MAR 04, 2024

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

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

Created: Mar 03, 2024

Last Modified: Mar 04, 2024

PROTOCOL integer ID: 96069

Creating a Manual Consensus Sequence from FASTQ with UGENE

Stephen Douglas Russell[1]

[1]Mycota Lab / The Hoosier Mushroom Society

Stephen Douglas Russell

ABSTRACT

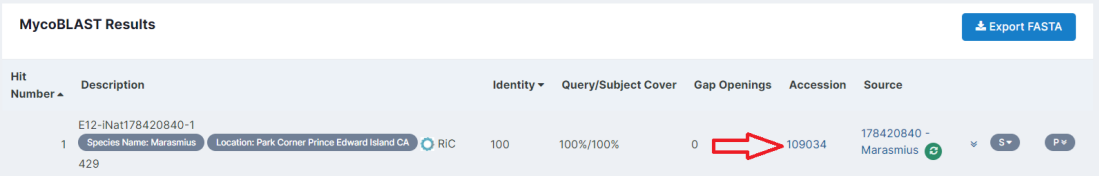
Current versions of automated nanopore amplicon pipelines sometimes (rarely) produce erroneous consensus sequences based on low quality reads that are being incorporated into the final result. It is sometimes helpful to remove these reads from the consensus manually. Further, fungal amplicons can have 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.

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

Install Software, Retrieve Data, Import to UGENE

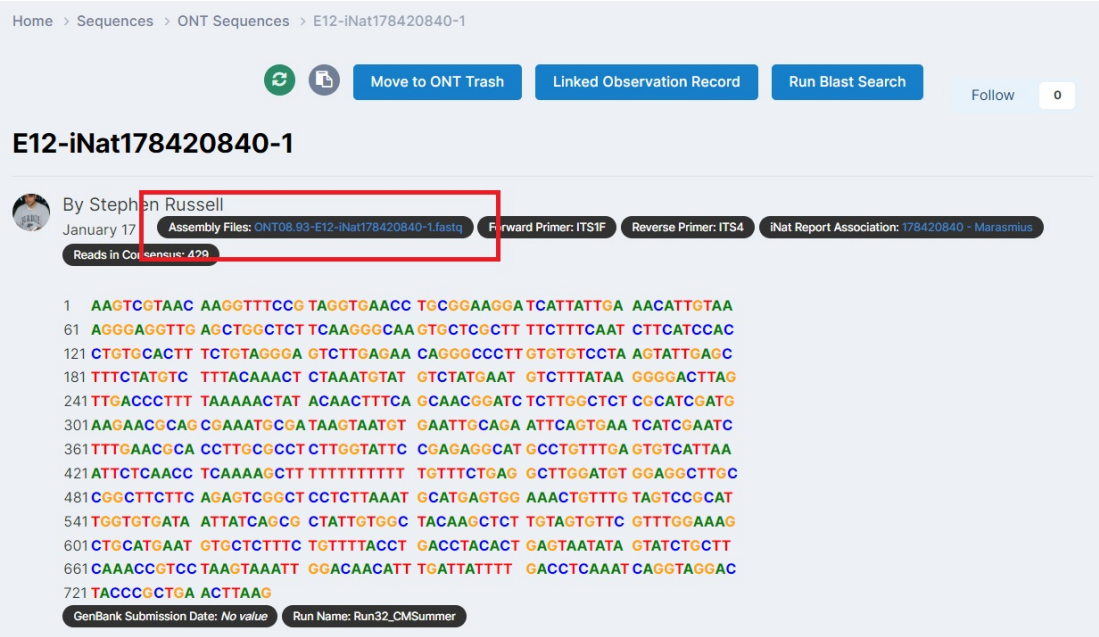
1 Download and install the latest version of UGENE at https://ugene.net/.

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



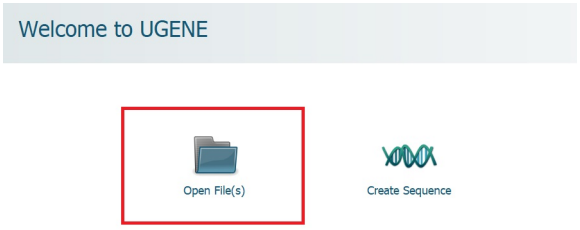
A line within the MycoBLAST results for a given sequence. The red arrow points to the link to the MycoMap sequence accession for that record.

Then download the FASTQ file from the accession page.



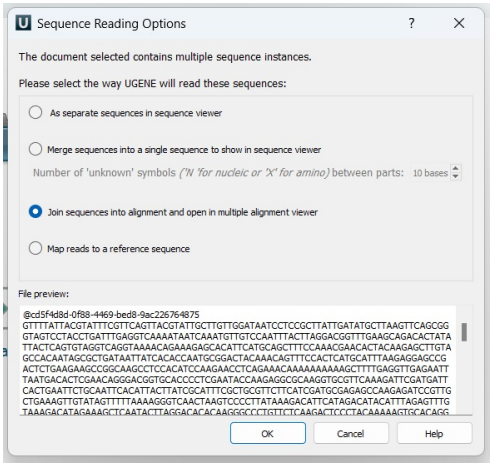
The MycoMap sequence accession page. The red box highlights the link to download the FASTQ file of the raw reads this consensus sequence was formed from.

3 Open your FASTQ file in UGENE.

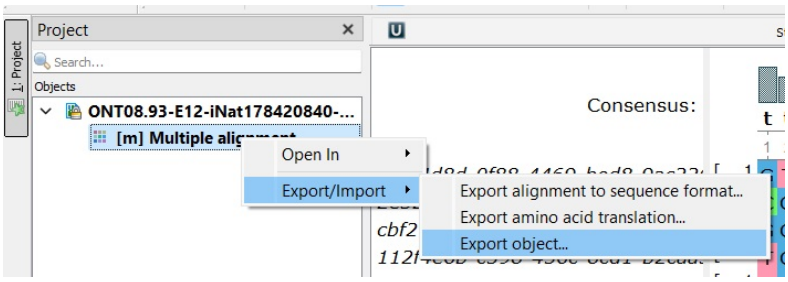


Create your initial alignment

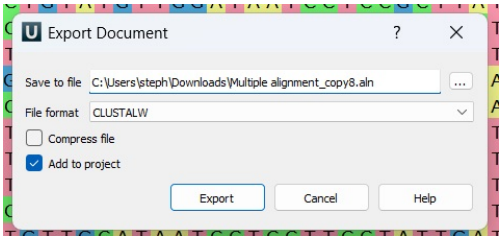
4 Select "Join sequences into alignment and open in multiple alignment viewer" and hit "OK."



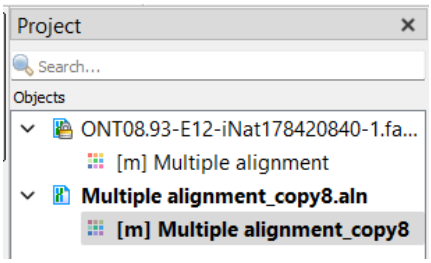
5 Right click on the "Multiple Alignment" text in the Objects box on the left hand side. Export/Import -> Export object.



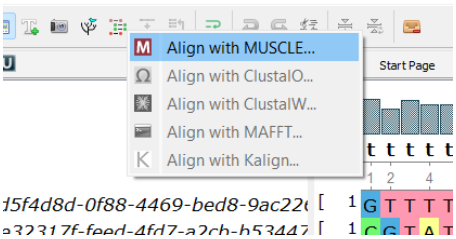
A new "Export Document "dialog box will appear. Hit Export



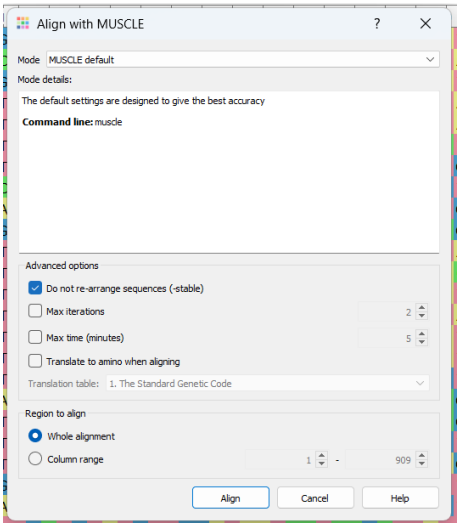
6 Highlight the new copy you just made of this document.



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.

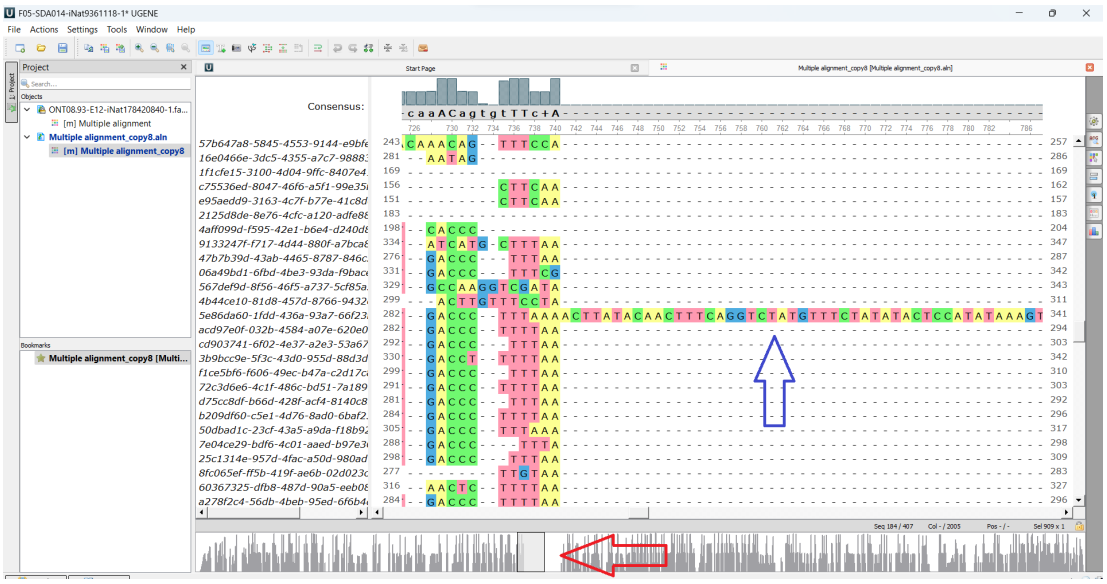


If you have a large number of reads in your FASTQ file and/or you have a slow computer, the initial alignment may take a few minutes to generate.

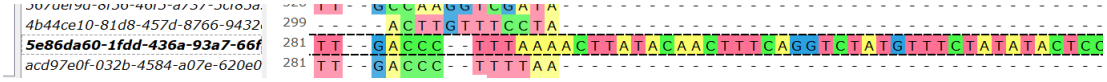
7

Start performing some preliminary triage on your alignment.

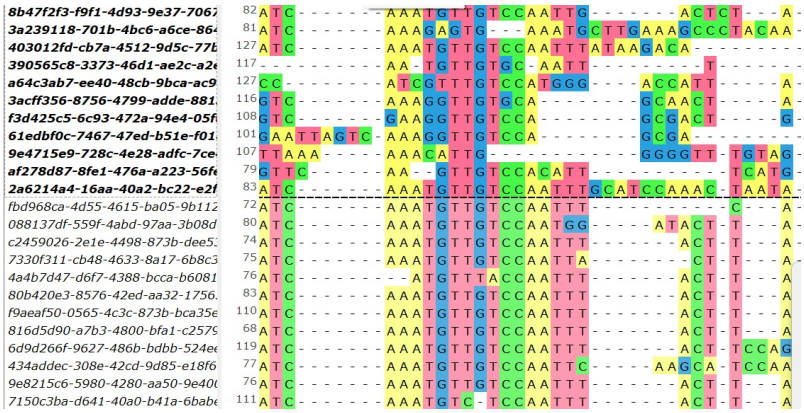
8 Begin by examining any large gaps in your alignment (red arrow). They are often caused by erroneous reads (blue arrow).



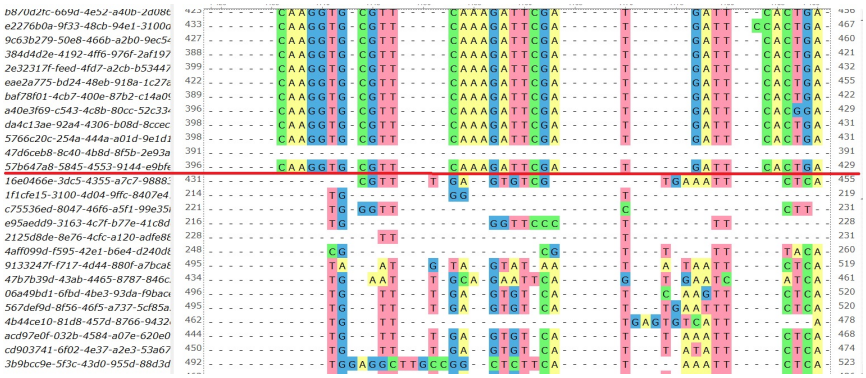
You can highlight the line and hit delete to remove this read from your alignment. Do this for any of the large gaps in your alignment. This should only take a minute or two.



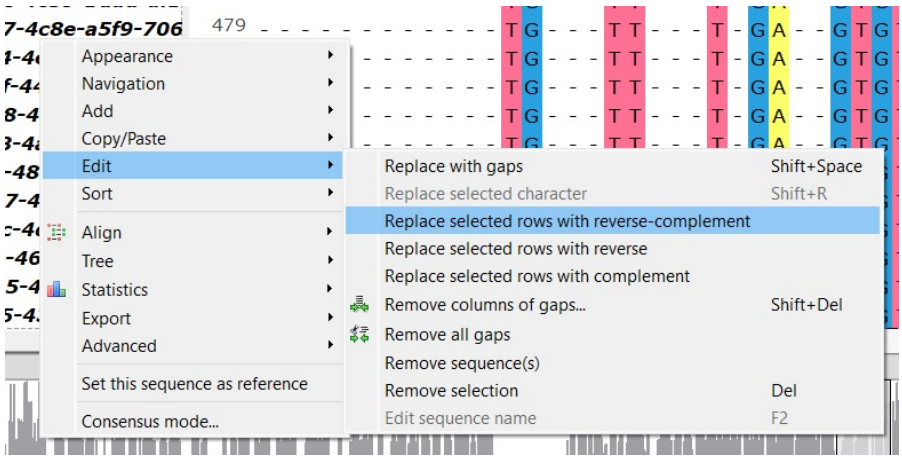
9 The top 5-10% of reads are typically the worst aligning reads in the batch. I will typically just delete them as a batch. At the end of this protocol, we may only be left with the top 10-20 best aligning reads, and that is fine, so if you are working with many reads, just err on the side of deletion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



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).



Highlight all of the reads below 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.

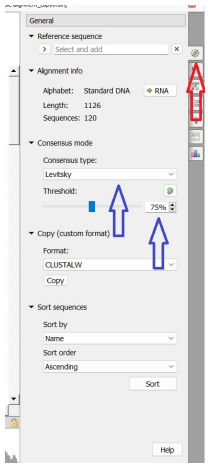
Rerun your alignment with MUSCLE (uncheck do not re-arrange sequences).

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

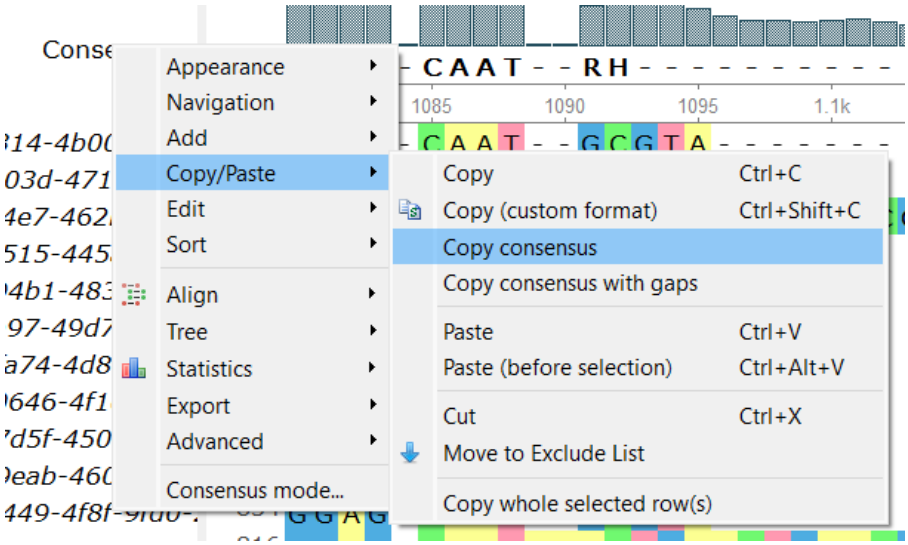
The alignment should be substantially better now. This one in particular went from 2,005 columns to 1,126 columns in the alignment.

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
GCCTCCACATCCAAGCCTCAGAAACAAAAAAAAAAAGCTTTTGAGGTTGAGAATTTAATGACACTCAAACAGGCA
TGCCTCTCGGAATACCAAGAGGCGCAAGGTGCGTTCAAAGATTCGATGATTCACTGAATTCTGCAATTCACATTA
CTTATCGCATTTCGCTGCGTTCTTCATCGATGCGAGAGCCAAGAGATCCGTTGCTGAAAGTTGTATAGTTTTTAAA
AGGGTCAACTAAGTCCCCTTATAAAGACATTCATAGACATACATTTAGAGTTTGTAAAGACATAGAAAGCTCAATA
CTTAGGACACACAAGGGCCCTGTTCTCAAGACTCCCTACARAAAGTGCACAGGTGGATGAAGATTGAAAGAAAA
GCGAGCACTTGCCCTTGAAGAGCCAGCTCAACCTCCCTTTACAATGTTTCAATAATGATCCTTCCGCAGGTTCAC
CTACGGAAACCTTGTTACGACTTTTACTTCCTCTAAATGACCAAGCGGCCAATCTCSGAGCAAT