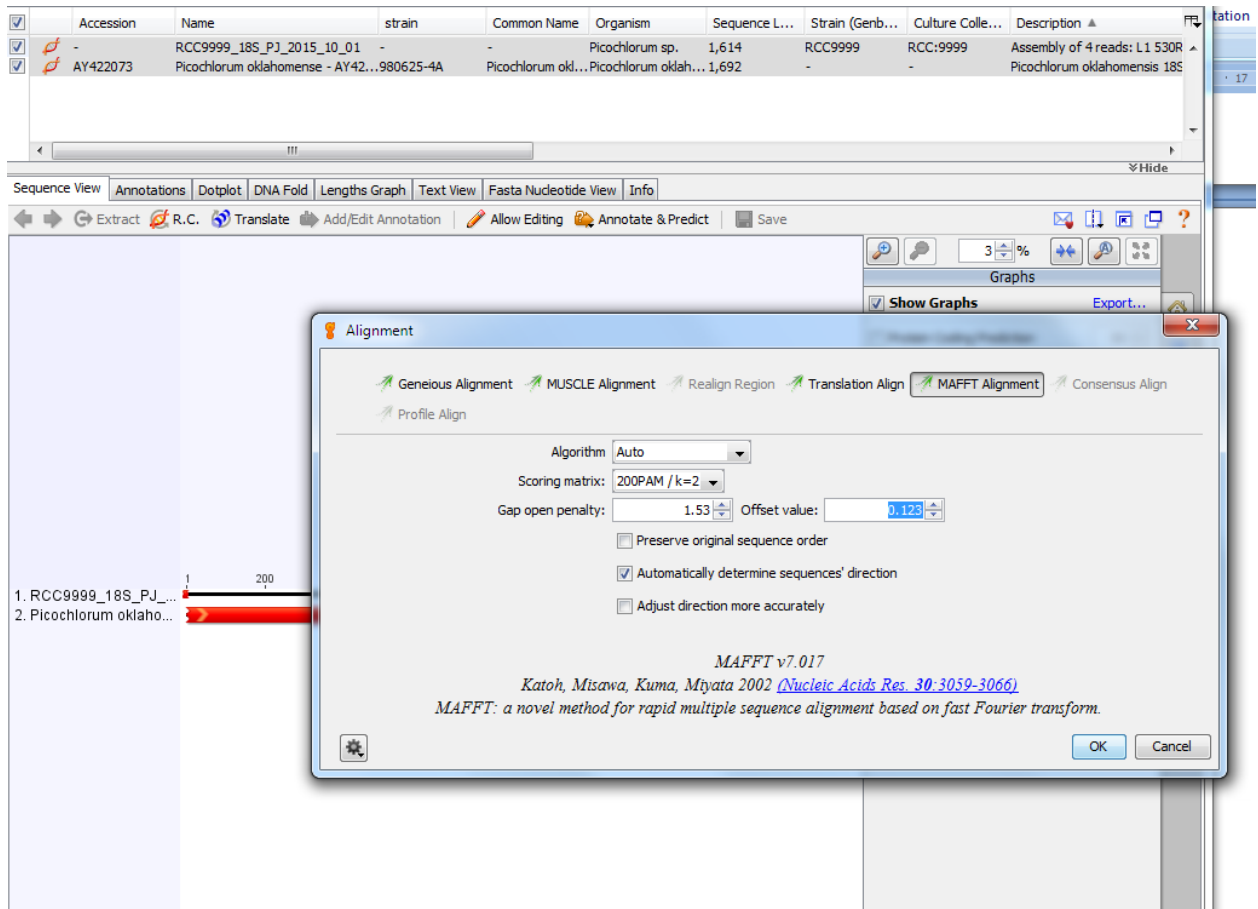


Figure 16: Alignment parameters

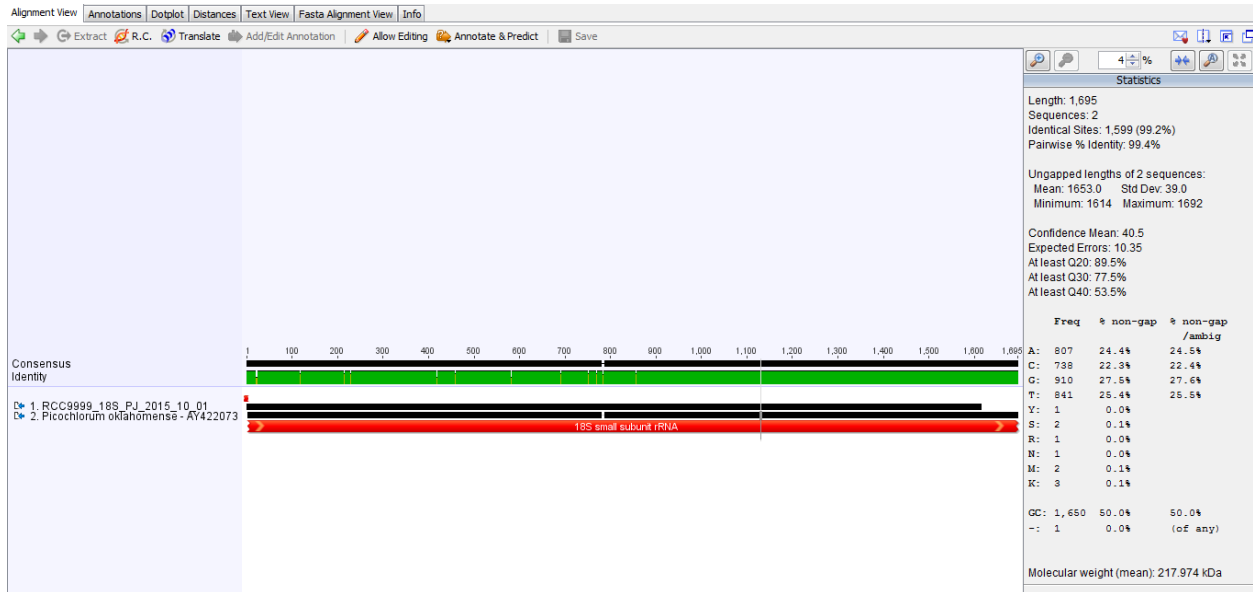


Figure 17: Alignment results

3.2 Gene annotation



This step is NOT necessary for 16S, 18S, ITS

- With the mouse, select your sequence, add notation
- Parameters to be changed (look the picture below)
 - **Name:** name of the gene
 - **Type:**
 - * select rRNA for 18S, ITS, 16S platidial and 28S
 - * CDS or gene coding sequence for example rbcL
- Add property using the 1st ADD: name = product, value = name of gene, for example 18S rRNA.
- Add annotation using the 2nd ADD (click in INTERVALS to see it): click in truncated left end and truncated right end. This is to tell that the sequence is not complete. For example, the 18S in this tutorial had the extremities before the primer removed, so is incomplete.



Make sure you do not have two annotations for the same gene !

Accession	Name	strain	Common Name	Organism	Sequence L...	Strain (Gerb...	Culture Colle...	Description	Notes
-	Nucleotide alignment	-	-	-	1,695	-	-	Alignment of 2 sequences: RCC9999_18S_P1_2015_10_01, Picochlorum ...	-
RCC9999_18S_P1_2015_10_01	-	-	-	Picochlorum sp.	1,614	RCC9999	RCC:9999	Assembly of 4 reads: L1 530R, 165EUAA046.ab1 (reversed), L1 1818R 1...	Here anything that you need to
AY422073	Picochlorum oklahomense - AY42...980625-4A	-	Picochlorum okl...	Picochlorum oklah...	1,692	-	-	Picochlorum oklahomensis 18S small subunit ribosomal RNA gene, partial ...	-

Figure 18: Annotate genes

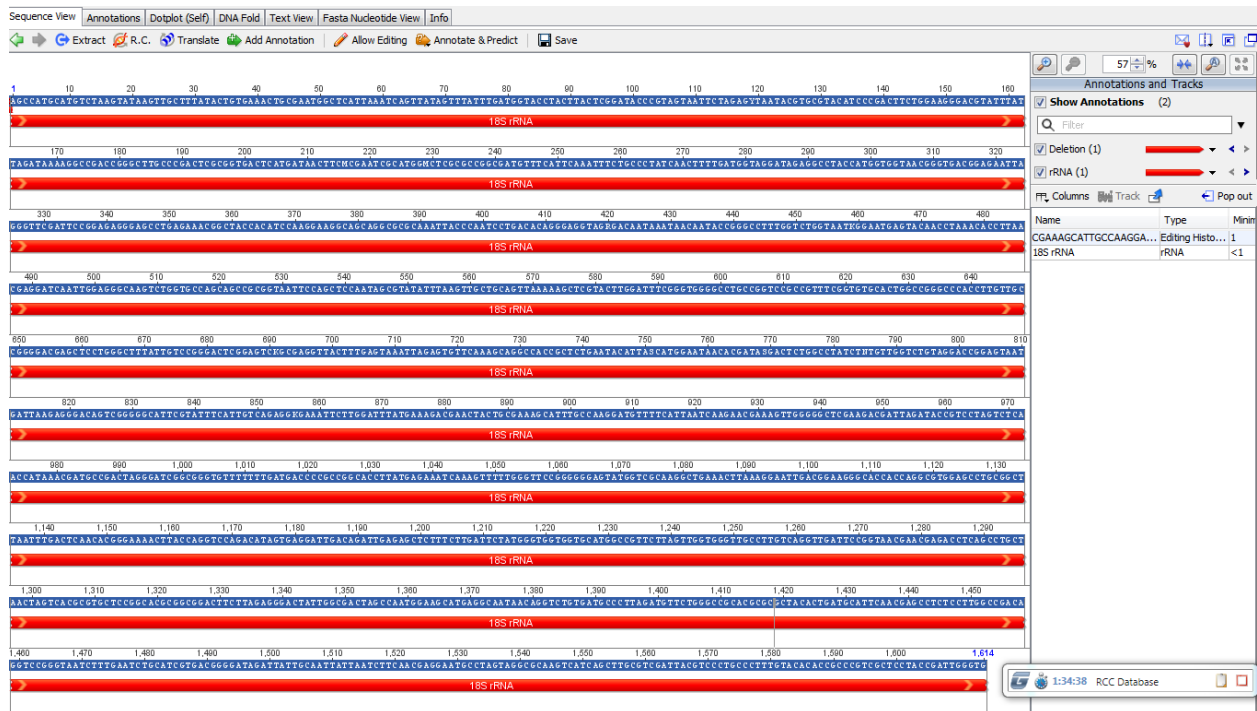


Figure 19: Annotated genes

3.3 Metadata

- Add two new type of metadata (it has to be done only once) in the GenBank submission category:
 - **Strain**
 - **Culture_collection** Edit Meta data Types > Genbank Submission > click on the + on the right side > write Culture Collection on the new field - > ok



Make sure that these new fields are in the Genbank Submission category. Do not recreate a new category.



Use exactly the orthograph for names especially with underscores “Culture_collection” and not as before “Culture Collection”.

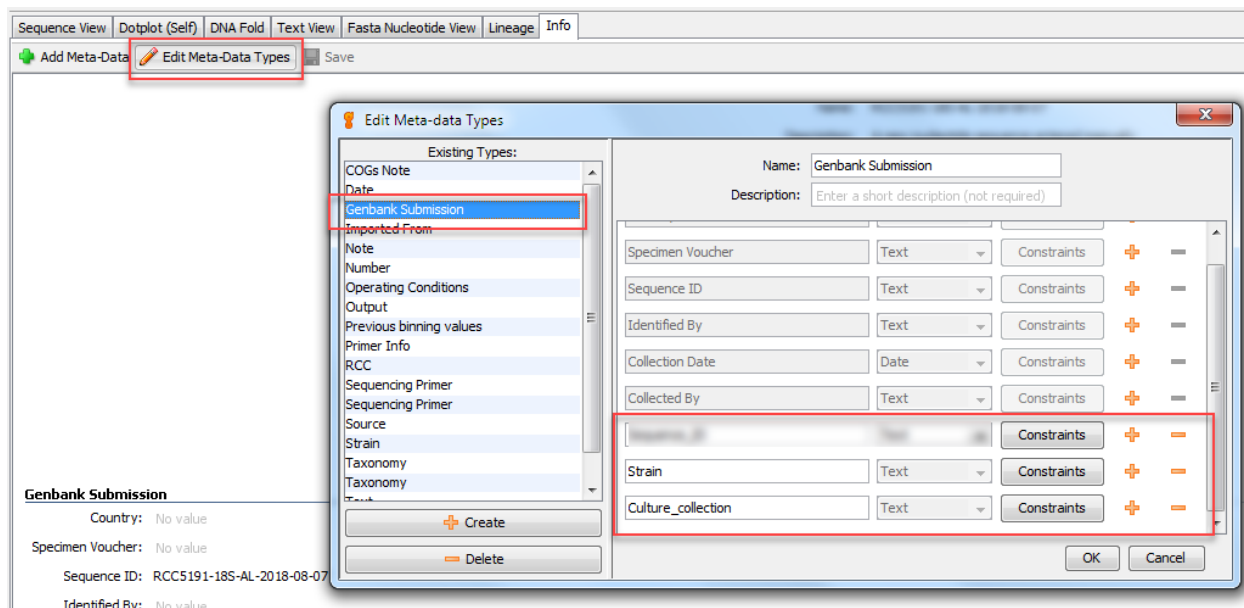




Figure 20: Add new meta-date type: Culture Collection

- Click on the final sequence, go to info and change or correct the following fields.
 - **Name** : RCC####_gene-name_your-initials_date, e.g. RCC9999_18S_PG_2015_10_01 (change if it is not in this format at this point).
 - * This will be the ID of the sequence submitted to GenBank.
 - * This name must not contain any space
 - * This name must be unique. For example if you submit 2 sequences for the same strain and same gene you must use different names e.g.RCC9999_18S_PG_2015_10_01_A and RCC9999_18S_PG_2015_10_01_B
 - **Organism** : *Picochlorum* sp. or *Trebouxiophyceae*.
 - * Enter the genus name or, if not known, the lowest taxonomic level known.
 - *  Only use the species name if **you are absolutely sure** of the species as determined by microscopy or ITS. Do not rely on BLAST!!
 - * DO NOT add the RCC number at the end of the organism name. 
 - * For levels above the genus, do not use sp. For example use *Trebouxiophyceae* and not

- Trebouxiophyceae sp. or Chlorophyta and not Chlorophyta sp. NEW
- **Strain** : This is the RCC code as **RCCxxxx** without space between RCC and number e.g. RCC1236.
 - **Culture_collection** : This is the RCC number as **RCC:xxxx** with “:” between RCC and number e.g. RCC:1236.

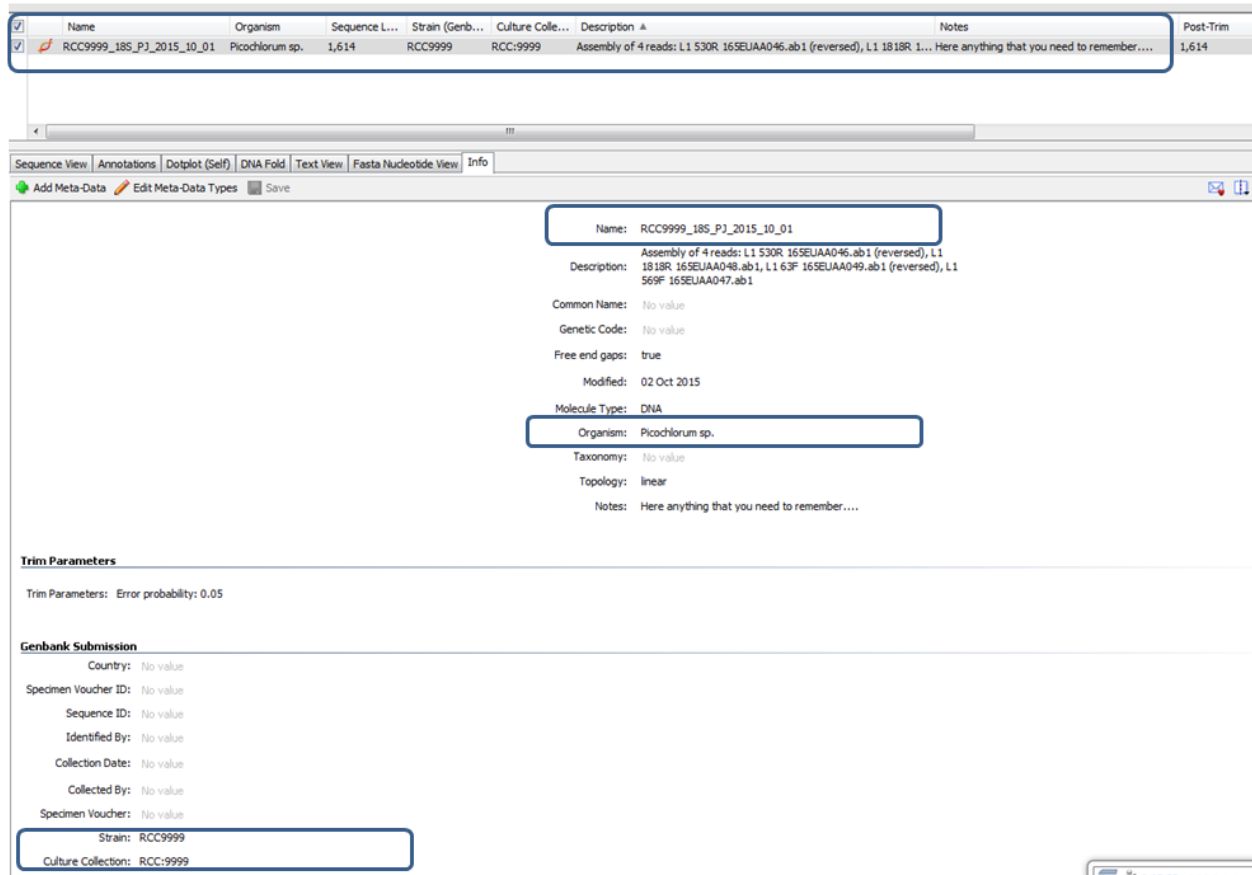


Figure 21: Update the different meta-data fields

- It is possible to quickly change metadata for a set of sequences using the Batch edit mode. For example you can :
 1. Copy the Strain field to the Culture Collection field
 2. Add the “:” automatically for all sequences by replacing “RCC” by “RCC:”.

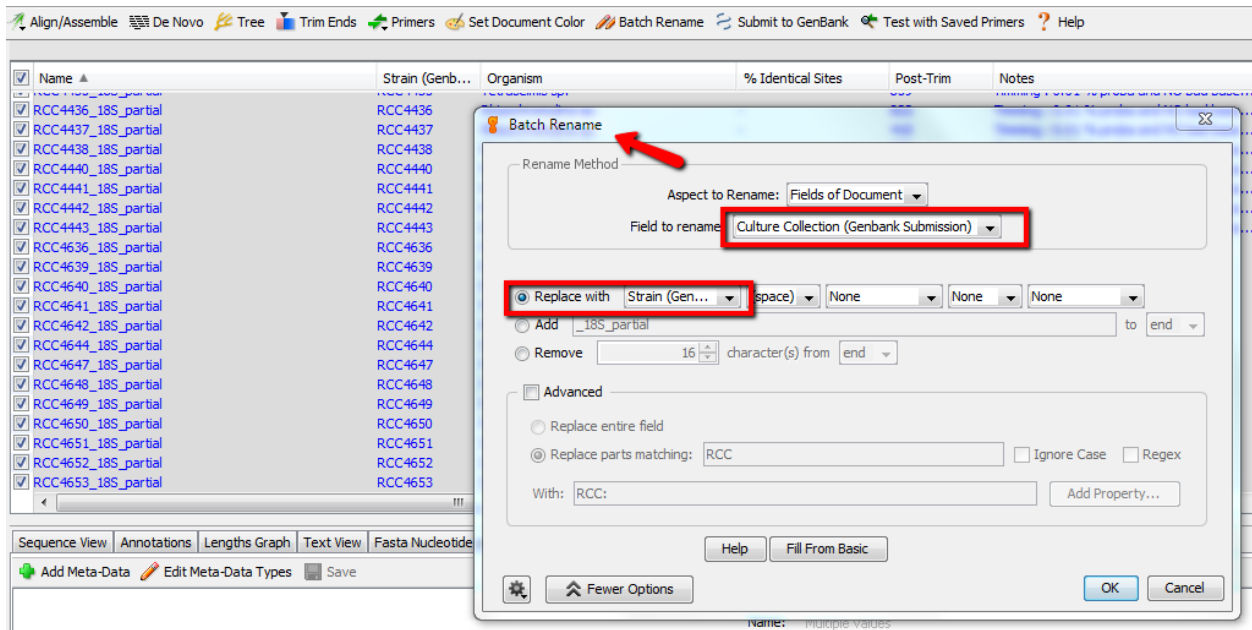


Figure 22: Batch edit - simple

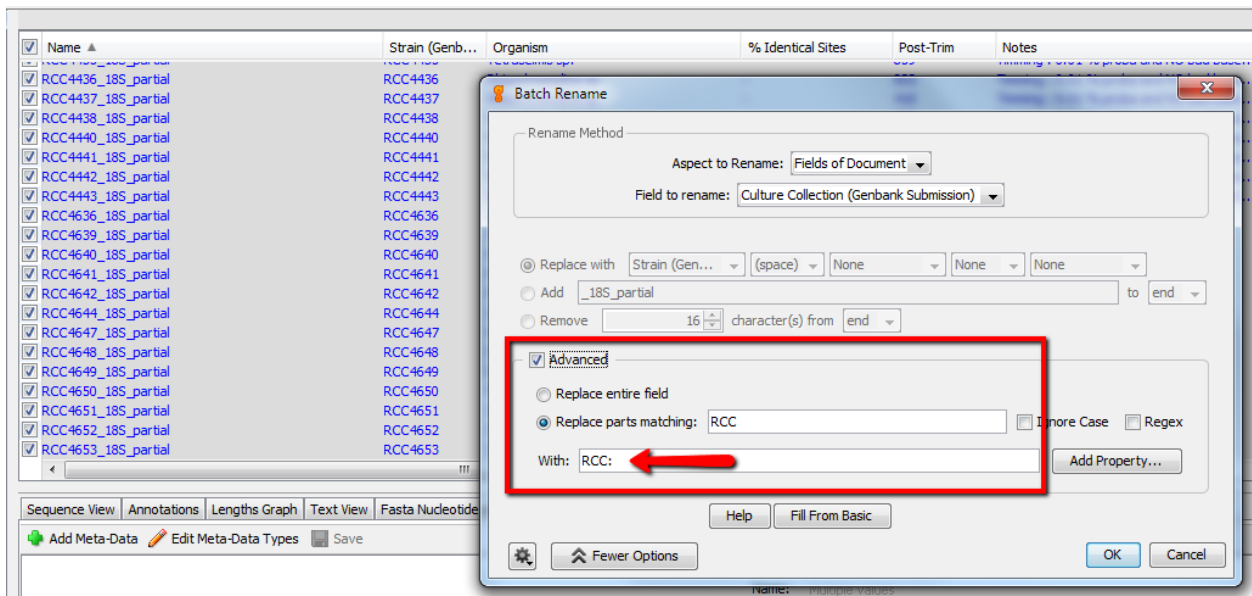


Figure 23: Batch edit - advanced

3.4 Primers information

! This step is Optional for 18S, but must be added for ITS, 28S and other genes

Edit Meta data Types > Sequencing Primer > OK

You can also use Batch edit to go faster

Sequencing Primer

Forward PCR Primer Name: 18S 63F

Forward PCR Primer Sequence: ACGCTTGTCTCAAAGATTA

Reverse PCR Primer Name: 18S 1818R

Reverse PCR Primer Sequence: ACGGAAACCTTGTTACGA

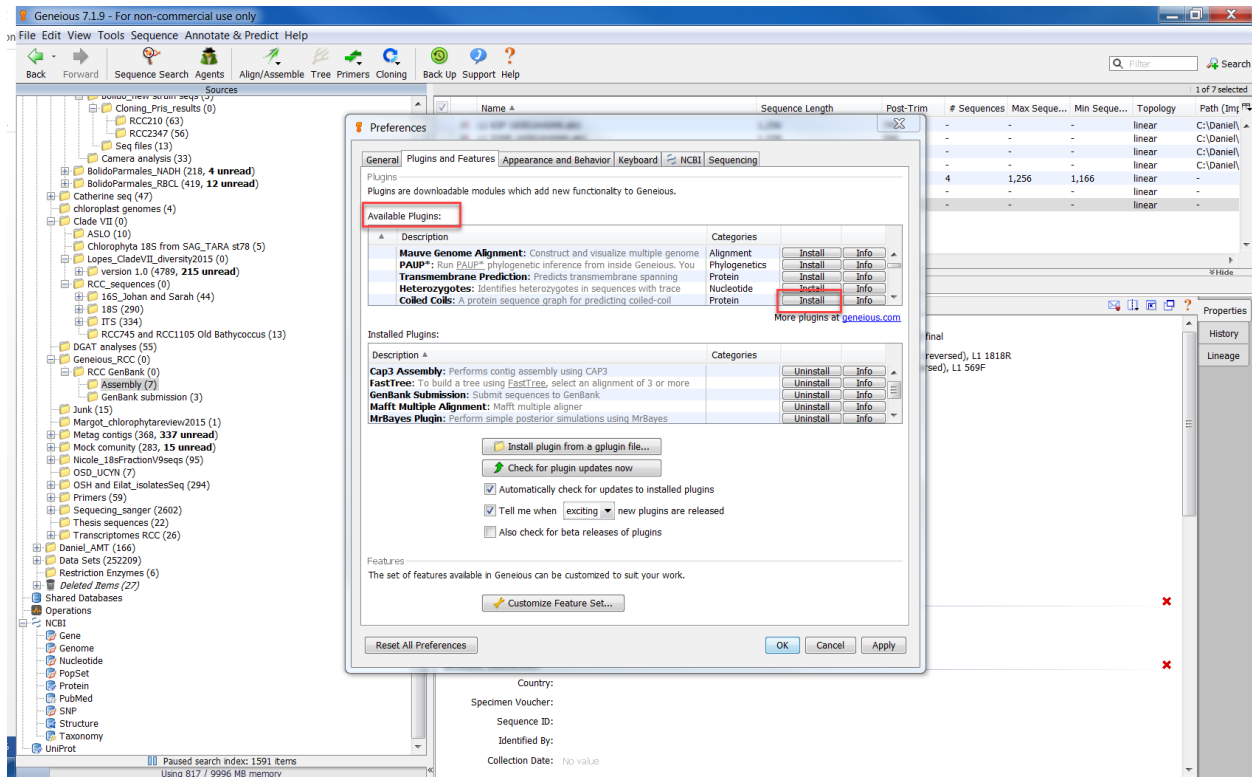
Forward Sequencing Primer Name: No value

Forward Sequencing Primer Sequence: No value

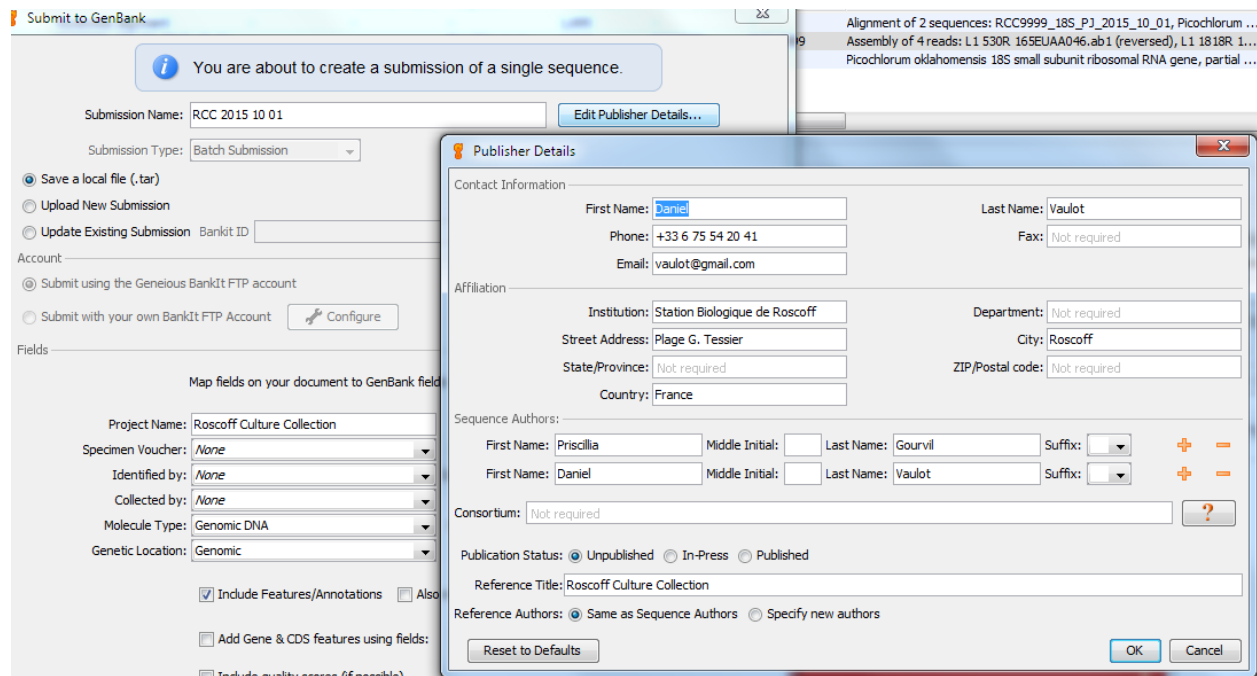
Figure 24: Add primer information to meta-data

4 GenBank submission - General case (not for 16S, 18S or ITS, see next part)

- Note that since August 2018 16S, 18S, 28S and ITS cannot be submitted by BankIt and must be submitted through a web interface.
- Install plugin GenBank submission
 - Tools > plugin > choose the plugin and click in install



- Select the sequences you want to submit
- Select GenBank submission
- Enter first the **Publisher details** (add the info like the picture bellow, except that the sequence authors is Daniel Vaultot + who did the sequence)
 1. **Name**
 2. **email**
 3. **Address**
 4. **Sequence authors**
 5. Select **Unpublished**
 6. **Reference** should be “Roscoff Culture Collection”



- Check very carefully all the fields
 - **Submission name** : the name of the file to be saved (this should be kept on the Databases computer)
 - **Save a local file** (only upload when everything is OK)
 - **Project name** : Roscoff Culture Collection
 - **Molecule type** : Genomic DNA
 - **Genetic location** : in general Genomic but can also be Plastid or Mitochondrion or Nucleomorph for Cryptophytes
 - **Sequence ID** : Name
 - **Organism** : Organism
 - **Include features/annotation** : Yes
 - **Include other fields** : Yes
 - * **Culture_collection** : Culture_collection (GenBank submission)
 - * **Strain** : Strain (GenBank submission).
 - **Primers** : You can put the primers if necessary but they need to be entered Sequencing primers

Submit to GenBank

You are about to create a submission of a single sequence.

Submission Name:

Submission Type:

Save a local file (.tar)

Upload New Submission

Update Existing Submission

Account

Submit using the Geneious BankIt FTP account

Submit with your own BankIt FTP Account

Fields

Map fields on your document to GenBank fields

Project Name: Country:

Specimen Voucher: Sequence ID:

Identified by: Collection Date:

Collected by: Organism:

Molecule Type: Genetic Code:

Genetic Location:

Include Features/Annotations Also include annotations in tracks

Add Gene & CDS features using fields:

Include quality scores (if possible)

Include extra fields:

Include structured comments:

✖ Primers

Include Primers

Forward PCR Primer Name: Reverse PCR Primer Name:

Forward PCR Primer Sequence: Reverse PCR Primer Sequence:

✔ Primers

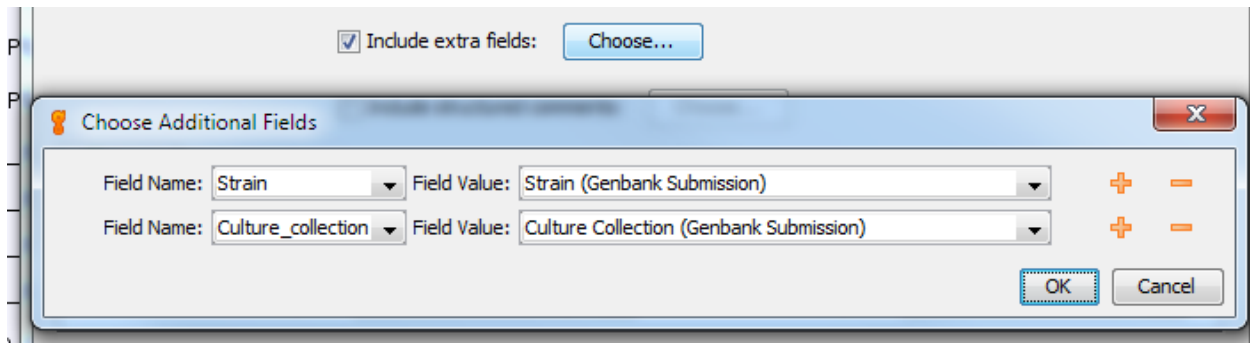
Include Primers

Forward PCR Primer Name:

Reverse PCR Primer Name:

Forward PCR Primer Sequence:

Reverse PCR Primer Sequence:



- Check submission in the Preview mode
 - If Errors you need to correct
 - Ignore warning about Organism not found and Collection

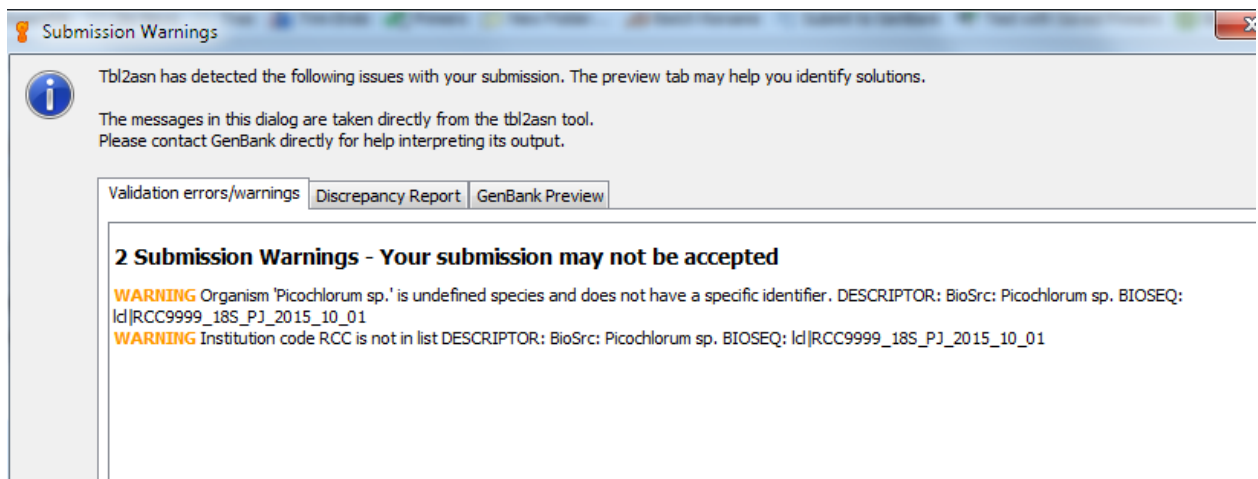


Figure 25: Warnings - Ignore

Submission Warnings

Tbl2asn has detected the following issues with your submission. The preview tab may help you identify solutions.

The messages in this dialog are taken directly from the tbl2asn tool. Please contact GenBank directly for help interpreting its output.

Validation errors/warnings | Discrepancy Report | GenBank Preview

```

LOCUS       RCC9999_18S_PJ_2015_10_011614 bp   DNA       linear     02-OCT-2015
DEFINITION  Picochlorum sp. strain RCC9999.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Picochlorum sp.
  ORGANISM  Picochlorum sp.
            Unclassified.
REFERENCE   1 (bases 1 to 1614)
  AUTHORS   Gourvil,P. and Vaultot,D.
  TITLE     Roscoff Culture Collection
  JOURNAL   Unpublished
REFERENCE   2 (bases 1 to 1614)
  AUTHORS   Gourvil,P. and Vaultot,D.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-OCT-2015) Station Biologique de Roscoff, Plage G.
            Tessier, Roscoff, France
FEATURES    Location/Qualifiers
     source          1..1614
                    /organism="Picochlorum sp."
                    /mol_type="genomic DNA"
                    /strain="RCC9999"
                    /isolate="RCC9999"
                    /culture_collection="RCC:9999"
                    /PCR_primers="fwd_name: 18S 63F, fwd_seq:
                    acgcttgctcctcaaagatta, rev_name: 18S 1818R, rev_seq:
                    acggaaccttggttacga"
     rRNA            <1..>1614
                    /product="18S ribosomal RNA"
ORIGIN
1  agccatgcat  gtctaagtat  aagttgcttt  atactgtgaa  actgccaatg  gctcattaaa
61  tcagttatag  tttatttgat  ggtacctact  tactcggata  cccgtagtaa  ttctagagyt
121 aatacgtgcg  tacatccoga  cttctggaag  ggacgtattd  attagataaa  aggcgcgaccg
181 ggcttgcccg  actcgcgggtg  actcatgata  acttcmcgaa  tcgcatggmc  tcgcgccggc
241 gatgtttcat  tcaaatttct  gccctatcaa  cttttgatgg  taggatagag  gcctaccatg
301 gtgtaacgg  gtgacggaga  attagggttc  gattccggag  agggagcctg  agaaacggct
361 accacatcca  aggaaggcag  caggcgcgca  aattacccaa  tcctgacaca  gggaggtagr
421 gacaataaat  aacaataccg  ggcctttggt  ctggaatkg  gaatgagtac  aacctaataa
481 ccttaacgag  gatcaattgg  agggcaagtc  tggtgccagc  agccgcggta  attccagctc
541 caatagcgta  tatttaagtt  gctgcagtta  aaaagctcgt  acttgattt  cgggtggggc
601 ctgccggtc  gccgtttcgg  tgtgcactgg  ccgggccac  ctggttgccg  gggacgagct
661 cctgggcttt  attgtccggg  actcggagtc  kgcgaggtta  ctttgagtaa  attagagtgt
721 tcaaagcagg  ccaccgctct  gaatacatta  scatggaata  acacgatasg  actctggcct
781 atctntgttg  gtctgtagga  ccggagtaat  gattaagagg  gacagtcggg  ggcattcgta

```

Figure 26: Genbank record preview

- Save as tar file.
 - The submission has to be done before processing a new one starts because Geneious keep at the memory the info from the last .tar file you saved.

- The tar file can be uncompressed to an .asn file which can be opened with Sequin which can be downloaded from NCBI.
- Finally submit using the Geneious BankIt account and record the BankIt number

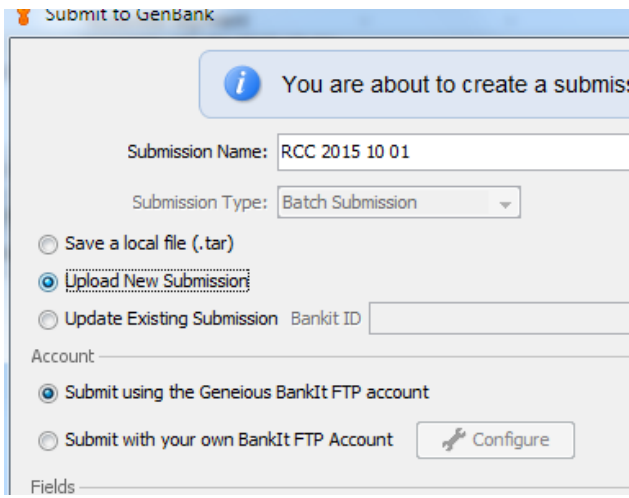


Figure 27: Submit to GenBank

5 GenBank submission - 16S, 18S or ITS

Submission must now be done at <https://submit.ncbi.nlm.nih.gov/subs/genbank/>. If you do not have a login you must create one.

Information about the NCBI submission portal is here. We recommend to read these instructions very carefully before submitting the sequences.

The main steps are :

1. Create fasta file with unique **Name** for each sequence.
 - Sequence **Name** (Sequence_ID) cannot contain spaces. The Sequence_ID identifies the same specimen in all the steps of a submission. We use a convention of the following type **RCC9999_18S_PG_2015_10_01** (see above)
 - Sequence **Name** must be unique within the set and may not contain spaces.
 - Sequence **Name** may contain only the following characters - letters, digits, hyphens (-), underscores (_), periods (.), colons (:), asterisks, and number signs(#).
2. Create a tabulated file as Text (tsv - tab-delimited) containing all the information about the sequence. See this link for the description of all the modifiers. This file can be easily exported from Geneious and finalized with Excel. For the RCC, the following columns are necessary (fields in bold are **mandatory**):
 - **Name** - This field will be used as the **Sequence_ID** for submission
 - **Organism** - *Picochlorum* sp. or *Trebouxiophyceae* sp.

 - **Genbank Submission : Strain** - e.g. RCC1236.
 - **Genbank Submission : Culture_collection** - e.g. RCC:1236.
 - Fwd_primer_name - name of forward PCR primer
 - Fwd_primer_seq - nucleotide sequence of forward PCR primer
 - Rev_primer_name - name of reverse PCR primer
 - Rev_primer_seq - nucleotide sequence of reverse PCR primer
3. The columns of the tabulated file must be edited, not forgetting the underscores. This is best done with an editor such as Notepad++ or with Excel. In the latter case the file must be saved as a text tabulated file.
 - Name -> Sequence_ID
 - Genbank Submission : Strain -> Strain
 - Genbank Submission : Culture_collection -> Culture_collection
 - name of forward PCR primer -> Fwd_primer_name
 - nucleotide sequence of forward PCR primer -> Fwd_primer_seq
 - name of reverse PCR primer -> Rev_primer_name
 - nucleotide sequence of reverse PCR primer -> Rev_primer_seq

Example of header for the tsv file : Sequence_ID Culture Collection Strain Organism
Fwd_primer_name Fwd_primer_seq Rev_primer_name Rev_primer_seq

5.1 Prepare files

We will do a simple case but you can add more columns (see list of modifiers).

- Fasta file
 - Select sequences
 - Export as fasta

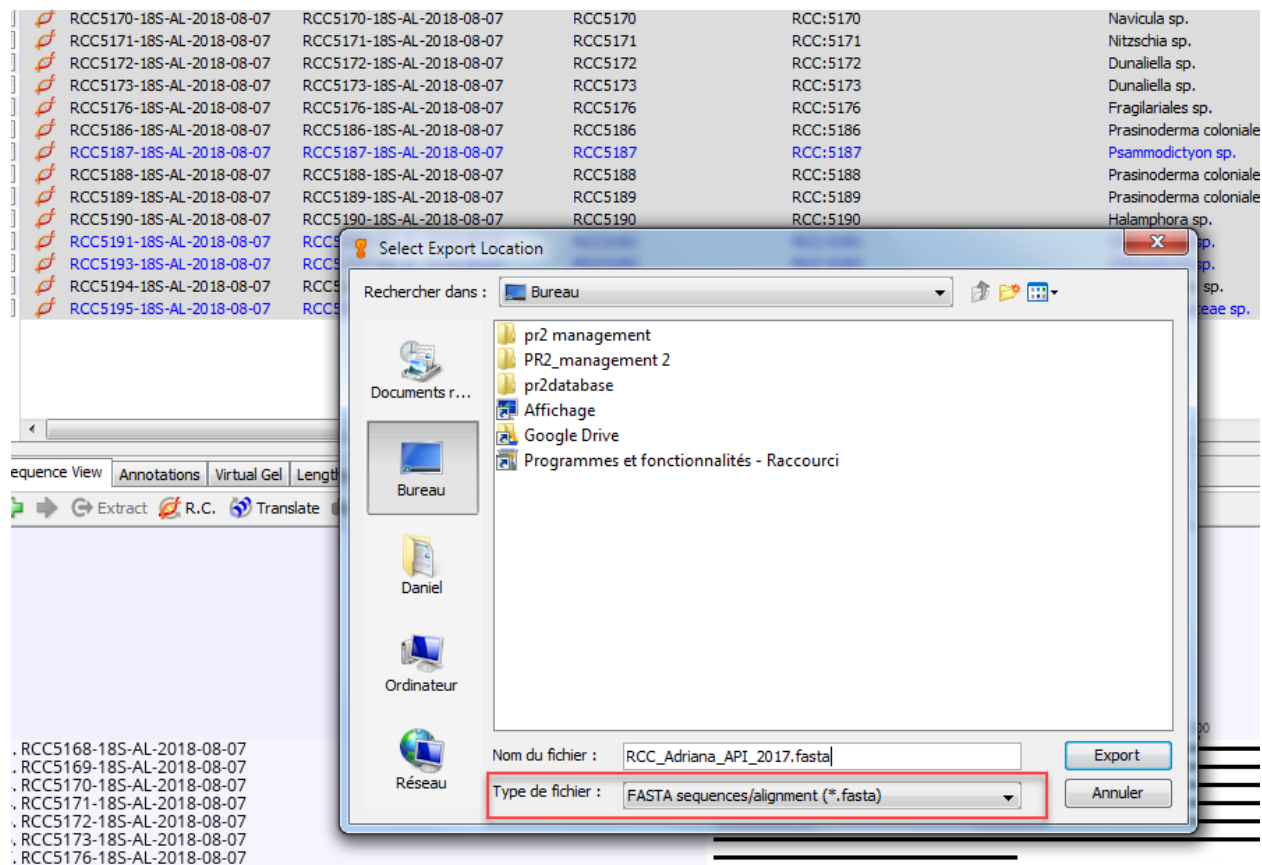


Figure 28: Export to Fasta

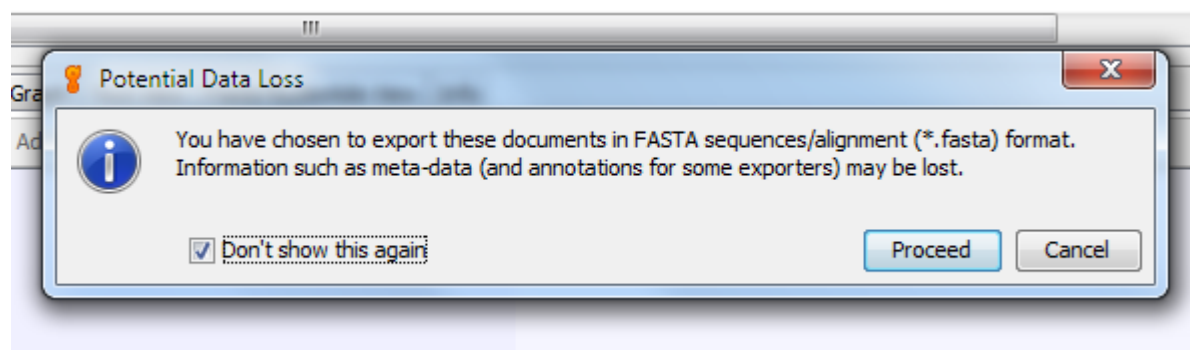


Figure 29: Ignore this warning

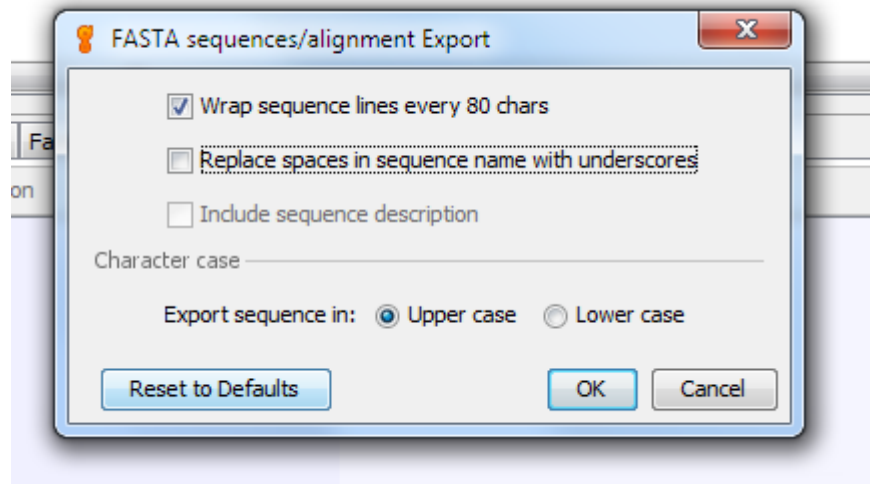


Figure 30: Wrap sequences to 80 characters


```
1 >RCC5168-18S-AL-2018-08-07
2 CTCATTATATCAGTTATAGTTTCTTTGATAGTCCCTCACTACTTGGATAACCGTAGTAATTCTAGAGCTAATACATGCGT
3 CGATACCCCTTTGGGGTAGTATTTATTAGATGAAACCAACCCCTTCGGGGTAGTGTGGTGAATCATAATAAGCTTGCGGA
4 TCGCCGGTGGCGATGGATCATTCAAGTTCTGCCCTATCAGCTTTGGACGTTTGGGTATTGGCCAAACGTGGCTTTAACG
5 GGTAAACGGGAATTAGGGTTCGATTCCGGAGAGGGAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGT
6 AAATTACCCAATCTTGACACAAGGAGGTAGTGACAATAAATAACAATGCCGGCCTTTGTAGGTCTGGCAATTGGAATGA
7 GAACAATTTAAACCCCTTATCGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAG
8 CGTATATTAAGTTGTTGCAGTTTAAAAGCTCGTAGTTGGACTTGTGGTGGTTCCCTGAGGTCCGTGTTGGTACTTTTG
9 GGACTGCCATCCTTGGGTGGATCCTGTGTGGCATTAGGTTGTGTCGAGGGGATGCCCATCGTTTACTGTGAAAAATA
10 GAGTGTTCAAAGCAGGCTTATGCCGTTGAATATAATTAGCATGGAATAATGAGATAGGACCTTGGTACTATTTTGTGGTT
11 TGCGCACCGAGGTAATGATTAATAGGGACAGTTGTGGGTATTCGTATTCCATTGTCAGAGGTGAAATTCCTGGATTTCTG
12 GAAGACGAACGAATGCGAAAACGATTTACCAAGGATGTTTTCAATTAATCAAGAACGAAAAGTTAGGGGATCGAAGATGATTA
13 GATACCATCGTAGTCTTAACCATAAACTATGCCGACAAGGGATTGGCGGGGTTTCGTTACGTCTCCGTGACACCTTATG
14 AGAAATCACAAGTTTTCGGGTTCCGGGGGAGTATGGTTCGCAAGGCTGAAACTTAAAGAAATTGACGGAAGGGCACACC
15 AGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACCTACCAGGTCCAGACATAGTGAAGATTGACAGATTGA
16 GAGCTCTTTCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGTATTGTCGTGTTAATTCGGTTAA
17 CGAACGAGACCCCTGCTGCTAAATAGTCCCTTTGAGTGATTTTCACTGATTGGGGCTTCTTAGAGGGACGTGCATTCTAT
18 TAGATGCAGGAAGATAGGGCAATAACAGGTCGTGTATGCCCTTAGATGTTCTGGGCCGCACGCGCGCTACACTGATGCA
19 TTCAACGAGTTTCCTTGGCCGAGAGGCTGGGCAATCTTTGGAACGTGCATCGTGATAGGGATAGATTATTGCAATTATT
20 AATCTTGAACGAGGAAATCCTAGTAAACGACATCATCAATCTGCATTGATTACGTCCCTGCCTTTGTACACACCGCCC
21 GTCGCACCTACCGATTGGATGGTCCGGTGAAGCCTCGGATGTGACCGGAGCCTTTACGGG
22 >RCC5169-18S-AL-2018-08-07
23 ACTGCGAACGGCTCATTATATCAGTTATAGTTTATTTGATAGTCCCTTACTATTTGGATAACCGTAGTAATTCTAGAGCT
24 AATACATGCGTCAATACCCTTCTGGGGTAGTATTTATTAGATAGAAACCAACCCCTTCGGGGTAGTGTGGTGAATCATAA
25 TAAGCTTGGGATCGCATGGCCTCGGCCGGCGACGGATCATTCAAGTTTCTGCCCTATCAGCTTTGGATGGTAGGGTATT
26 GGCCTACCATGGCTTTAACGGGTAACGGGAAATTAGGGTTTGATTCCGGAGAGGGGGCCTGAGAGATGGCCACCACATCC
27 AAGGAAGGCAGCAGGCGCTAAATTACCCAATCTGACACAGGGAGGTAGTGACAATAAATAACAATGCCGGGCTTTTAT
28 AGGTCTGGCAATTGGAATGAGAACAATTTAAATCCCTTATCGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCG
29 GGTAAATCCAGCTCCAATAGCGTATAATTAAGTTGTTGCAGTTAAAAGCTCGTAGTTGGATTGTGGTTTACGGCGTGT
30 ACCAGGCACTTGTGTCTGAGTTTATGCCGTTGCCATCCTTGGGTGGAACCTGTGTGGCATTAGGTTGTCGTGCAGGG
31 ATGCCCATCGTTACTGTGAAAAAATTAGAGTGTCAAAGCAGGCTTATGCCGTTGAATATAATTAGCATGGAATAATAAG
32 ATAGGACTTTTTCGCTATTTGTGGTGTGGCGAAGAGGTAAATGATTAATAGGGACAGTTGGGGGATTCGTATTCCAT
33 TGTCAGAGGTGAAATTCCTGGATTTTTGGAAGACGAACTACTGCGAAAGCATTACCAAGGATGTTTTCAATTAATCAAGA
34 ACGAAAAGTTAGGGGATCGAAGATGATTAGATACCATCGTAGTCTTAACCATAAACTATGCCGACAAGGGATTGGTGGGGT
35 CTCGTTACGTCTCCATCAGCACCTTATGAGAAATCACAAGTCTTTGGGTTCCGGGGGGAGTATGGTTCGCAAGGCTGAAAC
36 TTAAGAAAATTGACGGAAGGGCACACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACCTACCAGGT
37 CCAGACATAGTGAGGATTGACAGATTGAGAGCTCTTTCTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTG
38 GAGTGATTTGCTGGTTAATTCGTTAACGAACGAGACCCTGCCTGCTAAATAGCCAGTGAGTGAATCTTCACTGACT
39 GCTGGCTTCTTAGAGGGACGTGCATTCTATCAGATGCAGGAGGATAGTGGCAATAACAGGTCTGTGATGCCCTTAGATGT
40 CCTGGGCCGCACGCGCTACACTGATGCATTCAACGAGTTTTACCTTGGCCGAGAGGCTGGGCAATCTTTGAAACGTG
41 CATCGTATAGGGATAGATTATTGCAATTAATAATCTGAAACGAGGAAATCCTAGTAAACGAAATCATCAATTTGCATT
42 GATTACGTCCCTGCCCTTTGTACACACCGCCCGTCACCTACCGATTGAATGGTCCGGTGAAGCCTCGGGATTGTGATC
43 TGTTTCCTTTATTGGAGATG
44 >RCC5170-18S-AL-2018-08-07
45 CGGCTCATTATATCAGTTATAGTTTATTTGATAGTCCCTTACTATTTGGATAACCGTAGTAATTCTAGAGCTAATACATG
46 CGTCAATACCCTTCTGGGGTAGTATTTATTAGATAGAAACCAACCCCTTCGGGGTAGTGTGGTGAATCATAATAAGCTTG
47 CGGATCGCATGGCCTCGGCCGGCGACGGATCATTCAAGTTTCTGCCCTATCAGCTTTGGATGGTGGGTATTGGCCTACC
48 ATGGCTTTAACGGGTAACGGGAAATTAGGGTTTGATTCCGGAGAGGGGGCCTGAGAGATGGCCACCACATCCAAGGAAGG
49 CAGCAGGCGCGTAAATTACCCAATCCTGACACAGGGAGGTAGTGACAATAAATAACAATGCCGGGCTTTTATAGTCTGG
50 CAATTGGAATGAGAACAAATTAATCCCTTATCGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTC
51 CAGTCCAATAGCGTATAATAAGTTGTTGCAGTTAAAAGCTCGTAGTTGGATTGTGGTTCAGGCGGTACCAGGCA
```

Figure 31: Final fasta file

- Source information file tab-delimited (tsv file)

Name ▲	Sequence_ID (Genbank Submission)	Strain (Genbank Submission)	Culture_collection (Genbank Submission)	Organism
RCC5168-18S-AL-2018-08-07	RCC5168-18S-AL-2018-08-07	RCC5168	RCC:5168	Halamphora sp.
RCC5169-18S-AL-2018-08-07	RCC5169-18S-AL-2018-08-07	RCC5169	RCC:5169	Navicula sp.
RCC5170-18S-AL-2018-08-07	RCC5170-18S-AL-2018-08-07	RCC5170	RCC:5170	Navicula sp.
RCC5171-18S-AL-2018-08-07	RCC5171-18S-AL-2018-08-07	RCC5171	RCC:5171	Nitzschia sp.
RCC5172-18S-AL-2018-08-07	RCC5172-18S-AL-2018-08-07	RCC5172	RCC:5172	Dunaliella sp.
RCC5173-18S-AL-2018-08-07	RCC5173-18S-AL-2018-08-07	RCC5173	RCC:5173	Dunaliella sp.
RCC5176-18S-AL-2018-08-07	RCC5176-18S-AL-2018-08-07	RCC5176	RCC:5176	Fragilariiales sp.
RCC5186-18S-AL-2018-08-07	RCC5186-18S-AL-2018-08-07	RCC5186	RCC:5186	Prasinoderma coloniale
RCC5187-18S-AL-2018-08-07	RCC5187-18S-AL-2018-08-07	RCC5187	RCC:5187	Psammodictyon sp.
RCC5188-18S-AL-2018-08-07	RCC5188-18S-AL-2018-08-07	RCC5188	RCC:5188	Prasinoderma coloniale
RCC5189-18S-AL-2018-08-07	RCC5189-18S-AL-2018-08-07	RCC5189	RCC:5189	Prasinoderma coloniale
RCC5190-18S-AL-2018-08-07	RCC5190-18S-AL-2018-08-07	RCC5190	RCC:5190	Halamphora sp.
RCC5191-18S-AL-2018-08-07	RCC5191-18S-AL-2018-08-07	RCC5191	RCC:5191	Chlorophyta sp.
RCC5193-18S-AL-2018-08-07	RCC5193-18S-AL-2018-08-07	RCC5193	RCC:5193	Chlorophyta sp.
RCC5194-18S-AL-2018-08-07	RCC5194-18S-AL-2018-08-07	RCC5194	RCC:5194	Prorocentrum sp.
RCC5195-18S-AL-2018-08-07	RCC5195-18S-AL-2018-08-07	RCC5195	RCC:5195	Heterocapsaceae sp.

Figure 32: Select sequences. Check the 4 fields (Sequence_ID, Strain, Culture_collection and Organism) are correct.

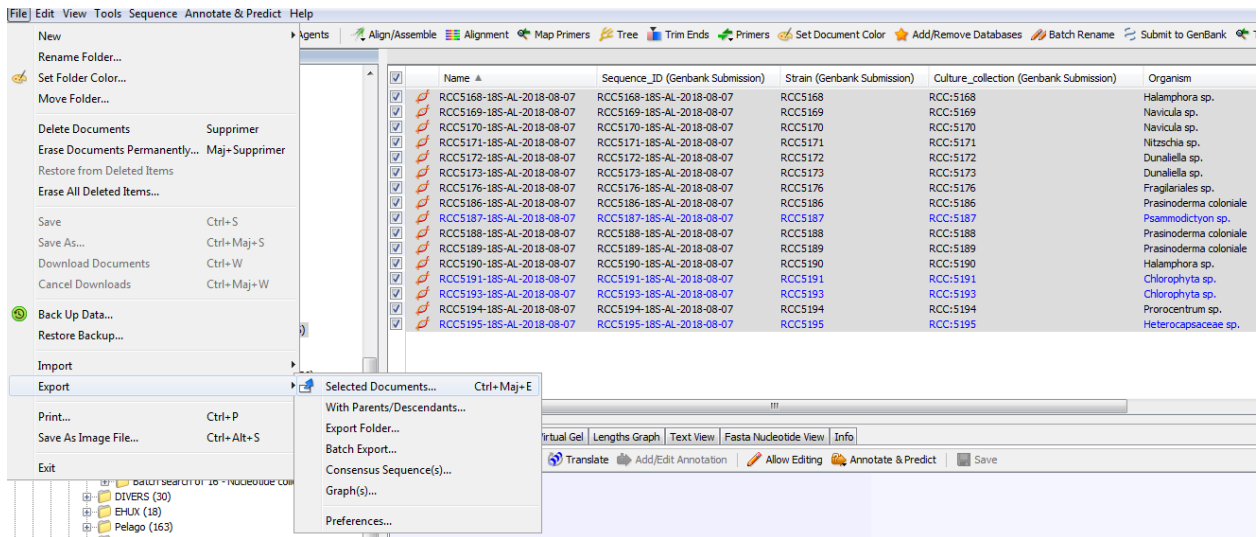


Figure 33: Export selected documents

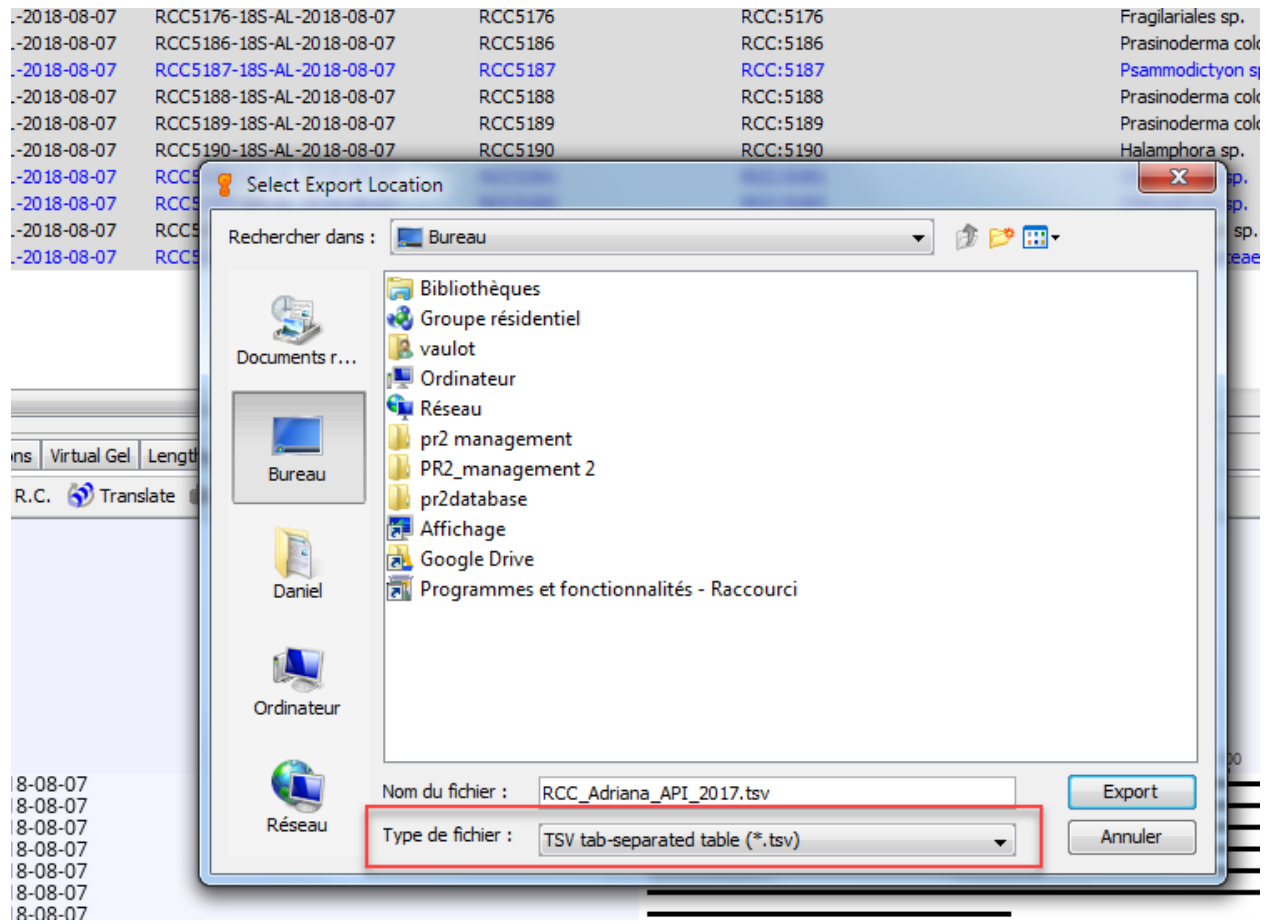


Figure 34: Export as tsv

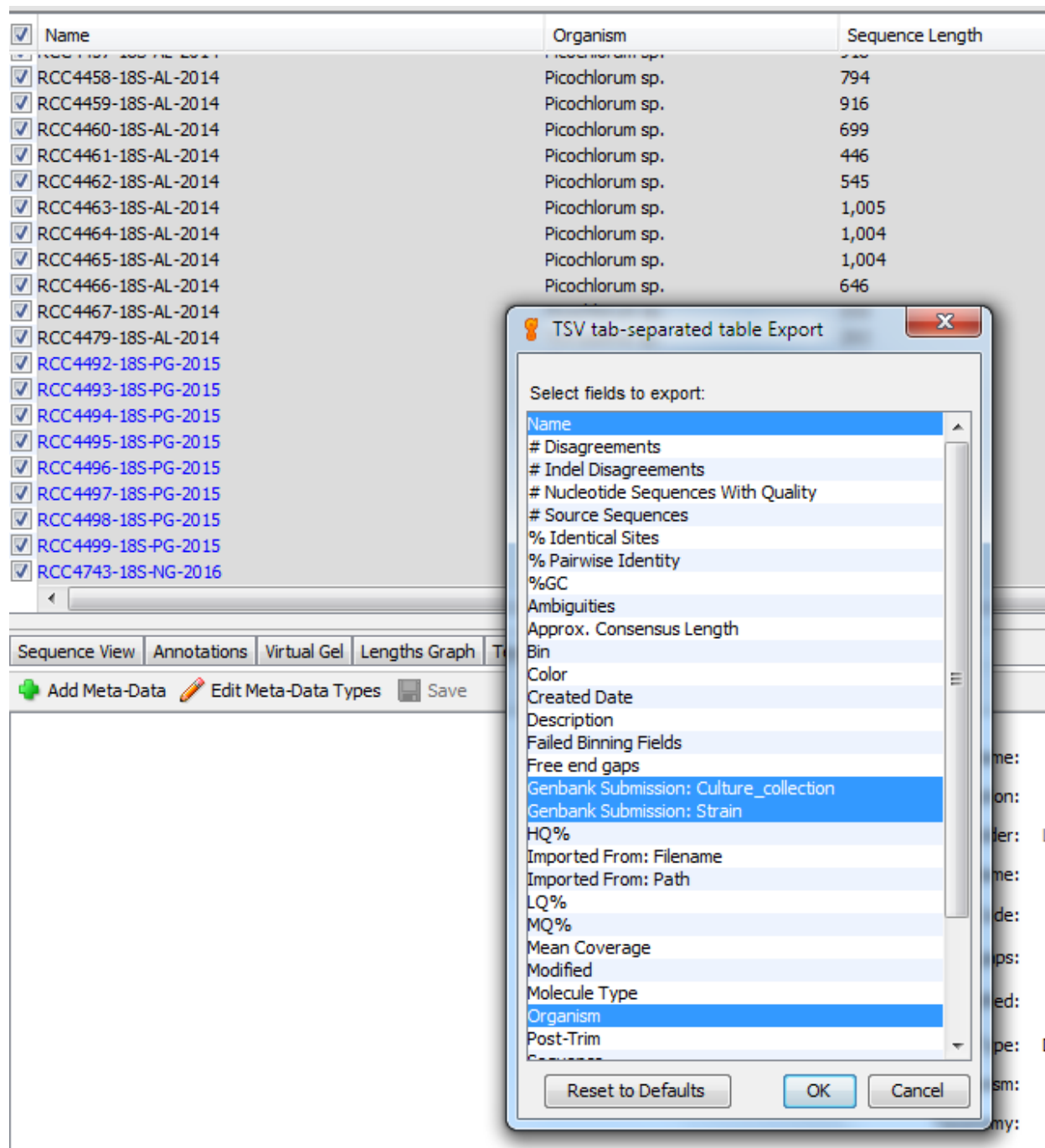


Figure 35: Select the columns to be exported ; Name, Culture_collection, Strain, Organism

Name	Genbank Submission:	Culture collection	Genbank Submission:	Strain	Organism	CRLE
RCC521-18S-AL-2014	→RCC:521	RCC521	→	Pycnococcus	.sp.	CRLE
RCC853-18S-JD-2013	→RCC:853	RCC853	→	Stramenopile	.sp.	RCC853 CRLE
RCC862-18S-JD-2013	→RCC:862	RCC862	→	Stramenopile	.sp.	RCC853 CRLE
RCC3632-18S-IP-2015	→RCC:3632	→RCC3632	Chloropicon	roscoffensis	CRLE	
RCC3633-18S-IP-2015	→RCC:3633	→RCC3633	Chloropicon	roscoffensis	CRLE	
RCC4207_SC2_528F-18S-RE-2014	→RCC:4207	→RCC4207	Isochrysis	.sp.	CRLE	
RCC4208_SC3_528F-18S-RE-2014	→RCC:4208	→RCC4208	Chaetoceros	.sp.	CRLE	
RCC4209_SC6_528F-18S-RE-2014	→RCC:4209	→RCC4209	Pyramimonas	.sp.	CRLE	
RCC4210_SC11_528F-18S-RE-2014	→RCC:4210	→RCC4210	Chrysochromulina	.sp.	CRLE	
RCC4212_SC16_528F-18S-RE-2014	→RCC:4212	→RCC4212	Pyramimonas	.sp.	CRLE	
RCC4213_SC18_528F-18S-RE-2014	→RCC:4213	→RCC4213	Minidiscus	.sp.	CRLE	
RCC4214_SC27_528F-18S-RE-2014	→RCC:4214	→RCC4214	Dicrateria	.sp.	CRLE	
RCC4216_30SC_1818R_RC-18S-RE-2014	→RCC:4216	→RCC4216	Hemiselmia	.sp.	CRLE	
RCC4217_31SC_1818R_RC-18S-RE-2014	→RCC:4217	→RCC4217	Pyramimonas	.sp.	CRLE	
RCC4218_PG-2014	→RCC:4218	→RCC4218	Pyramimonas	.sp.	CRLE	
RCC4219_PG-2014	→RCC:4219	→RCC4219	Thalassiosira	.sp.	CRLE	
RCC4220_PG-2014	→RCC:4220	→RCC4220	Thalassiosira	.sp.	CRLE	
RCC4444-18S-AL-2014	→RCC:4444	→RCC4444	Rhodomonas	.sp.	CRLE	
RCC4445-18S-AL-2014	→RCC:4445	→RCC4445	Picochlorum	.sp.	CRLE	
RCC4447-18S-AL-2014	→RCC:4447	→RCC4447	Picochlorum	.sp.	CRLE	

Figure 36: Edit the tsv file to remove ‘GenBank Submission:’ in the titles of the columns and change ‘Name’ to ‘Sequence_ID’. This is best done with an editor such as [Notepad++](https://notepad-plus-plus.org/fr/) or with Excel. In the latter case the file must be saved as a text tabulated file.

Sequence ID	Culture Collection	Strain	Organism	CRLE
RCC521-18S-AL-2014	→RCC:521	RCC521	→	Pycnococcus .sp. CRLE
RCC853-18S-JD-2013	→RCC:853	RCC853	→	Stramenopile .sp. RCC853 CRLE
RCC862-18S-JD-2013	→RCC:862	RCC862	→	Stramenopile .sp. RCC853 CRLE
RCC3632-18S-IP-2015	→RCC:3632	→RCC3632	Chloropicon	roscoffensis CRLE
RCC3633-18S-IP-2015	→RCC:3633	→RCC3633	Chloropicon	roscoffensis CRLE
RCC4207_SC2_528F-18S-RE-2014	→RCC:4207	→RCC4207	Isochrysis	.sp. CRLE
RCC4208_SC3_528F-18S-RE-2014	→RCC:4208	→RCC4208	Chaetoceros	.sp. CRLE
RCC4209_SC6_528F-18S-RE-2014	→RCC:4209	→RCC4209	Pyramimonas	.sp. CRLE
RCC4210_SC11_528F-18S-RE-2014	→RCC:4210	→RCC4210	Chrysochromulina	.sp. CRLE
RCC4212_SC16_528F-18S-RE-2014	→RCC:4212	→RCC4212	Pyramimonas	.sp. CRLE
RCC4213_SC18_528F-18S-RE-2014	→RCC:4213	→RCC4213	Minidiscus	.sp. CRLE
RCC4214_SC27_528F-18S-RE-2014	→RCC:4214	→RCC4214	Dicrateria	.sp. CRLE
RCC4216_30SC_1818R_RC-18S-RE-2014	→RCC:4216	→RCC4216	Hemiselmia	.sp. CRLE
RCC4217_31SC_1818R_RC-18S-RE-2014	→RCC:4217	→RCC4217	Pyramimonas	.sp. CRLE
RCC4218_PG-2014	→RCC:4218	→RCC4218	Pyramimonas	.sp. CRLE
RCC4219_PG-2014	→RCC:4219	→RCC4219	Thalassiosira	.sp. CRLE
RCC4220_PG-2014	→RCC:4220	→RCC4220	Thalassiosira	.sp. CRLE
RCC4444-18S-AL-2014	→RCC:4444	→RCC4444	Rhodomonas	.sp. CRLE
RCC4445-18S-AL-2014	→RCC:4445	→RCC4445	Picochlorum	.sp. CRLE
RCC4447-18S-AL-2014	→RCC:4447	→RCC4447	Picochlorum	.sp. CRLE
RCC4448-18S-AL-2014	→RCC:4448	→RCC4448	Picochlorum	.sp. CRLE

Figure 37: After editing and removing GenBank Submission:

The image shows the Microsoft Excel interface with the following data in the spreadsheet:

	A	B	C	D	E
1	Culture_collection	Sequence_ID	Strain	Organism	
2	RCC:5168	RCC5168-18S-AL-2018-08-07	RCC5168	Halamphora sp.	
3	RCC:5169	RCC5169-18S-AL-2018-08-07	RCC5169	Navicula sp.	
4	RCC:5170	RCC5170-18S-AL-2018-08-07	RCC5170	Navicula sp.	
5	RCC:5171	RCC5171-18S-AL-2018-08-07	RCC5171	Nitzschia sp.	
6	RCC:5172	RCC5172-18S-AL-2018-08-07	RCC5172	Dunaliella sp.	
7	RCC:5173	RCC5173-18S-AL-2018-08-07	RCC5173	Dunaliella sp.	
8	RCC:5176	RCC5176-18S-AL-2018-08-07	RCC5176	Fragilariales sp.	
9	RCC:5186	RCC5186-18S-AL-2018-08-07	RCC5186	Prasinoderma coloniale	
10	RCC:5187	RCC5187-18S-AL-2018-08-07	RCC5187	Psammodictyon sp.	
11	RCC:5188	RCC5188-18S-AL-2018-08-07	RCC5188	Prasinoderma coloniale	
12	RCC:5189	RCC5189-18S-AL-2018-08-07	RCC5189	Prasinoderma coloniale	
13	RCC:5190	RCC5190-18S-AL-2018-08-07	RCC5190	Halamphora sp.	
14	RCC:5191	RCC5191-18S-AL-2018-08-07	RCC5191	Chlorophyta sp.	
15	RCC:5193	RCC5193-18S-AL-2018-08-07	RCC5193	Chlorophyta sp.	
16	RCC:5194	RCC5194-18S-AL-2018-08-07	RCC5194	Prorocentrum sp.	
17	RCC:5195	RCC5195-18S-AL-2018-08-07	RCC5195	Heterocapsaceae sp.	
18					
19					

Figure 38: Editing with Excel (save file as tab-delimited tsv)

5.2 Submit to NCBI web portal

- Go to web portal : <https://submit.ncbi.nlm.nih.gov/subs/genbank/>

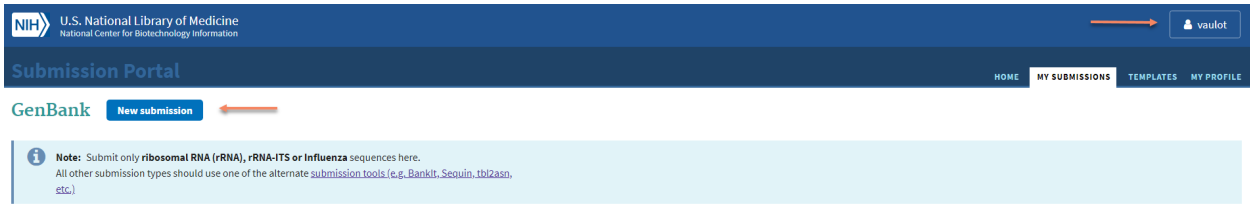


Figure 39: Web portal. Register or login if you have already an ID

GenBank submission: SUB4388090

New



Submission Type

Submission type

★ **What do your sequences contain?**

- Prokaryotic rRNA/IGS
- Eukaryotic Nuclear rRNA/ITS
- Eukaryotic Organelle rRNA
- Influenza virus

★ **What do these Prokaryotic rRNA/IGS sequences contain?**

- small subunit rRNA only (16S rRNA)
- large subunit rRNA only (23S rRNA)
- intergenic spacer (16S-23S rRNA IGS)

i If none of the options above describe your sequence, you can provide a title for the Submission Portal. This title records. Limit your title to 60 characters.

Submission title (Optional, not displayed in final records) ?

RCC Adriana 2018 Prokaryotes

Continue

Figure 40: Enter the type of sequence

Submitter

Affiliation

i The information you give here will be displayed in the final sequence records.
For address details, provide the primary address where work was done to generate the data in this submission.

★ **Submitting organization**
 ★ **Department**

★ **Street** ★ **City** ★ **State/Province**
 ★ **Postal code** ★ **Country**

Contact information

i GenBank may use this information to contact you about your submission, it will not be displayed in the final sequence records.

★ **E-mail (primary)** ★ **E-mail (secondary)** **i** Please provide an alternate email address to ensure that messages are received

★ **First (given) name** ★ **Middle name**
 ★ **Last (family) name**

Phone **Fax**

Continue Update my contact information in profile

Figure 41: Enter submitter information

Submission Portal

GenBank submission: SUB4388090

Prokaryotic 16S rRNA / RCC Adriana 2018 Prokaryotes

1 SUBMISSION TYPE

2 SUBMITTER

3 SEQUENCING TECHNOLOGY

4 SEQUENCES

5 SEQUENCE PROCE

Sequencing Technology

Method

★ What methods were used to obtain these sequences? ?

Sanger dideoxy sequencing



454

Helicos

Illumina

IonTorrent

Pacific Biosciences

SOLiD

Other

Assembly State

These sequences are:

Unassembled sequence reads

Assembled sequences (each sequence was assembled from two or more overlapping sequence reads)

Continue

Figure 42: Enter the sequence technology. In almost all cases choose Sanger and Assembly

Sequences

Release date

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

★ When should this submission be released to the public?

Release immediately following processing ←

Release on specified date or upon publication, whichever is first

Chimera check

★ Did you check and remove low-quality and chimeric sequences from your FASTA file prior to preparing this submission?

Yes ← **Only for prokaryotes**

No

Note: Please provide the name and version of the chimera checking program. BLAST alone is not sufficient as a chimera checking program.

Program Name	Version
Geneious	10

Cultured or Uncultured

Select whether your sequences were obtained from cultured or uncultured samples.

★ Bacterial/archaeal Sequences: How were they obtained?

Pure-cultured strains (axenic cultures containing only one microbial species each) ← **Only for prokaryotes**

Uncultured, bulk environmental DNA (PCR-amplified directly from environmental sample or host; samples were not grown in culture)

Sequences

★ Upload a nucleotide FASTA formatted file.

Choisir un fichier | Aucun fichier choisi ←

Note: If you have multiple sequences, all of your sequences need to be in one file. [Help on FASTA file.](#)

Example FASTA nucleotide format:

```
>Seq1
aaccgatagagatagtgatccgatagagagga
>Seq2
gtacgataaagagatagtgatccgatagagagga
```

Note: Use the latest version of the [Aspera Connect plugin](#) for faster file uploads. If a pop-up box about 'fasp protocol' is displayed, click 'Allow' or 'Open' to let [Aspera Connect](#) handle file uploads more efficiently.

Figure 43: Sequences. Release date: Choose immediate release in most cases, there is really no need to delay release.- The chimera question is only for Prokaryotes. - Chose pure cultures for cyanos. - Upload the fasta sequence file

GenBank submission: SUB4388090

cultured Prokaryotic 16S rRNA / RCC Adriana 2018 Prokaryotes

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SOURCE INFO 6 SOURCE MODIFIERS 7 REFERENCES 8 OVERVIEW

Source Information

i The first few sequence IDs that we found are:

RCC5180-185-AL-2018-08-07
RCC5183-185-AL-2018-08-07
RCC5185-185-AL-2018-08-07

***** Do your sequence IDs represent one of these? **i** Values for these are typically alpha-numeric sample codes used in your laboratory to track individual samples. Select 'NONE of these' if it does not describe your sequence IDs or the sequence IDs contain more information than the [description/scope](#) of that field.

Strain

NONE of these ←

Continue

Figure 44: Source information. Since it will be loaded in the text file, choose - NONE of these

GenBank submission: SUB4388090

cultured Prokaryotic 16S rRNA / RCC Adriana 2018 Prokaryotes

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SOURCE INFO 6 SOURCE MODIFIERS 7 REFERENCES 8 OVERVIEW

Source Modifiers

▼ For each sequence, GenBank requires the following source information:

- scientific name of the **Organism** that was sequenced and
- strain** (NOT from a top BLAST hit, NOT the species name).

If you have already provided all the required information, you can press Continue to proceed.

i More help: [how to provide source modifiers](#), [description of each modifier](#), [what is a source modifier?](#)

★ How do you want to apply source modifiers?

Use an editable table

Upload a tab-delimited file (use our provided template) ←

Continue

Figure 45: Upload tsv file saved from Geneious

GenBank submission: SUB4388090

cultured Prokaryotic 16S rRNA / RCC Adriana 2018 Prokaryotes

- 1 SUBMISSION TYPE
- 2 SUBMITTER
- 3 SEQUENCING TECHNOLOGY
- 4 SEQUENCES
- 5 SOURCE INFO
- 6 SOURCE MODIFIERS
- 7 REFERENCES
- 8 OVERVIEW

Source Modifiers

For each sequence, GenBank requires the following source information:

- scientific name of the **Organism** that was sequenced and
- strain** (NOT from a top BLAST hit, NOT the species name).

If you have already provided all the required information, you can press Continue to proceed.

More help: [how to provide source modifiers](#), [description of each modifier](#), [what is a source modifier?](#)

Current Source Modifiers - what you have provided so far

* How do you want to apply source modifiers?

- Download [source modifier template](#) with any source information provided so far.
 - Edit the downloaded table in Microsoft Excel or any other editor.
 - [See an example Source Modifiers table](#)
 - Save the table as a tab-delimited text file.
 - Upload Source Modifiers file.
RCC_Adriana_API_2017_Prok.tsv ✓
- Choisir un fichier | Aucun fichier choisi
- Click Continue to validate the information and follow the instructions.

Continue


Figure 46: After uploading tsv file

GenBank submission: SUB4388231

Eukaryotic Nuclear rRNA/ITS / RCC Adriana 2018 Eukaryotes

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SOURCE INFO 6 SOURCE MODIFIERS 7 REFERENCES 8 OVERVIEW

Source Modifiers

 **Warning:** One or more of the organism names listed below are not in the NCBI Taxonomy database. Confirm that the spelling of the listed organism name(s) is correct. If any are not correct, provide corrected names. If they are correct, click the Continue button.

Organism Name

Trebouxiophyceae sp.

Chlorophyceae sp.

Fragilariales sp.

Chlorophyta sp.

Heterocapsaceae sp.

Figure 47: Taxonomy error - This error is due to the addition of sp. to taxa at the rank above the genus. You need to correct and remove sp. to the tsv file. If the error comes from a new taxon not yet described you can ignore and GenBank will contact you probably to add this taxon to their database.

GenBank submission: SUB4388090

cultured Prokaryotic 16S rRNA / RCC Adriana 2018 Prokaryotes

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SOURCE INFO 6 SOURCE MODIFIERS 7 REFERENCES 8 OVERVIEW

References

Sequence authors

Who should be publicly credited as the submitter of this sequence data?

* First (given) name	MI	* Last (family) name	Delete
Adriana		Lopes dos Santo	<input type="checkbox"/>
Daniel		Vaulot	<input type="checkbox"/>

Names will appear in your records as:
Lopes dos Santos, A. and Vaulot, D.

Reference

Please provide the title and relevant publication details of **your paper** that discusses **this submission**.

* Publication status
 Unpublished In-press Published

Reference title
Roscoff Culture Collection

Select Reference Authors
 Same as sequence authors Specify new authors

Figure 48: Add the reference. For the Roscoff Culture Collection just fill as indicated with the name of the person who produced the sequence first.

Overview

Submit

You have requested that your sequence data be released **immediately following processing**.

Submitter

Submitter: Daniel Vaulot
vaulot@roscoff.fr
vaulot@gmail.com
Institution: Sorbonne Universite, CNRS, Station Biologique
Department: UMR7144
Street: Place Georges Teissier
City: Roscoff
Postal code: 29680
Country: France

Sequence authors

- Adriana Lopes dos Santos
- Daniel Vaulot

References

Publication status: unpublished
Reference title: Roscoff Culture Collection
Authors: same as sequence authors

Sequencing Technology

Methods: - Sanger dideoxy sequencing
Assembly state: assembled
Chimera tool used?: Geneious 10

Uploaded files

To proceed please review your submission, make changes if necessary by clicking on the tabs/steps above, then click the Submit button below.

Figure 49: Final check

Submission Portal

HOME MY SUBMISSIONS TEMPLATES MY PROFILE

GenBank [New submission](#)

Note: Submit only ribosomal RNA (rRNA), rRNA-ITS or Influenza sequences here. All other submission types should use one of the alternate submission tools (e.g. BankIt, Sequin, tbl2asn, etc.)

Attention: If you have corrections to an existing submission with status:

- Processed-error: use the FIX button to correct a submission.
- Queued or Processing: email your request with the Submission ID.
- Processed: follow these directions and email your request with the Accession numbers.

Do not create a new submission to fix or update an existing submission whose status is Queued, Processed-error, Processing, or Processed!

Filter / Search

From date: [] To date: [] Status: Not deleted [v] Sort by: [] desc []

Data archives: Show

Query: [] Search Clear

Short description and brief instructions

Submission	Title	Group	Status	Updated
SUB4388090	cultured Prokaryotic 16S rRNA / RCC Adriana 2018 Prokaryotes		Submitted Awaiting processing	01:51

Figure 50: Submission status. When you press submit you should arrive at the final screen showing your submission.

GenBank MH732916-MH732918



De gb-admin@ncbi.nlm.nih.gov
à vaulot@gmail.com, vaulot@sb-roscoff.fr

Dear GenBank Submitter:

Thank you for your direct submission of sequence data to GenBank. We have provided GenBank accession number(s) for your nucleotide sequence(s):

SUB4388090	RCC5180-18S-AL-2018-08-07	MH732916
SUB4388090	RCC5183-18S-AL-2018-08-07	MH732917
SUB4388090	RCC5185-18S-AL-2018-08-07	MH732918

GenBank accession numbers should appear in any publication that reports or discusses the data, as it gives the community a unique label with which they may retrieve your data from our online servers.

Based on the data submitted to us, the scheduled release date for your submission is:

Aug 15, 2018

Figure 51: A few time latter (from a few minutes to a few days), you should receive an email with the accession numbers. Please forward to rcc@sb-roscoff.fr.

- In case of errors in the submission, you may receive an email as follows.

GenBank Submissions grp 6714438 (SUB4396743)



De gb-admin@ncbi.nlm.nih.gov
à vaulot@gmail.com, vaulot@sb-roscoff.fr

 FixSubmission_FAQ.txt (2 kB)

From: gb-admin@ncbi.nlm.nih.gov
To: vaulot@gmail.com, vaulot@sb-roscoff.fr
Subject: GenBank Submissions grp 6714438 (SUB4396743)
Date: Mon, 20 Aug 2018 12:04:33 -0400 (Eastern Daylight Time)

-----=_NextPart_A7E0E8E4=_247E1497

Dear Dr. Vaulot:

We have received the following 48 sequence resubmissions from you:

SUB4396743 : (48)

However, you have not resolved the misassembled sequence. We suggest that you trim/remove nucleotides 1-118 from RCC4219_PG-2014.

BLAST similarity search analysis indicates the following sequences may be misassembled:

RCC4219_PG-2014

This could be due to:

- a mixture of plus and minus strand sequence
- incorrect order of the assembled fragments
- or
- missing sequence indicated by gaps in BLAST results

Please correct the sequences and send a revised submission file through the Submission Portal:

1. Go to submission portal:
<https://submit.ncbi.nlm.nih.gov/subs/?search=SUB4396743>
(login required)
2. Use the 'Fix' button to enable editing of your submission and to correct the errors listed above
3. After you have corrected the information, click 'Submit' at the end of the submission forms.

See attached FixSubmission_FAQ.txt for more information about editing your submission.

Figure 52: This email explain which sequence(s) is(are) incorrect.

- You will need to carefully examine the sequence that Genbank has identified as problematic. Do not disregard their analysis because they really seem to catch errors. In most cases this is due to **bad** assembly. You need to go back to your trace files (see above) and use a higher threshold to trim the sequences (e.g. 0.02 instead of 0.05) and redo the assembly.
- Once this is done, remove the old sequence which was bad and add the new sequence (using the same sequence name) to your fasta file

- Update the .tsv file if necessary.
- Follow the link in the email to go back to the NCBI web site. You will need to reload **ALL** sequences in the submission that had problem as well as the tsv file with all the information

U.S. National Library of Medicine
National Center for Biotechnology Information

Submission Portal Home **My submissions** Templates My profile

Your submissions 1

Start a new submission:

- BioProject
- BioSample
- GenBank
- Sequence Read Archive
- TSA
- Genome
- GTR
- Variation
- ClinVar
- Supplementary Files
- API

Filter / Search

From date: To date: Status: Not deleted Sort by: desc

Apps [Show](#) Data archives [Show](#)

Query

1 submission

Submission	Title	App	Group	Status	Updated	
SUB4396743	Eukaryotic Nuclear rRNA/ITS / RCC Misc 2018 08 12 SSU	GenBank		GenBank: Error see email	<input type="button" value="Fix"/>	Aug 22

Figure 53: The web page to correct your submission. When you click on the link you will be asked to reload the fasta and tsv files.

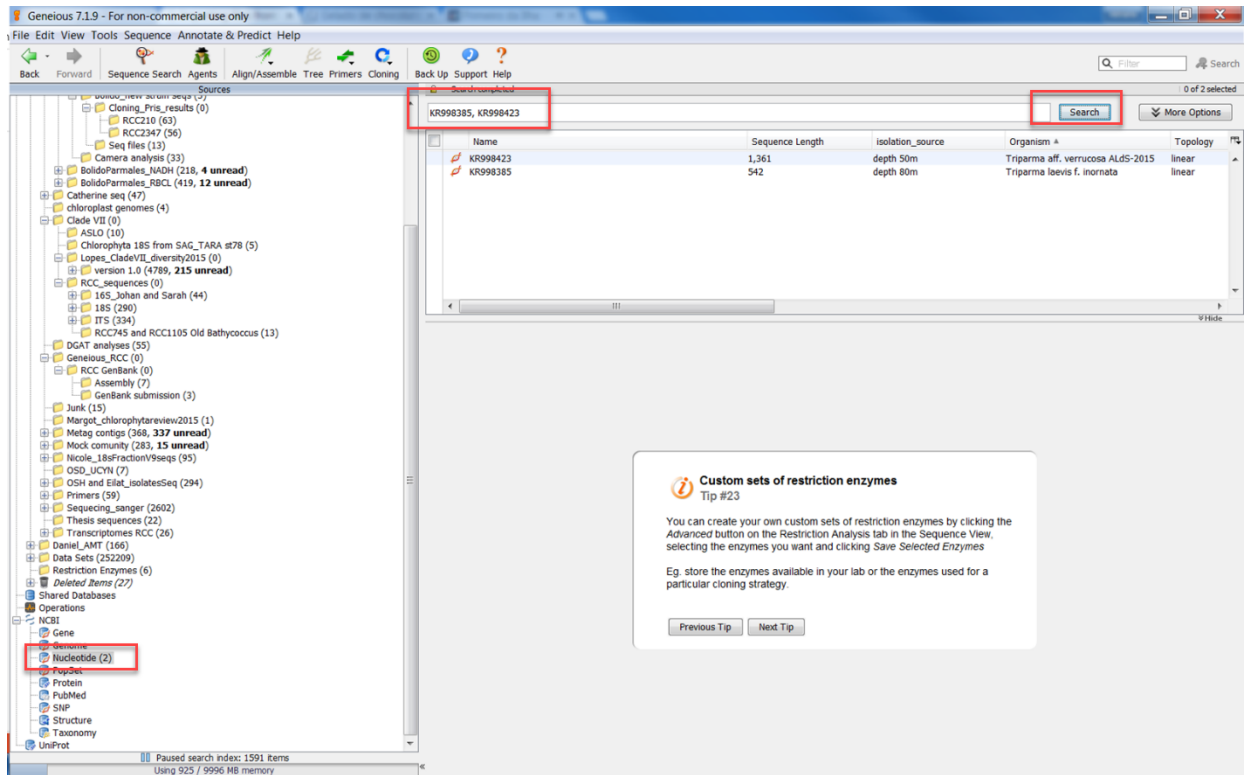
Submission	Title	Group	Status
SUB4396743	Eukaryotic Nuclear rRNA/ITS / RCC Misc 2018 08 12 SSU		Submitted Awaiting processing.
SUB4396254	Eukaryotic Nuclear rRNA/ITS / RCC Pris ITS 2018 08		GenBank: Processed MH781400-MH781403 3 files: • AccessionReport.tsv • flatfile • email.txt
SUB4388372	Eukaryotic Nuclear rRNA/ITS / RCC Catherine GE 2018 08		GenBank: Processed MH764681-MH765044 3 files: • AccessionReport.tsv • flatfile • email.txt
SUB4396248	Eukaryotic Nuclear rRNA/ITS / RCC Prs SSU 2018 08		GenBank: Processed MH764612-MH764680 3 files: • AccessionReport.tsv • flatfile • email.txt
SUB4388231	Eukaryotic Nuclear rRNA/ITS / RCC Adriana 2018 Eukaryotes		GenBank: Processed MH743117-MH743134 3 files: • AccessionReport.tsv • flatfile • email.txt
SUB4396253	Eukaryotic Nuclear rRNA/ITS / RCC Prs LSU 2018 08		GenBank: Processed MH734606 3 files: • AccessionReport.tsv • flatfile • email.txt
SUB4388090	cultured Prokaryotic 16S rRNA / RCC Adriana 2018 Prokaryotes		GenBank: Processed MH732916-MH732918 3 files: • AccessionReport.tsv • flatfile • email.txt

Figure 54: After doing the correction you should see both the validated submissions as well as the one you just fixed.

6 Appendixes

6.1 Retrieve sequences from Genbank using Geneious

- **For a list:** Go to nucleotide, type the numbers separate by coma and click search. The results will appear in the bottom panel. You must drag the file into a folder in your local database if you wish to retain the file and/or modify it.



*

For consecutive accessions numbers: type the first and last numbers separated by ;, click in more options, change All fields to Accession

The screenshot shows the Geneious 7.1.9 software interface. On the left is a 'Sources' tree with various folders like 'Cloning_Pris_results', 'BolidoParmales_NADH', etc. The 'Nucleotide (39)' folder is highlighted with a red box. The main window displays a search results table with columns: Name, Sequence Length, Isolation_source, Organism, and Topology. A search filter is applied: 'Accession is KR998385;KR998423'. The 'Search' button is also highlighted with a red box. A 'Custom sets of restriction enzymes' tip is visible at the bottom.

Name	Sequence Length	Isolation_source	Organism	Topology
KR998394	531	depth 20m	Bolidomonas mediterranea	linear
KR998397	1,665	depth 20m	Bolidomonas mediterranea	linear
KR998419	1,234	depth 20m	Bolidomonas mediterranea	linear
KR998389	574	depth 15m	Bolidomonas pacifica	linear
KR998390	575	depth 15m	Bolidomonas pacifica	linear
KR998393	601	depth 5m	Bolidomonas pacifica	linear
KR998398	1,698	depth 15m	Bolidomonas pacifica	linear
KR998399	1,686	depth 5m	Bolidomonas pacifica	linear
KR998409	458	depth 15m	Bolidomonas pacifica	linear
KR998416	1,358	depth 15m	Bolidomonas pacifica	linear
KR998417	1,200	depth 15m	Bolidomonas pacifica	linear
KR998418	1,200	depth 5m	Bolidomonas pacifica	linear
KR998391	513	depth 5m	Bolidomonas pacifica var. eleuthera	linear
KR998392	746	depth 15m	Bolidomonas pacifica var. eleuthera	linear
KR998396	536	depth 5m	Bolidomonas pacifica var. eleuthera	linear
KR998400	1,640	depth 5m	Bolidomonas pacifica var. eleuthera	linear
KR998408	584	depth 5m	Bolidomonas pacifica var. eleuthera	linear
KR998410	462	depth 15m	Bolidomonas pacifica var. eleuthera	linear
KR998412	458	depth 5m	Bolidomonas pacifica var. eleuthera	linear
KR998413	607	depth 25m	Bolidomonas pacifica var. eleuthera	linear
KR998421	1,200	depth 5m	Bolidomonas pacifica var. eleuthera	linear
KR998422	1,015	depth 25m	Bolidomonas pacifica var. eleuthera	linear
KR998395	560	English Channel depth 10m	Bolidomonas sp. ALDS-2015	linear
KR998411	692	English Channel depth 10m	Bolidomonas sp. ALDS-2015	linear
KR998420	1,386	English Channel depth 10m	Bolidomonas sp. ALDS-2015	linear
KR998388	546	depth 50m	Triparma aff. verrucosa ALDS-2015	linear
KR998403	1,739	depth 50m	Triparma aff. verrucosa ALDS-2015	linear
KR998407	691	depth 50m	Triparma aff. verrucosa ALDS-2015	linear
KR998423	1,361	depth 50m	Triparma aff. verrucosa ALDS-2015	linear