



MAR 04, 2024

Creating a Manual Consensus Sequence from FASTQ with UGENE

Stephen Douglas Russell¹¹Mycota Lab / The Hoosier Mushroom Society

Stephen Douglas Russell

ABSTRACT

Current versions of automated nanopore amplicon pipelines sometimes (rarely) produce erroneous consensus sequences based on low quality reads that are being incorporated into the final result. It is sometimes helpful to remove these reads from the consensus manually. Further, fungal amplicons can have multiple haplotypes of the same ribosomal sequence region within a single organism, and it can be helpful to have them outlined or otherwise flagged with ambiguous nucleotides. This protocol can help to assist sequence analysis when these conditions are present.

OPEN  ACCESS

Protocol Citation: Stephen Douglas Russell 2024. Creating a Manual Consensus Sequence from FASTQ with UGENE.

protocols.io

<https://protocols.io/view/creating-a-manual-consensus-sequence-from-fastq-wi-c93dz8i6>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
We use this protocol and it's working

Created: Mar 03, 2024**Last Modified:** Mar 04, 2024**PROTOCOL integer ID:** 96069

Keywords: dna barcoding, dna sequencing, ont, nanopore, sequence analysis

Install Software, Retrieve Data, Import to UGENE

- 1 Download and install the latest version of UGENE at <https://ugene.net/>.
- 2 Retrieve the FASTQ file of your sequence. If you are using MycoMap, first go to your MycoMap sequence accession. The easiest way to find it is from the MycoBLAST results page. The file I used for this protocol can be found here:

 ONT08.93-E12-iNat178420840-1.fastq 675KB

MycoBLAST Results							Export FASTA	
Hit Number	Description			Identity	Query/Subject Cover	Gap Openings	Accession	Source
1	E12-iNat178420840-1 Species Name: Marasmius Location: Park Corner Prince Edward Island CA RIC			100	100%/100%	0	109034	178420840 - Marasmius
429								

A line within the MycoBLAST results for a given sequence. The red arrow points to the link to the MycoMap sequence accession for that record.

Then download the FASTQ file from the accession page.



Move to ONT Trash

Linked Observation Record

Run Blast Search

Follow

0

E12-iNat178420840-1



By Stephen Russell

January 17

Assembly Files: [ONT08_93-E12-iNat178420840-1.fastq](#)

Forward Primer: ITS1F

Reverse Primer: ITS4

iNat Report Association: [178420840 - Marasmius](#)

Reads in Consensus: 429

```

1  AAGTCGTAAC AAGGTTTCCG TAGGTGAACC TCGGGAAGGATCATTATTGA AACATTGTAA
61  AGGGAGGTTG AGCTGGCTCT TCAAGGGCAA GTGCTCGCTT TTCTTTCAAT CTTATCCAC
121 CTGTGCACCT TCTGTAGGGA GTCTTGAGAA CAGGGCCCTT GTGTGTCTTA AGTATTGAGC
181 TTTCTATGTC TTACAAACT CTAAATGTAT GTCTATGAAT GTCTTTATAA GGGGACTTAG
241 TTGACCCCTT TAAAAACTAT ACAACTTTCA GCAACGGATC TCTTGGCTCT CCGATCGATG
301 AAGAACGCAG CGAAATGCGA TAAATAATGT GAATTGCGA ATTCAGTGAA TCATCGAATC
361 TTTGAACGCA CCTTGGCCT CTTGGTATTC CGAGAGGCAT GCCTGTTTGA GTGTCATTAA
421 ATTCTCAACC TCAAAAGCTT TTTTTTTTTT TGTTCCTGAG GCTTGGATGT GGAGGCTTGC
481 CGGCTTCTTC AGAGTCGGCT CCTCTTAAAT GCATGAGTGG AAAGTGTTC TAGTCCGCAT
541 TGGTGTGATA ATTATCAGCG CTATTGTGCG TACAAGCTCT TGTAGTGTTC GTTTGAAAG
601 CTGCATGAAT GTGCTCTTTC TGTTTTACCT GACCTACACT GAGTAATATA GTATCTGCTT
661 CAAACCGTCC TAAGTAAATT GGACAACATT TGATTATTTT GACCTCAAAT CAGGTAGGAC
721 TACCCGCTGA ACTTAAG
  
```

GenBank Submission Date: No value

Run Name: Run32_CMsummer

The MycoMap sequence accession page. The red box highlights the link to download the FASTQ file of the raw reads this consensus sequence was formed from.

3 Open your FASTQ file in UGENE.

Welcome to UGENE



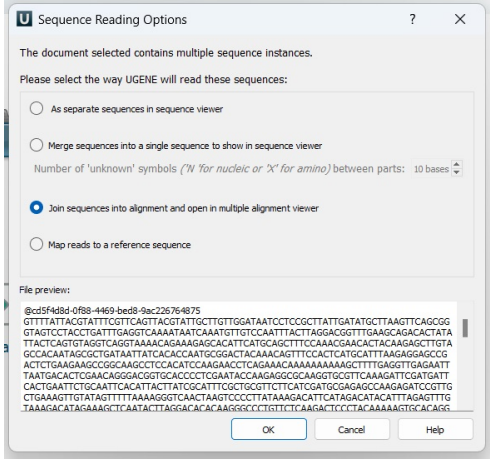
Open File(s)



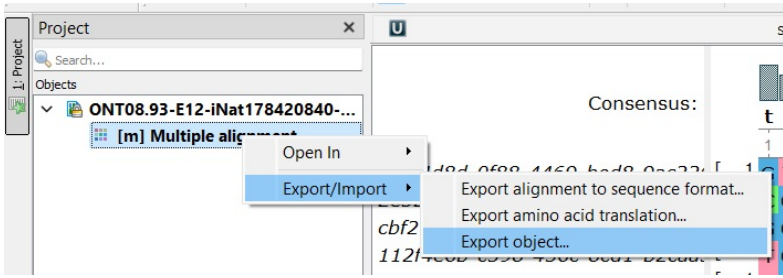
Create Sequence

Create your initial alignment

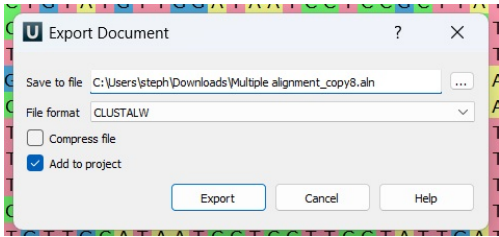
4 Select "Join sequences into alignment and open in multiple alignment viewer" and hit "OK."



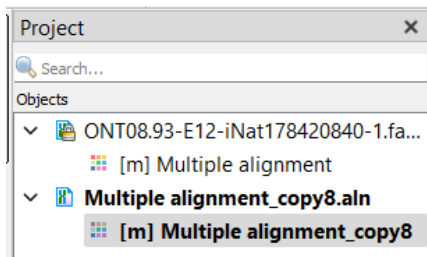
- 5** Right click on the "Multiple Alignment" text in the Objects box on the left hand side. Export/Import -> Export object.



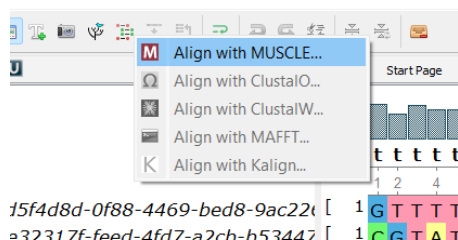
A new "Export Document "dialog box will appear. Hit Export



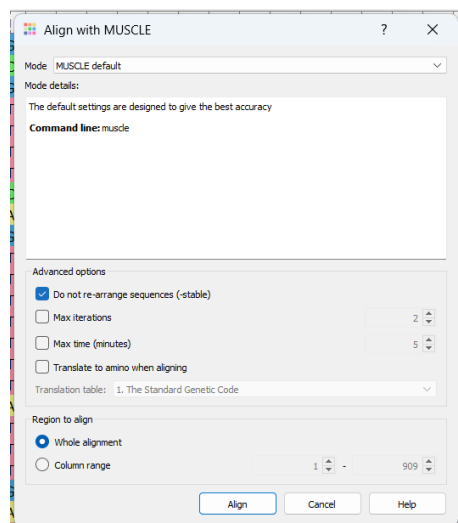
- 6** Highlight the new copy you just made of this document.



On the top toolbar, select the alignment icon and hit Align with MUSCLE (or whatever your favorite algorithm is. They all work reasonably well for this.



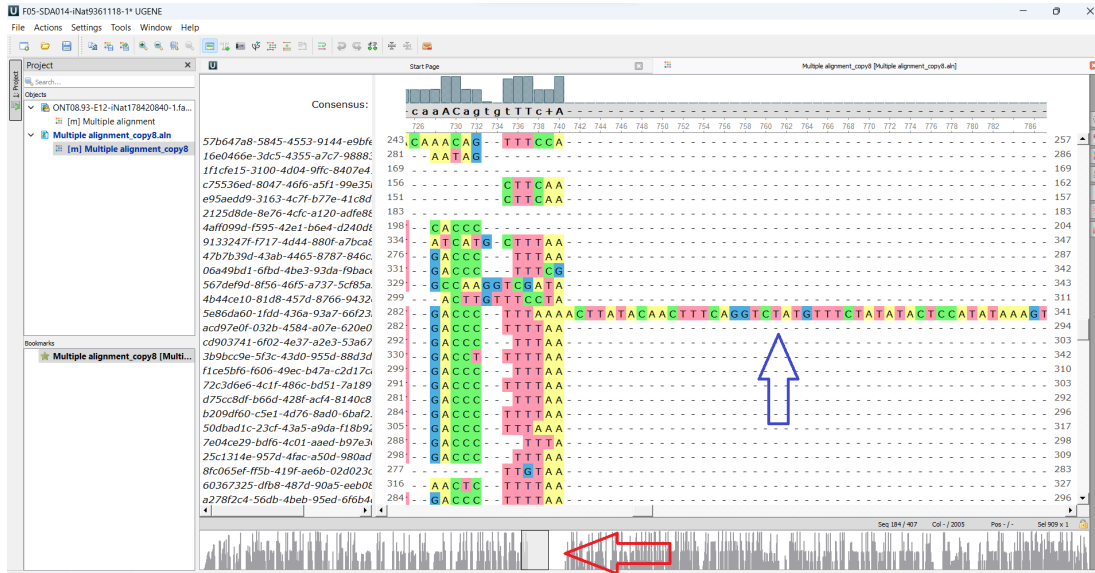
A new dialog box will appear. Uncheck the "Do not re-arrange sequences" button. You want them to cluster with other sequences they align with best.



If you have a large number of reads in your FASTQ file and/or you have a slow computer, the initial alignment may take a few minutes to generate.

Start performing some preliminary triage on your alignment.

- 8 Begin by examining any large gaps in your alignment (red arrow). They are often caused by erroneous reads (blue arrow).



You can highlight the line and hit delete to remove this read from your alignment. Do this for any of the large gaps in your alignment. This should only take a minute or two.

```

5070e19d-6130-4613-a737-3c163a. 281 - - - - - C C C A A G G T C C A T A
4b44ce10-81d8-457d-8766-9432. 299 - - - - - A C T T G T T T C C T A
5e86da60-1fdd-436a-93a7-66f 281 T T - - - G A C C C - - I I T A A A A C T I A T A C A A C T T T C A G G T C T A T G T T C T A T A T A C T C C A T A T A A A G T
acd97e0f-032b-4584-a07e-620e0 281 T T - - - G A C C C - - T T T T A A

```

- 9 The top 5-10% of reads are typically the worst aligning reads in the batch. I will typically just delete them as a batch. At the end of this protocol, we may only be left with the top 10-20 best aligning reads, and that is fine, so if you are working with many reads, just err on the side of deletion for reads that do not align well.

```

8b47f2f3-f9f1-4d93-9e37-7067 82 A T C - - - - - A A A T G T G T C C A A T T G - - - - - A C T C T - - A
3a239118-701b-4bc6-a6ce-864 81 A T C - - - - - A A A G A G T G - - - - - A A A T G C T T G A A A G C C C T A C A A
403012fd-cb7a-4512-9d5c-77b 127 A T C - - - - - A A A T G T T G T C C A A T T T A T A A G A C A - - - - -
390565c8-3373-46d1-ae2c-ac2 117 - - - - - - - - - - A A - T G T T G T G C - A A T T - - - - - T - - - - -
a64c3ab7-ee40-48cb-9bca-ac9 127 C C - - - - - A T C G T T T G T C C A T G G G - - - - - A C C A T T - - A
3acf356-8756-4799-adde-881 116 C T C - - - - - A A A G G T T G T G C A - - - - - G C A A C T - - - A
f3d425c5-6c93-472a-94e4-05f 108 C T C - - - - - G A A G G T T G T C C A - - - - - G C G A C T - - - G
61edbf0c-7467-47ed-b51e-f01 101 C A A T T A G T C - A A A G G T T G T C C A - - - - - G C G A - - - - -
9e4715e9-728c-4e28-adfc-7ce af278d87-8fe1-476a-a223-56f 107 T T A A A - - - - - A A A C A T T G - - - - - G G G G T T - T G T A G
af278d87-8fe1-476a-a223-56f 79 G T T C - - - - - A A - G T T G T C C A A T T - - - - - T C A T G
2a6214a4-16aa-40a2-bc22-e2f 83 A T C - - - - - A A A T G T T G T C C A A T T T G C A T C C A A A C T A A T A
fbd968ca-4d55-4615-ba05-9b112 72 A T C - - - - - A A A T G T T G T C C A A T T T - - - - - C - - - - -
088137df-559f-4ab0-97aa-3b08d 80 A T C - - - - - A A A T G T T G T C C A A T T G G - - - - - A T A C T - T - A
c2459026-2e1e-4498-873b-dee5 74 A T C - - - - - A A A T G T T G T C C A A T T T - - - - - A C T - T - A
7330f311-cb48-4633-8a17-6b8c3 75 A T C - - - - - A A A T G T T G T C C A A T T A - - - - - C T - T - A
4a4b7d47-d6f7-4388-bcca-b6081 76 A T C - - - - - A - T G T T T A C C A A T T T - - - - - A C T - T - A
80b420e3-8576-42ed-aa32-1756 83 A T C - - - - - A A A T G T T G T C C A A T T T - - - - - A C T - T - A
f9aeaf50-0565-4c3c-873b-bca35e 110 A T C - - - - - A A A T G T T G T C C A A T T T - - - - - A C T - T - A
816d5d90-a7b3-4800-bfa1-c2575 68 A T C - - - - - A A A T G T T G T C C A A T T T - - - - - A C T - T - A
6d9d266f-9627-486b-bdbb-524e 119 A T C - - - - - A A A T G T T G T C C A A T T T - - - - - A C T - T C C A G
434addec-308e-42cd-9d85-e18f6 77 A T C - - - - - A A A T G T T G T C C A A T T C - - - - - A A G C A T C C A A
9e8215c6-5980-4280-aa50-9e40f 76 A T C - - - - - A A A T G T T G T C C A A T T T - - - - - A C T - T - A
7150c3ba-d641-40a0-b41a-6babe 111 A T C - - - - - A A A T G T C - T C C A A T T T - - - - - A C T - T - A

```

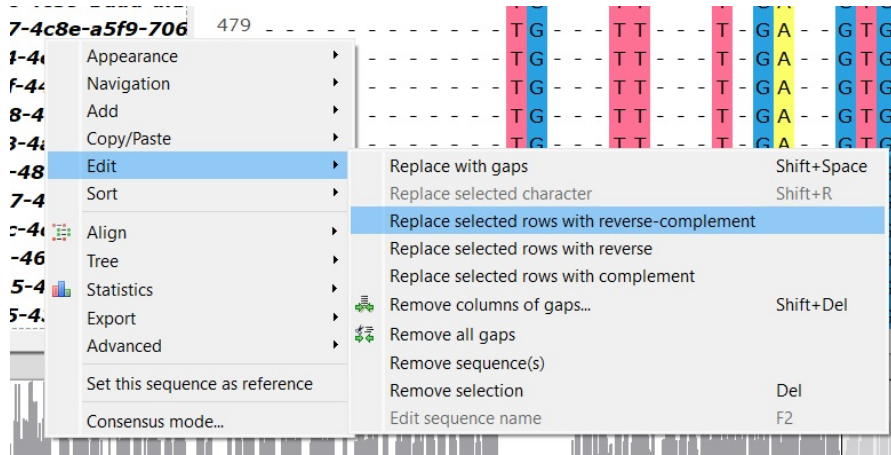
10 Many FASTQ files have both forward and reverse reads within the final demultiplexed FASTQ. If you jump to a random point in the middle of your alignment, at about the midway point of the read count, you will likely see the break point where the alignments are very similar above and very similar below (see red line in the picture).

```

b8f0d2fc-669d-4e52-a40b-2d08e 431 C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 430
e2276b0a-9f23-48cb-94e1-3100a 433 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 467
96c3b279-50e8-46eb-a2b0-9e5c 427 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 460
38404d2e-4192-4ff6-976f-2af197 388 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 421
2e32317f-feed-4fd7-a2cb-b53447 399 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 432
eae2a775-bd24-48eb-918a-1c27e 422 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 455
baf78f01-4cb7-400e-87b2-c14a0f 389 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 422
a40e3f69-c543-4c8b-80cc-52c33 396 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C G G A - 429
da4c13ae-92a4-4306-b08d-8ccce 398 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 431
5766c20c-254a-444a-a01d-9e1d1 398 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 431
47d6ceb8-8e40-4b6d-8f50-2ef3a 391 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 391
57b642a8-5845-4553-9144-e9bfb 396 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 429
1e0d466e-3dc5-4355-a7c7-9888 431 - - - - - C A A G G T G - C G T T - - - - - C A A A G A T T C G A - - - - - T - - - - - G A T T - C A C T G A - 455
1f1cfe15-3100-4d04-9ffc-8407e4 214 - - - - - T G - - - - - G G - - - - - T - - - - - T G A A A T T - - - - - C T G A - 219
c75536ed-8047-46f6-asf1-99e35f 221 - - - - - T G - G G T T - - - - - - - - - - T - - - - - T - - - - - C T T - - - - - 231
e95aed9d-3163-4c7f-b77e-41c8d 216 - - - - - T G - - - - - G G T T C C G - - - - - T - - - - - T T - - - - - 228
2125d8de-8e76-4dc-a120-adfe8e 228 - - - - - T T - - - - - - - - - - T - - - - - T T - - - - - 231
4aff099d-f995-42e1-b6e4-d240d 248 - - - - - C G - - - - - G - - - - - C G - - - - - T - - - - - T T - - - - - T A C A - 260
9133247f-f717-4d44-880f-a7bca8 495 - - - - - T A - - - - - G - T A - G A T T A A - - - - - T - - - - - A T A A T T - - - - - C T C A - 519
47b7b3bd-43ab-4465-8787-846c 434 - - - - - T G - A A T - - - - - T G C A - G A A T T C A - - - - - G - - - - - T - - - - - A T C A - 461
0649b9d1-6fbd-4be3-93da-f9bac 495 - - - - - T G - T T - - - - - T G A - G T G I - C A - - - - - T - - - - - G - A A G T T - - - - - C T C A - 520
567def9d-8f56-46f5-a737-5cf85a 495 - - - - - T G - T T - - - - - T G A - G T G I - C A - - - - - T - - - - - T G A A T T T - - - - - C T C A - 520
4b44ce10-81d8-457d-8766-9432 462 - - - - - T G - T T - - - - - T G - - - - - T G A G T G T C A T T - - - - - - - - - - A - 478
acd97e0f-032b-4584-a07e-620e0 444 - - - - - T G - T T - - - - - T G A - G T G T - C A - - - - - T - - - - - T - A A A T T - - - - - C T C A - 468
cd903741-6f02-4e37-a2e3-53a67 450 - - - - - T G - T T - - - - - T G A - G T G T - C A - - - - - T - - - - - T - A T A T T - - - - - C T C A - 474
3b9bcc9e-5f3c-43d0-955d-88d3d 492 - - - - - T G G A G G C T T G C C G G - - - - - C T C T T C A - - - - - T - - - - - A A A T T - - - - - C T C A - 523

```

Highlight all of the reads above this break point. Right click -> Edit -> Replace selected rows with reverse-complement.



For this alignment, the majority of the bases only had a consensus base for about half of the reads (the gray bar is only going halfway to the top for each position). This is because about half of the reads in this pool were in the wrong direction.

11 Now rerun your alignment with MUSCLE (uncheck do not re-arrange sequences).

Remove the top and bottom 10% or so of reads. These will once again be the worst aligning.

Rerun your alignment with MUSCLE (uncheck do not re-arrange sequences).

Spend a minute or two giving a quick stroll through a number of the largest gaps that are still in your alignment. Remove any sequences that are not aligning nicely. It is fine to be heavy-handed when removing them, assuming you started with a large number of seqs.

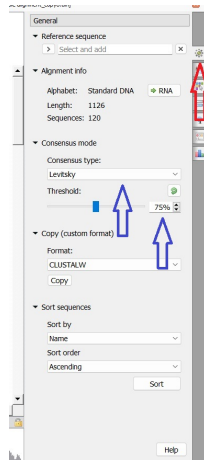
Rerun your alignment with MUSCLE (uncheck do not re-arrange sequences).

12 I will typically repeat step 11 two or three times.

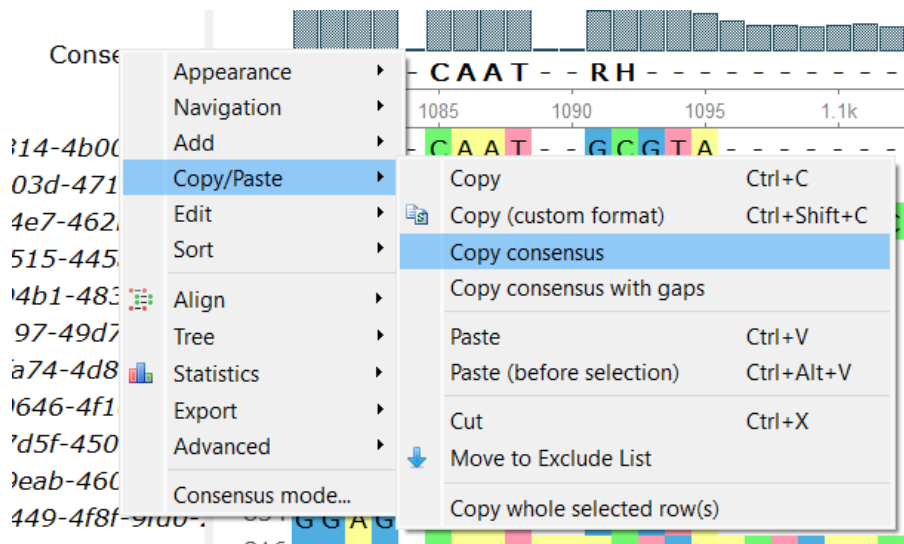
The alignment should be substantially better now. This one in particular went from 2,005 columns to 1,126 columns in the alignment.

Create your Consensus

- 13** On the right hand side, click the gear icon (red arrow in image). Change the Consensus Type to "Levitsky" and the Threshold to "75%." This model will incorporate ambiguous nucleotides into the final consensus.



- 14** Up at the top near where it says "Consensus," right-click -> Copy/Paste -> Copy Consensus.



- 15** Remove any ambiguous bases that may be at the beginning or end of your sequence.

Paste your manually edited sequence into notepad or otherwise to it's final destination. I will often give it a quick MycoBLAST to make sure there is nothing fishy.

<https://mycomap.com/genetics/blast-search/protocol-r115234/>

```
GTATTGCTGTATGTTGGATAATCCTCCGCTTATTGATATGCTTAAGTTCAGCGGGTAGTCCTACCTGATTTGAGGT
CAAAATAATCAAATGTTGTCCAATTTACTTAGGACGGTTTGAAGCAGAYACTATATTACTCAGTGTAGGTCAGGTA
AAACAGAAAGAGCACATTCATGCAGCTTTCCAAACGAACACTACAAGAGCTTGTAGCCACAATAGCGCTGATAAT
TATCACACCAATGCGGACTACAAACAGTTTCCACTCATGCATTTAAGAGGAGCCGACTCTGAAGAAGCCGGCAA
GCCTCCACATCCAAGCCTCAGAAACAAAAAAAAAAGCTTTTGAGGTTGAGAATTTAATGACACTCAAACAGGCA
TGCCTCTCGGAATACCAAGAGGCGCAAGGTGCGTTCAAAGATTTCGATGATTCACTGAATTCTGCAATTCACATTA
CTTATCGCATTTTCGCTGCGTTCTTCATCGATGCGAGAGCCAAGAGATCCGTTGCTGAAAGTTGTATAGTTTTTAAA
AGGGTCAACTAAGTCCCCTTATAAAGACATTCATAGACATACATTTAGAGTTTGTAAAGACATAGAAAGCTCAATA
CTTAGGACACACAAGGGCCCTGTTCTCAAGACTCCCTACARAAAGTGACACAGGTGGATGAAGATTGAAAGAAAA
GCGAGCACTTGCCCTTGAAGAGCCAGCTCAACCTCCCTTTACAATGTTTCAATAATGATCCTTCCGCAGGTTTAC
CTACGGAAACCTTGTTACGACTTTTACTTCCTCTAAATGACCAAGCGGCCAATCTCSGAGCAAT
```