Mitochondrial Annotation Protocol

using JX412834 as trial sample

Ivan Croydon Veleslavov & Jia Le Lim
Edited by: Dr Benjamin Linard

Last updated: November 5, 2016

Preface

Mitochondrial Annotation Protocol was written as a guide to those new in Geneious and as a reference for editing sequences to be submitted onto NCBI GenBank.

Acknowledgements

We would like to thank our supervisor, Dr Benjamin Linard, for his guidance throughout the project and everyone in the molecular write-up lab for their input and comments on the protocol. Lastly, we would like to thank Professor Alfried Vogler and Natural History Museum for making this internship possible.

Contents

Introduction

Geneious

Geneious is a very powerful piece of software which is used as a platform for a wide range of bioinformatical tasks. Importantly, the software works across the commonly used platforms (Windows, Mac, Linux) and is used by many universities and institutions world wide, meaning that large scale collaborations within Geneious are possible. The program is capable of performing sequence alignments, sequence assemblies and phylogenetic analyses and much more (Kearse et al., 2012). This protocol is based on Geneious 6.1.8.

Basics of using Geneious such as zooming in shortcuts and exporting documents are found in the tutorial under Help. Otherwise you can use the Geneious user manual for v6.1. You can also choose to set shortcuts for commonly used functions under Tools, Preferences.

Why are we writing this protocol?

Though very powerful, Geneious does take a little while to get used to, especially if users are unfamiliar with other bioniformatics software platforms. As such, researchers often need training to get the most out of the program. Here at the Natural History museum, students and new researchers arrive regularly, often to complete relatively short placements, and thus a lot of time has been spent by supervisors teaching the basics of Geneious.

By writing this protocol we are aiming to reduce the time invested by supervisors at this step so that more focus can be given to the goals of the project, rather than the technical operation of the software. This protocol was created to outline the basic functionality of Geneious, provide fixes for the common problems we encountered and also include techniques to streamline processes to improve user efficiency.

Mitochondrial genomes, our project

There has been a recent drive towards the collection, sequencing and identification of mitochondrial genomes to rapidly identify species from large collections. This has been particularly effective in invertebrate species where many individuals, some of which are morphologically indistinct, can be identified at a fraction of the cost and time of traditional methods. However, metamitochondrial genomics relies on a good database of sequences to compare field samples from - as well as streamlined bioinformatical processes to keep up with the rate of specimen collection.

Our project has been to improve this database by creating a taxonomic tree of well annotated reference mitochondria to rapidly and accurately transfer gene and tRNA annotations to unannotated mitochondrial sequences so that they can be submitted to GenBank and accessed by other scientists. A large proportion of the annotated invertebrate mitochondria, especially within the Coleoptera, have been submitted by the Natural History Musuem and as such we have a large amount of responsibility in designing an efficient and easily replicated protocol for future submissions.

Summary of Steps

1. Assembling sequences. Select all raw data and select De Novo Assemble using the correct configurations.

2. Blasting against Barcodes. Sequence search using the barcodes provided and change the supercontig names accordingly.

3. Editing supercontigs. Attempt to resolve mismatches and try to circularise sequences that are around 18,000 bps in length.

4. Annotating genomes. Find a reference complete mitochondria by blasting the COX1 gene on your supercontig. Transfer the gene annotations onto your supercontig. Manually edit the start and stop codons of all 13 genes.

5. Removing Annotations. Remove unneeded and irrelevant annotations transferred from the reference genome.

6. Adding tRNAs. Use Putty and log on to ctag. Thereafter, assemble the results with your supercontigs and transfer annotations.

7. Remove other Annotations. Check again to see if all unwanted annotations have been removed.

8. Submitting to GenBank. Generate three different files: Template file(suffix .sbt), Fasta file(suffix .fsa) and the Feature table file(suffix .tbl). Use tbl2asn to generate a .sqn file to be submitted to GenBank.

9. YAY. Pass the Geneious files and any other relevant documents to your supervisor. Good job! You deserve a beer! Cheers! ☺ ☺ ☺.

1 Assembling sequences

The whole genome shotgun (wgs) sequencing produces tens of millions of short reads of about 250bps. These reads are then assembled by different assemblies such as Celera, IDBA, Newbler or Ray into contigs that range from 1000 bps to 30,000 bps. These assemblies are computationally heavy and takes one or several weeks to run. Some work better for certain species, while others work better for other things. By assembling contigs from different assemblies, also known as supercontiging, we can obtain supercontigs, which are longer and more reliable than contigs. Supercontigs allow us to detect problems in contigs, where reads from different species may have been merged together.

Select all given contigs. Go to Tools and under Align/Assemble, click on De Novo Assemble... which will direct you to this popup window (Figure 1). Select Custom sensitivity. Under more options, select Save assembly report, Save list of unused contigs and Save in sub-folder. At the bottom left, change Minimum Overlap to 1000. Try different combinations between 1-5% of Maximum Per Read under Allow Gaps and Maximum Mismatches Per Read. You will obtain the following results (Figure 2) after pressing okay.

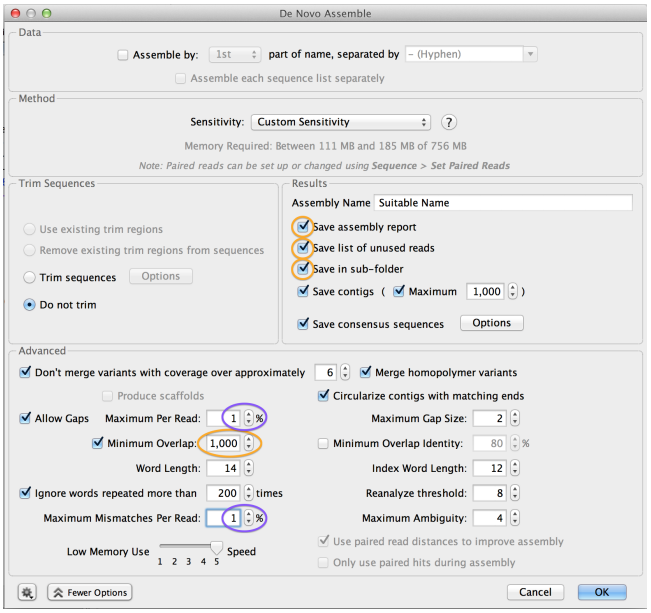


Figure 1: Assembly Popup Window

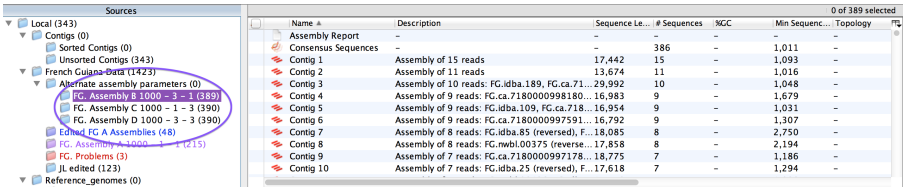


Figure 2: Assembly Results

For convenience, you can name your folders in the following way: 'Suitable Name Minimum Overlap - Maximum Per contig - Maximum Mismatches Per contig', as seen in the purple circle in Figure 2. In our project, we choose to use only FG. Assembly A 1000 - 1 - 1, as increasing both Maximum Per contig and Maximum Mismatches Per contig only gave us one extra supercontig.

2 Blasting against Barcodes

Barcodes are COX1 and CYTB gene sequences from individual samples. The contigs obtained are from a pool of all these samples. Using the barcodes, we can find out which contigs belong to which sample.

In Geneious, go to Tools, Sequence Search... Click on Add/Remove Databases, Add Sequence Database... Select Custom BLAST, give it an appropriate name and select Create from file on disk. Browse and select the file with your barcode sequences. Leave Type as Nucleotide and click OK.

Before barcoding, find contigs that are unable to assemble but are longer than 8,000 bps. These may be unique mitochondrial genomes on their own. You can find these in the file called Unused contigs. Left click on their name and select Extract Regions... to give you the file in the blue circle (Figure 3).

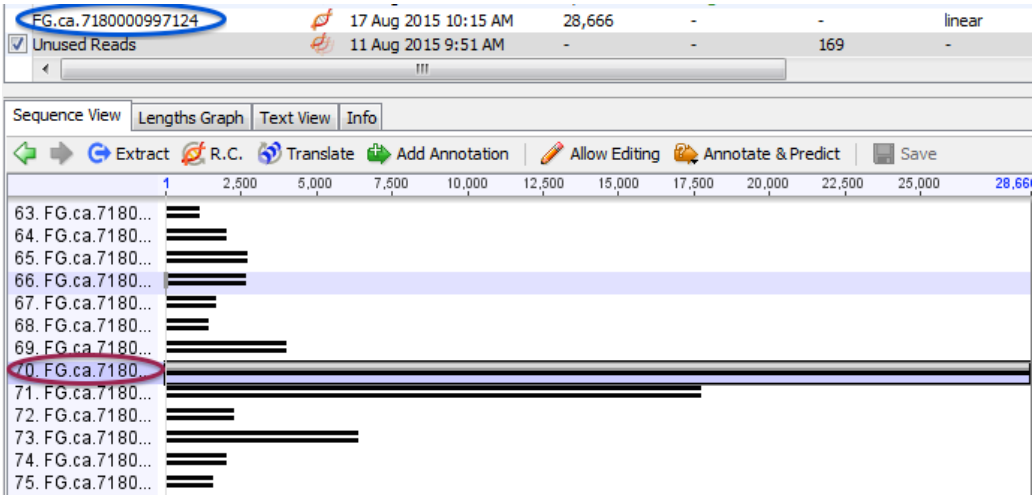


Figure 3: Extracting unused contigs.

For contigs that are longer than or approximately equal to 34000 bps, extract half of the sequence by highlighting the sequence and left click to select Extract Regions. Delete this region from the original sequence. Select the extracted sequence and the halved original sequence and go to Tools, Align, Pairwise alignment. This enables us to see if the two halves are from the same mitochondrion or if it is a concatenation of two mitochondria of two different species. As seen from the result in Example 4, the two halves do not belong to the same mitochondria and should be barcoded separately. Otherwise, delete one copy of the mitochondria.

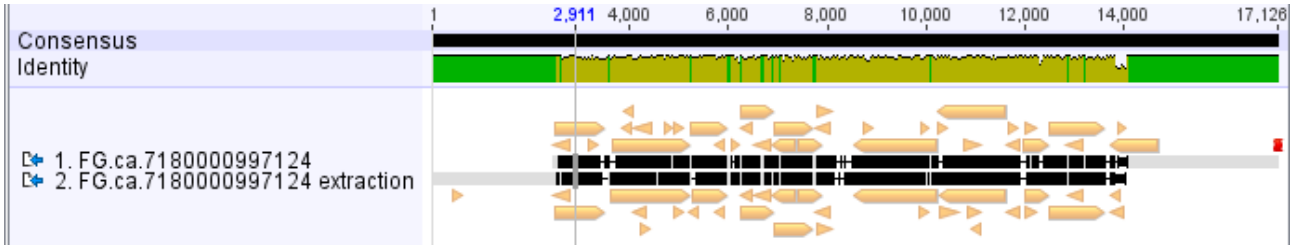


Figure 4: Eliminating chimeras. Example 4

Select all of the supercontigs to be barcoded. Go back again to Tools, Sequence Search... Select the respective Query and under Database, you should now be able to select under Custom BLAST your

newly created database. Leave other options as default and click Search (Figure 5).

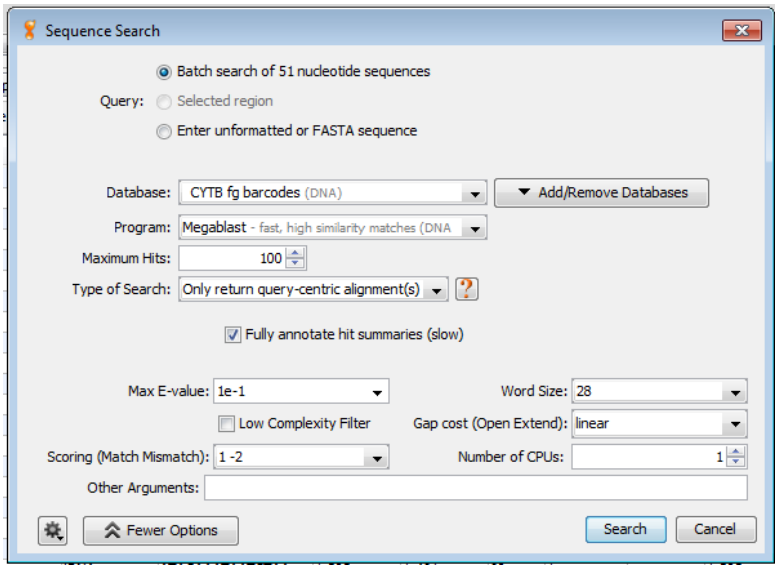


Figure 5: Blasting against barcodes

This should create new folders under Searches (Figure 6). Click on Columns to show % Identical Sites and % Pairwise Identity. Clicking on % Identical Sites and % Pairwise Identity will reorder the Search Hits to give you Hits with the highest similarity.You can then click on the Sequence Name of the highest similarity and return to Alignment View. This will highlight the Search Hit you have just selected in the Alignment View.

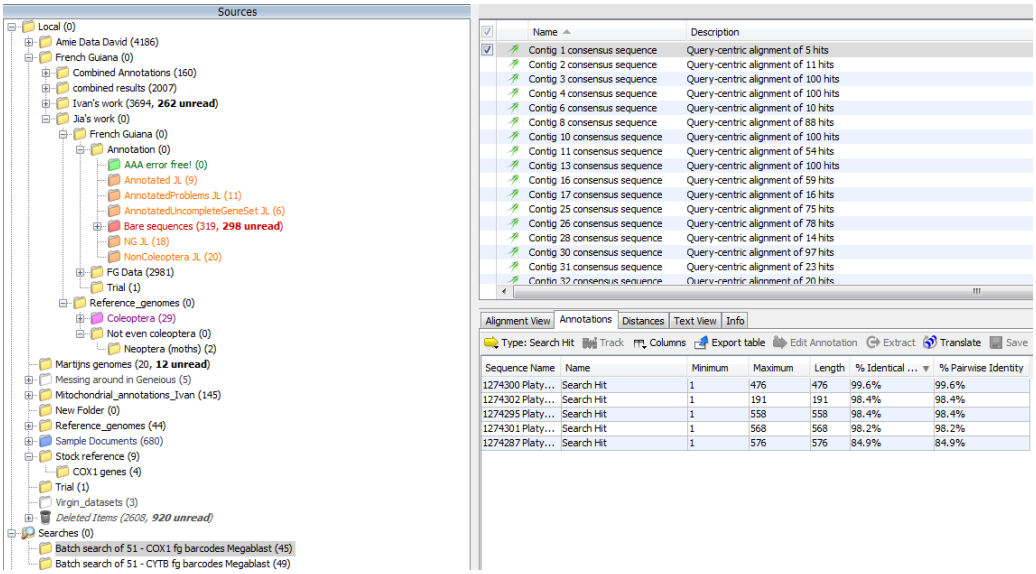


Figure 6: Blast search of barcodes

If it is a 100% similarity, Left click the Search Hit name, Copy Name and record it down in the data sheet. At the same time return to your query folder, find the corresponding supercontig and click on the supercontig name twice, enabling you to change the supercontig name. Alternatively one can click the supercontig once and use the keyboard shortcut F2 to edit the supercontig name. Left click and select Paste to easily change the supercontig name to the title found on the barcode with the highest similarity. This is usually a string of numbers_Name_P1 or P2. For example, 1274645_Staphylinidae_P2. Some supercontigs will correspond to COX1 barcodes or CYTB barcodes of a specimen or both.

supercontigs that have COX1 barcode of one specimen and CYTB barcode of another specimen are chimeras and should be removed.

2.1 Case Studies

However, in the case that a Search Hit has a high % similarity of over 99%, record it down and note down the number of mismatches. Check through the mismatches. In some cases, it may be a M mismatch with a C. However, M represents either cytosine or adenine and hence this is not considered a mismatch, but Geneious does not see that. Also, every number should represent one species and hence there should be no more than one supercontig for each species. Mismatches may be due to editing or sequencing errors and hence Blast matches with few number of mismatches should be correct as long as each number only corresponds to one supercontig.

Sometimes, more than one supercontig appears to come from the same species. In Example 7, both supercontigs gave a blast result of species number 216 for COX1. However, one gives the blast result of species number 126 for CYTB and the other 216 for CYTB. Hence, the former is a chimera which has the gene COX1 from species 216 but CYTB from species 126 and hence will be deleted. A pairwise alignment of the two sequences will give the result of Example 7.

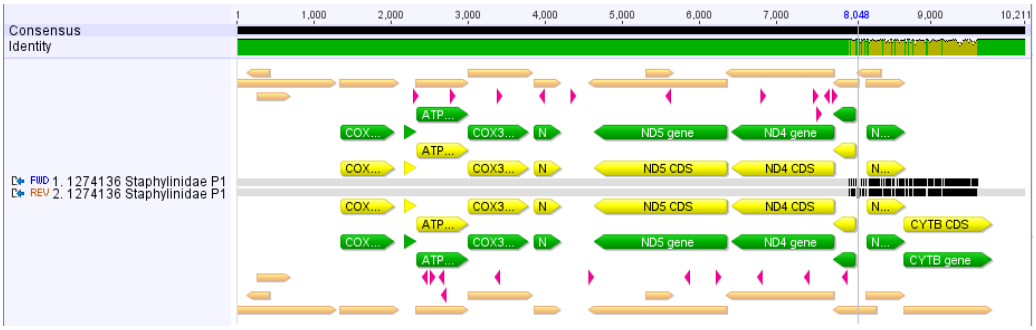


Figure 7: Pairwise alignment. Example 7.

2.2 Keeping a datasheet

Similarly, keep a record of the barcode numbers, whether they are present in CYTB and/or COX1 and which supercontigs came from which specimens, the number of mismatches it has with the Search hit and if it corresponded with both barcodes or only one (Figure 8). We did a second datasheet to record each supercontig's Search hit to detect chimeras (Figure 8).

BMNHE	CYTBBarcode	COX1Barcode	Contig	Mismatches	COX1 and CYTB
1274145	1	1	143	0	CYTB
1274146	1	0			
1274169	1	1			
1274170	1	1			
1274171	1	1			
1274172	1	0			
1274173	0	1			
1274176	1	1			
1274178	1	1			
1274179	1	0			
1274180	1	0	90	0	CYTB
1274181	1	1			
1274182	1	1			
1274183	1	1			
1274184	1	0			
1274186	1	0			
1274187	1	0			
1274188	1	0			
1274189	1	0			
1274190	1	1	52	0	1
1274192	1	0	53	1	CYTB

	A	M	N	O
1	Contig	COX1Barcode	CYTBBarcode	
2	1	Platypodinae/Scolytinae P1	01 Platypodinae/Scolytinae P1	
3	2	1274279 Platypodinae/Scolyti	1274279 Platypodinae/Scolytinae P1	
4	3	1274706 Staphylinidae P2	1274706 Staphylinidae P2	
5	4	NA	NA	
6	5			
7	6	NA	1274272 Curculionidae P1	
8	7			
9	8	1274246 Staphylinidae P1	1274246 Staphylinidae P1	
10	9			
11	10	1274704 Staphylinidae P2	1274704 Staphylinidae P2	
12	11	1274768 Elateroidea P2	1274768 Elateroidea P2	
13	12			
14	13	1274708 Staphylinidae P2	1274683 Staphylinidae P2	
15	14			
16	15			
17	16	1274780 Erotylidae P2	1274780 Erotylidae P2	
18	17	1274377 undescribed P1	1274377 undescribed P1	
19	18			
20	19			
21	20			
22	21			
23	22			
24	23			
25	24			
26	25	1274668 Staphylinidae P2	1274668 Staphylinidae P2	
27	26	1274226 Staphylinidae P1	1274226 Staphylinidae P1	

Figure 8: Recording barcodes

3 Editing supercontigs

Edit the supercontigs that are barcoded. These are the supercontigs that you are interested in.

Click on one of the supercontigs to see the supercontig sequence. At the right side menu, click on the orange arrow with a question mark and you can select or deselect Find ORFs. It is advised to increase the Minimum size to 200 and to change the Genetic code to the appropriate study organism i.e. Invertebrate Mitochondrial (Transl_table 5). Selecting ORFs can help with aligning contigs that seem out of place or to increase overlapping areas of contigs. Clicking on the title of the contig selects the entire contig which allows you to thereafter click and drag the entire contig to its preferred place (Figure 9). You can also delete the entire selected contig by pressing Delete.

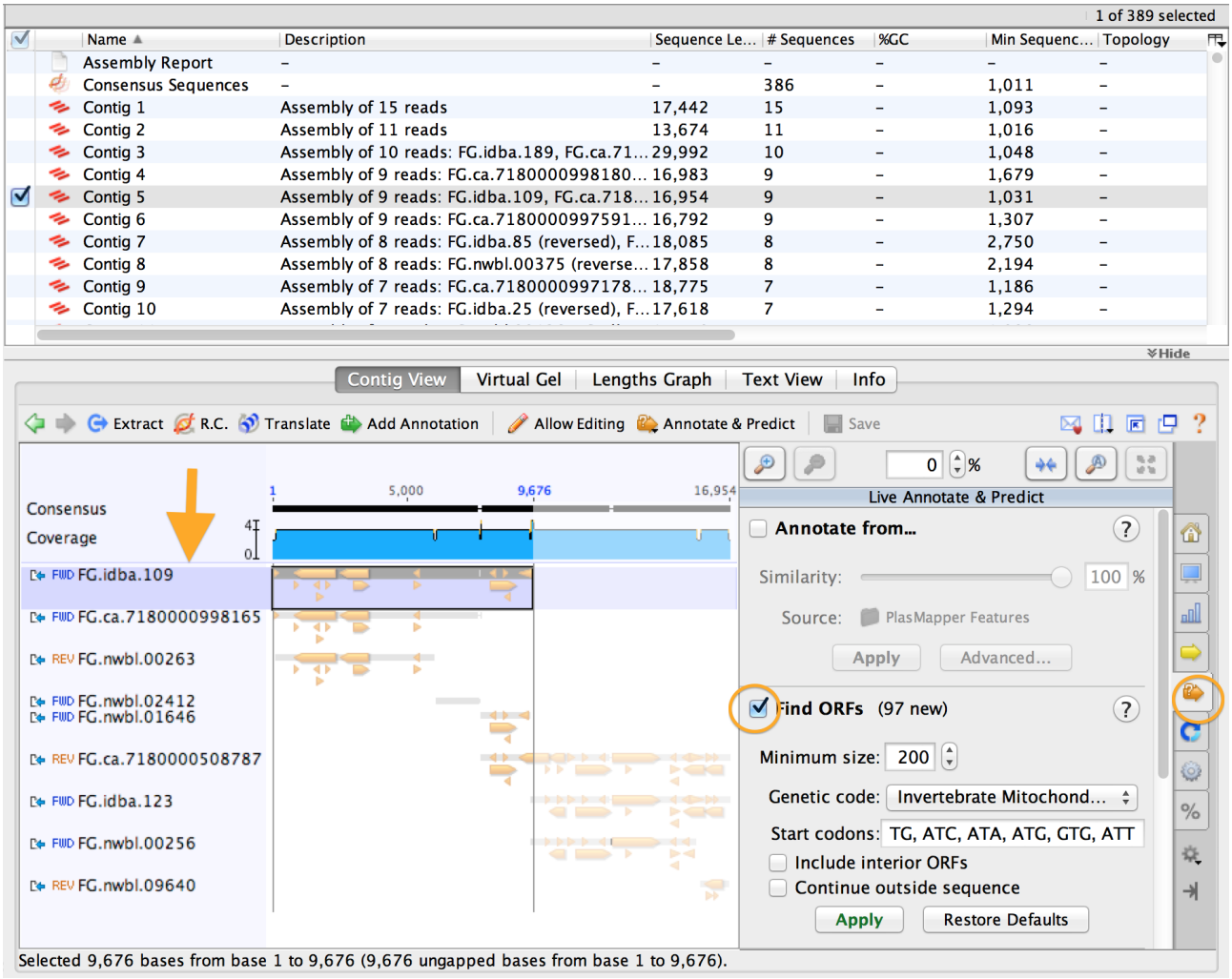


Figure 9: Selecting supercontigs

After aligning the contigs, deselect Find ORFs to reveal black strips which show bases on contigs that do not agree with other contigs. Ends of contigs usually have a higher error rate and hence, ends of contigs with black strips are usually deleted (Example 10). To delete, highlight the sequence and press Delete/Backspace. When deleting the left end of the contig, hold Alt and press Delete/Backspace or the entire contig will shift to the left and go out of sync with other contigs.

You can also add in additional bases by typing A,T,C or G. These bases will be underlined green. You can also delete bases and gaps by Delete/Backspace. Deleted bases will be underlined red.

One can check for overlapping areas of contigs by highlighting sequences and pressing Ctrl-C (Copy), Ctrl-F (Find) and Ctrl-V (Paste) to paste the sequence into the search box and click Find Next. To circularise consensus sequence of supercontigs when they have overlapping ends, you have to manually edit the sequence and delete the overlapping area. Take note that you can only circularise consensus sequences but it is easier to edit and find overlapping areas using supercontigs as you can make a better informed decision of what to delete when viewing the contig errors (Refer to Section 3.2).

To export consensus sequences, select all of the targeted supercontigs and go to File, Export, Consensus Sequence(s)... Choose 0% - Majority for Threshold, and tick Ignore Gaps. Leave other options as default and click OK.

3.1 Case Studies

Here are some case studies of editing supercontigs. Do remember to check for overlaps of ends after editing when sequences are ~15,000 bps, the usual length of invertebrate mitochondrion genome (Refer to Section 3).

3.1.1 Editing mismatches

In Example 10, we would delete the end of the contig in the orange circle and leave the rest of supercontig. Problems in the purple circle cannot be resolved as no third sequences can confirm which is correct. Welcome to Biology, where nothing is either black or white.

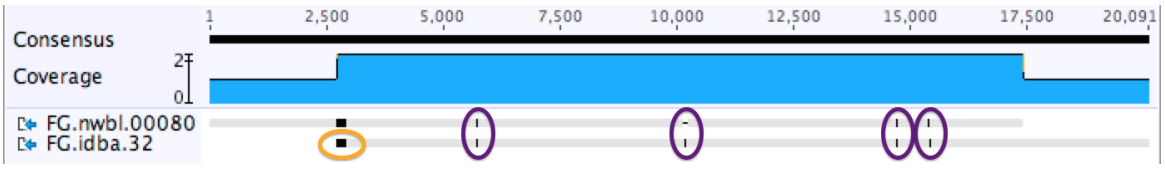


Figure 10: Editing supercontig Example 10

In Example 11, we would ignore the gap in the purple circle as the consensus sequence will include the correct base. You can otherwise zoom in and add in the correct base manually. In the orange circle, we would delete the gap in the consensus sequence as two contigs agree that there is no nucleotide at that base position. As the sequence is much less than 15,000 bps, there is no need to find overlaps.

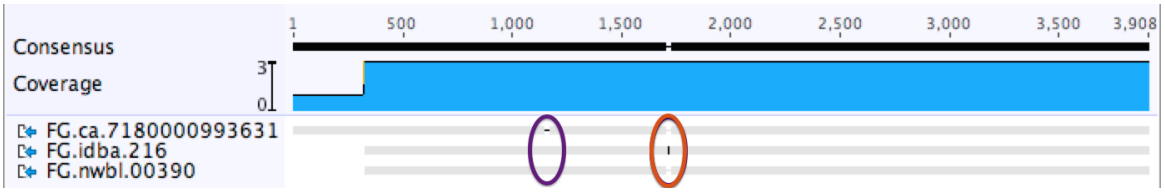


Figure 11: Editing supercontig Example 11

In Example 12, we would delete idba.12411 as it is a short sequence that provides an error. Furthermore, there are two or more contigs that confirms the consensus sequence and hence we do not need

idba.12411. We would ignore the rest of the errors as there are at least two contigs that agree with each other and thus these errors will be ignored by the consensus sequence.

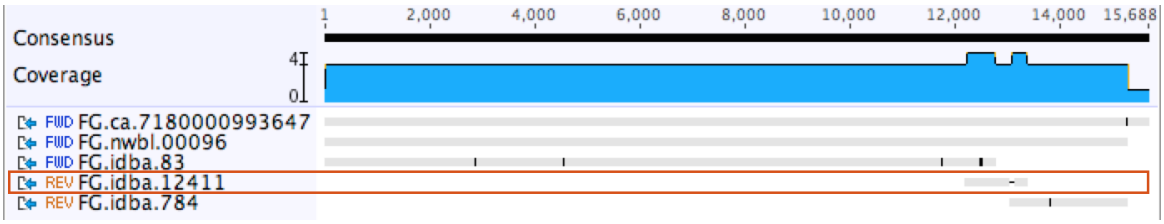


Figure 12: Editing supercontig Example 12

Example 13 is awesome because we do not have to edit anything ☺ but remember to check for overlaps (Refer to Section 3).

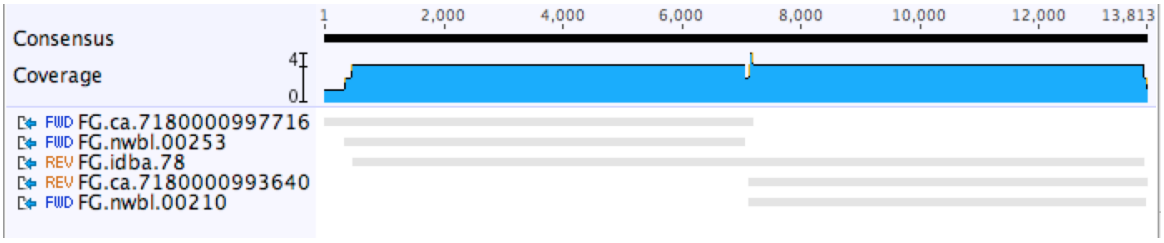


Figure 13: Editing supercontig Example 13

Example 14 is not awesome ☹ but you will come across many of these. Ends in the two orange circles can be deleted. You can choose to ignore ca.7180000507897 or delete it. Zoom into the purple circles to see if the black strips of different contigs are located on the same base. If they are, they cannot be edited, record them down onto a datasheet. If they do not coincide, you can ignore them and let the consensus sequence do its job. Again, check the ends for overlaps (Refer to Section 3).

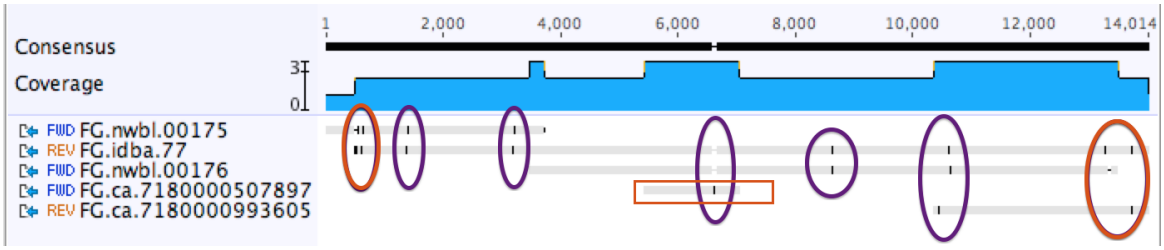


Figure 14: Editing supercontig Example 14

In Example 15, we would delete, in the orange circle, the extra base causing the gaps in two other contigs and delete the very end of nwbl.00195. The errors in the purple circles will be ignored. In the blue circle, we would delete the end of idba.91. Lastly, check for overlapping ends.

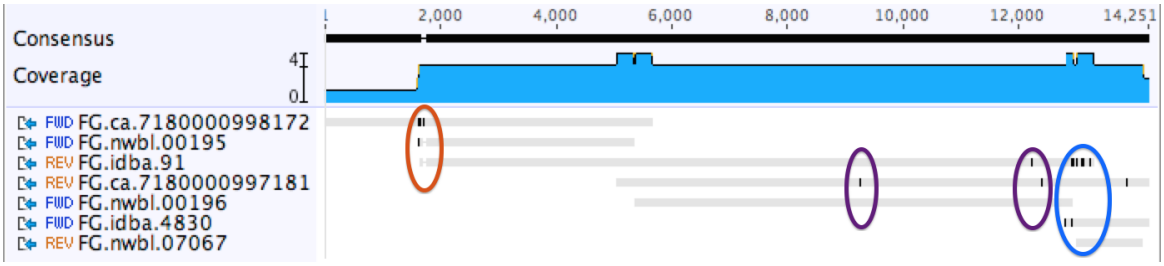


Figure 15: Editing supercontig Example 15

3.1.2 Editing long supercontigs

Example 16 shows how three products have been stacked to create a mammoth 36551 bp sequence. These very long supercontigs are a common problem during this process but can be easily sorted. It is advisable in these situations to ensure 'Find ORFs' is ticked for this procedure as it will help to visualise the mitochondrial regions for alignment. One can also select 'Find Annotations' to see what the ORF regions code for (Example 17). Both 'Find ORFs' and 'Find Annotations' function is found on the 'Live Annotate and Predict' Toolbar (Figure 34).

ORF stands for Open contiging Frames, which shows possible gene positions calculated by the program. They may not all represent real genes, but are frequently used as a guideline to find where actual genes are.

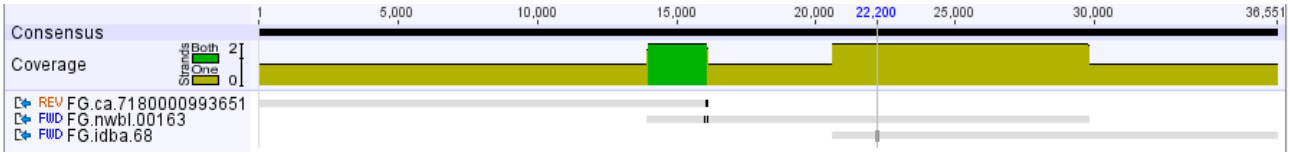


Figure 16: Dealing with long sequences Example 16

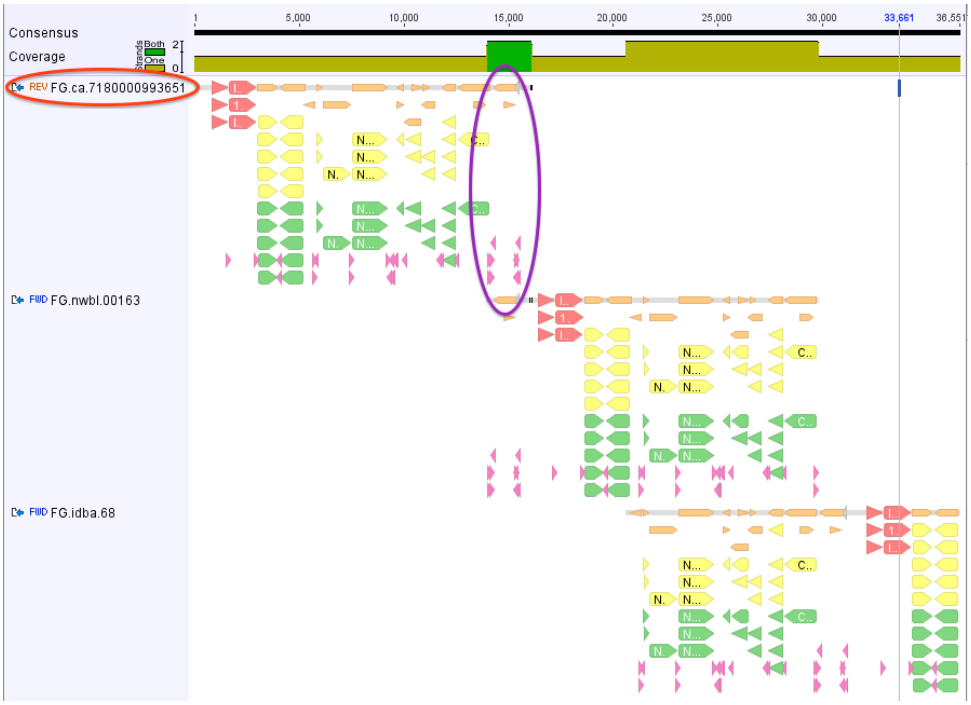


Figure 17: Dealing with long sequences Example 17

With these very large supercontigs we must confirm that the three contigs are from the same mitochondria. The overlap between nwbl.00163 and idba.68 is large, contains the same protein coding genes (as visualised by the ORFs) and is largely error free. As such we can be fairly confident that they do indeed represent the same mitochondria. However, the overlap between ca.7180000993651 is considerably smaller and might not indicate true mitochondrial identity.

To ensure all the contigs are from the same mitochondria it is advisable to shift them all to maximise overlap. This will increase the length over which we can look for mismatches, and thus allow us

to confirm or deny mitochondrial identity. To shift a contig, ensure 'Enable Editing' has been selected, click on the name of the contig on the left hand side and simply drag the contig to where you want it, relative to the other contigs. One can use the gene annotations and ORFs, as well as the zooming in function (Ctrl + Scroll) to correctly line up the contigs so that any mismatches are the result of sequence differences and not just poor alignment. In this example, ca.7180000993651 was shifted to the right to increase the overlap with nwbl.00163 and idba.68 (Example 18)

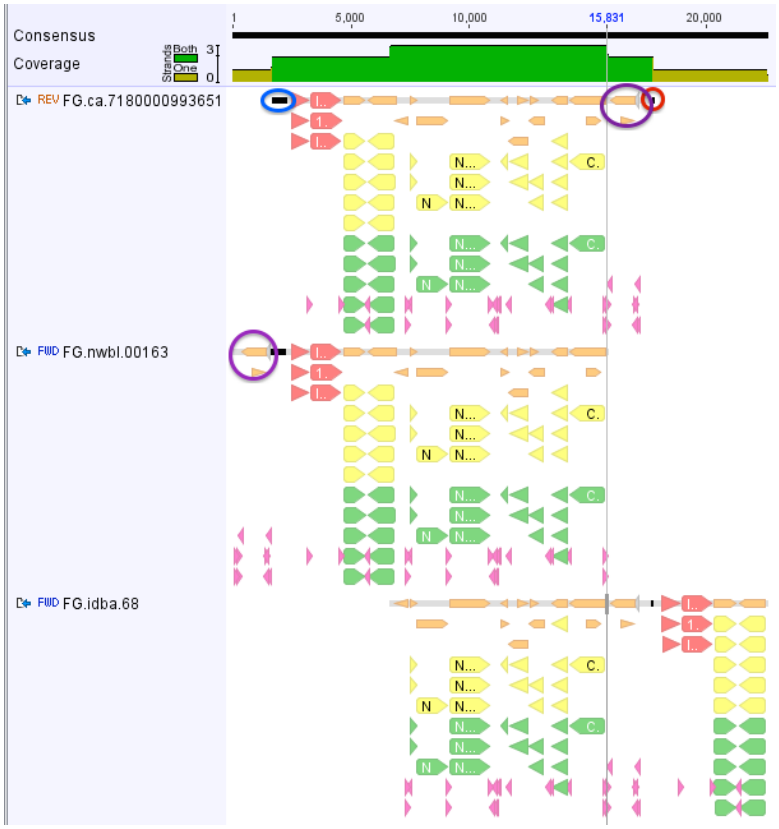


Figure 18: Dealing with long sequences Example 18

As can be seen in Example 17 the sequences in the purple circle are the same, however, the sequence in the purple circle does not match that in the blue circle found in Example 18. This means that the right end of ca.7180000993651 is not the same as its left end. As sequences on both side of the many mismatches area (indicated by a black banded area) in nwbl.00163 has been confirmed by the ca.7180000993651 and idba.68, the black banded area in nwbl.00163 is likely to be correct. Hence, we will delete the two ends of ca.7180000993651 found in the purple and red circle (Example 18).

Do remember to untick the 'Find ORFs' and the 'Find Annotations' box to see the mismatches shown as black bands that sometimes lie underneath (subsubsection 3.1.1).

In another more obvious example, the blue and red circle in Example 19 do not agree with each other. After shifting nwbl.00043 to the left, we now see that the left end of idba.20 agrees with the right end of nwbl.00043 but not with the left end of ca.7180000996827. As it is two against one, we will delete the control region present in the blue circle in ca.7180000996827 (Example 20). Control regions tend to be problematic, and this is one of the cases where the genes of one species has connected to the control region of another species, resulting in the chimeric ca.7180000996827.

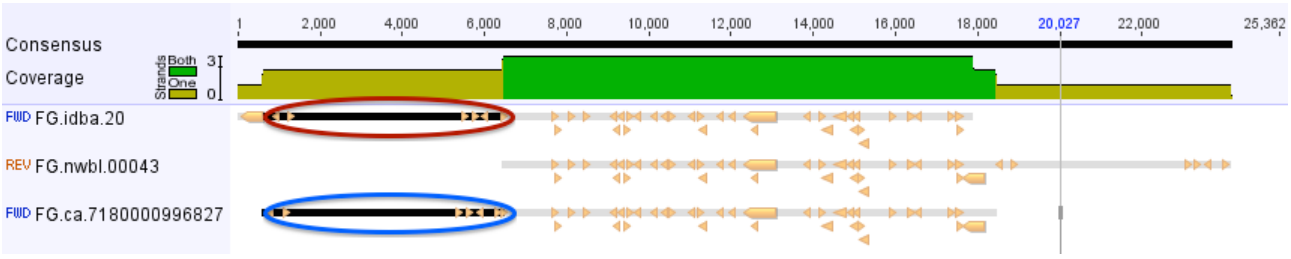


Figure 19: Dealing with long sequences Example 19

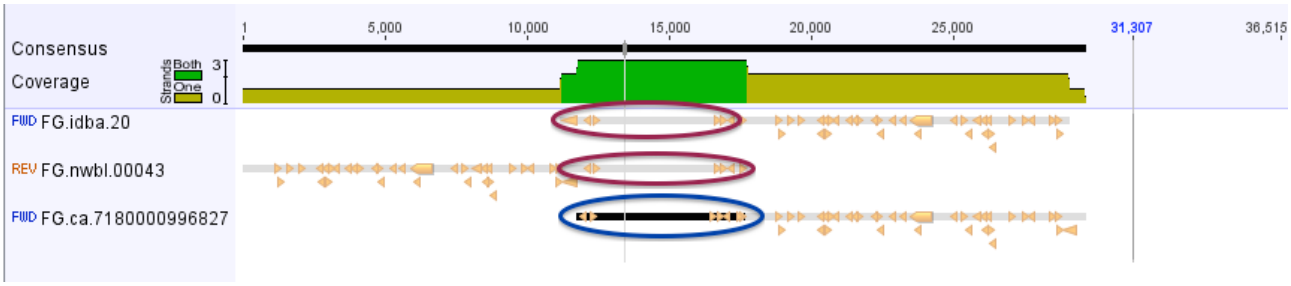


Figure 20: Dealing with long sequences Example 20

3.2 Circularising sequences

Circularising sequences is rather tricky. There are a few ways of doing this but we recommend using Geneious.

3.2.1 Using Python (or any other programming language).

Likewise, to make life easier as well as to tease out more overlaps, we used the programming software Python to help us find possible overlaps. Programming softwares can find overlapping areas that we may have missed as possible overlapping areas may have only one or two errors. The search function on Geneious only returns results that have a 100% match to the search sequence. However, about 1/4 of the time, it is possible to find matches such as the one below.

This particular supercontig had an overlapping area of 403 bps with only two mismatches at base positions 315 and 435. The supercontig overlaps at base positions 202-604 and 17332-17734 (Figure 21). It would not have been possible to find such a overlap using the search engine of Geneious alone. In fact, over half of our circular sequences were found using Python. Nonetheless, programming softwares are not easy to use, takes time to learn and one should not use it without knowing exactly what you are doing in case the wrong parameter was changed. They are also unforgiving on minor errors made.

```
Potential common AATAAATTTAATATTAAATTTACTATCCCTCAATAATAAATTACAAATAATGAAAATATTAAA
TTAAAAATTGAAATTTACTATCCCTCAATAATAAACCATAATAACTAAAAGATCCCTAATAAATTTAATATTAAATTTAC
TATCCCTCAATAATAAATTACAAATAATGAAAATATTAAATTAAAAATTGAAATTTACTATCCCTCAATAATAAACCATA
ATAACTAAAAGATCCCTAATAAATTTAATATTAAATTTACTATCCCTCAATAATAAATTACAAATAATGAAAATATTAAA
TTAAAAATTGAAATTTACTATCCCTCAATAATAAACCATAATAACTAAAAAATCCCTAATAAATTTAATATTAAATTTAC
TATCCCTCAATAATAAATTA at 202-604
401/403 at 17332-17734 differ at [315, 435]
```

Figure 21: Overlapping area found by Python.

Nonetheless, for those knowledgeable in programming, this is apparently a rather easy program to code for and can be done using any programming language. When using programming softwares, start from

20 bp overlaps. If these are not found, the chance of a overlap is near null. If found, increase number of bp overlaps until it cannot be found. Then, increase the number of mismatches allowed until none can be found any further to obtain the longest possible overlap. Then, double check this sequence with Geneious and decide whether it is an overlap and thereafter delete appropriately. Overlaps should be at least 300bps long with at most 3 errors.

Once overlapping areas have been found, you have to decide which overlapping area to delete. We would usually go for the area with the least contigs or with the most errors. Be very careful when deleting overlapping areas. In Example 22, delete starting from the first base of the overlap sequence at 2L and till the end of the supercontig.

On the other side, delete from the start of the supercontig to the first base before the overlapping sequence at 1L. If the keyboard Delete does not work, which happens at times, right-click and Delete Selected Bases. If the two sentences before seem gibberish, basically ensure that you delete only one overlapping area and not the other and not any other bases! Think before you start deleting and imagine the sequence going circular.

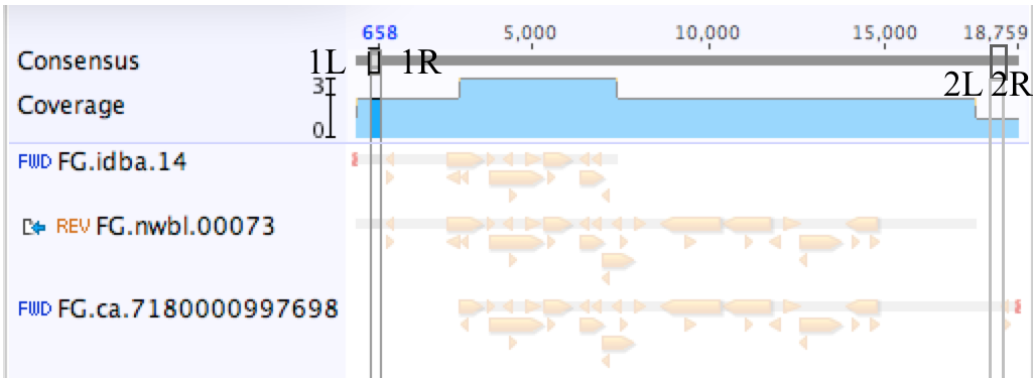


Figure 22: Deleting Overlaps Example 22

3.2.2 Using search function in Geneious.

First, you can try to find overlapping areas of at least 300 bps. Check the ends of the sequences for overlap by selecting about 300 bps at either end of the consensus sequence and press Ctrl+C, Ctrl+F, Ctrl+V and click Find Next. If no overlaps are found, try checking near the ends of the supercontigs. As ends of contigs are problematic, majority of overlaps has been found here.

3.2.3 Making use of other information in Geneious.

Otherwise, you can also make use of other available information. We recommend using this method to circularise sequences. Select any folder with a reference genome and Tick Annotate from... under the 'Live Annotate & Predict' Toolbar (Figure 34). This will show the rough location of all the genes (Figure 23).

As we prefer to start from ND2 gene, we would delete starting from the base before tRNA-Ile at the beginning of the supercontig (Figure 24) and from tRNA-lle onwards at the end of the supercontig. This will give us a complete circular sequence with all the required genes.

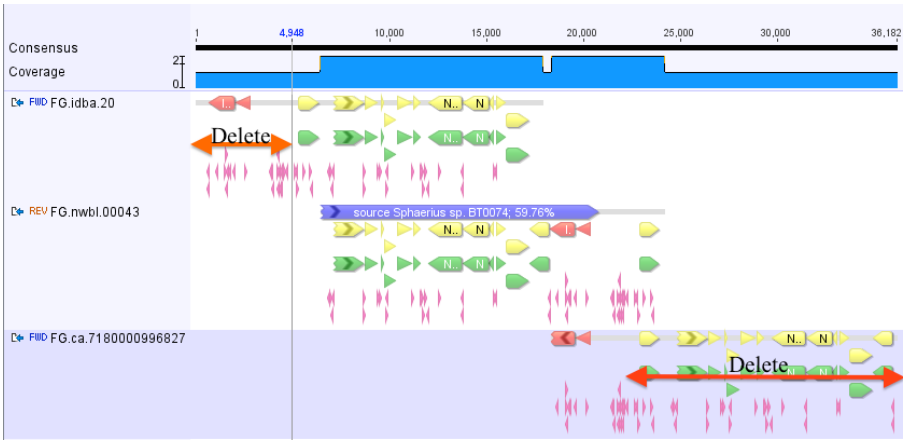


Figure 23: supercontig with annotations for reference

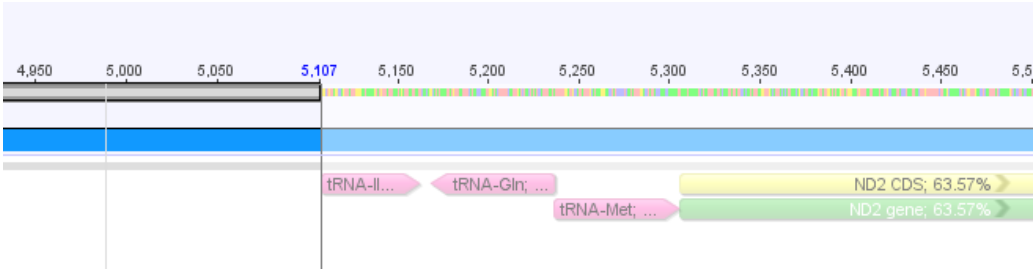


Figure 24: tRNA-Ile at beginning of supercontig

3.3 Keeping a datasheet

It is a good idea to keep a record of all the edits done on supercontigs. We recorded the initial and final number of contigs a supercontig has. We also recorded down the contigs that we deleted from each supercontig and the edits made on the supercontigs. Under edits, usual comments used were Trimmed R/L end off (contig.name) and Shifted R/L (contig.name). Number of final unresolved mismatches were recorded, if there were more than 7 mismatches, they were recorded as 100. supercontigs that have a reasonable overlap gives a 1 under able to circularise, and those unable a 0 (Figure 25).

	A	B	C	D	E	F	G	H	I	J	K
1	Contig	InitialReads	FinalSequences	NumberDeleted	ContigsDeleted	Edits	NumberOfTrims	NumberOfRTrims	NumberOfLTrims	MismatchNumber	Circularise
2	1	16	9	7	idba.15699, idb.3	Trimmed L off nwbl.0075, t	2	1	1	100	1
3	2	11	10	1	nwbl.00106	NA	0	0	0	2	0
4	3	10	4	6	idba.189, ca.7180	Trimmed R off nwbl.00038	3	2	1	0	1
5	4	9	5	4	nwbl.00592, idba.	Trimmed L off ca.7180000	4	3	1	100	1
6	5	9	7	2	nwbl.02412, nwbl	Trimmed R off ca.7180000	1	1	0	0	0
7	6	9	5	4	idba.551, idba.19	Trimmed RL off ca.718000	3	3	2	100	1
8	7	8	8	0	NA	Trimmed L off idba.175, tri	4	3	1	1	0
9	8	8	7	1	nwbl.00375	Trimmed L off ca.7180000	2	1	1	1	1
10	9	7	4	3	idba.11027, idba.	Trimmed R off idba.129, tri	2	2	1	1	1
11	10	7	4	3	nwbl.00310, nwbl	Trimmed R off ca.7180000	1	1	0	1	0
12	11	7	4	3	idba.1565, ca.718	NA	0	0	0	1	0
13	12	7	5	2	nwbl.07067, idba.	Trimmed L off nwbl.00195,	2	0	2	4	0
14	13	7	5	2	idba.5823, idba.2	Trimmed R off idba.110, tri	3	2	1	5	0
15	14	7	6	1	nwbl.02212	NA	0	0	0	1	0
16	15	7	5	2	idba.3601, nwbl.0	NA	0	0	0	0	0
17	16	6	5	1	idba.157	Trimmed L off idba.210	1	0	1	0	0
18	17	6	6	0	NA	Trimmed L off idba.102, tri	2	0	2	100	0
19	18	6	4	2	nwbl.00221, nwbl	Trimmed R off nwbl.00248	1	1	0	0	0
20	19	6	6	0	NA	Trimmed RL off ca.718000	5	2	4	0	1
21	20	6	5	1	nwbl.00666	Trimmed R off nwbl.00442	1	1	0	2	0
22	21	6	5	1	nwbl.00228	Trimmed R off ca.7180000	1	1	0	0	0
23	22	6	5	1	nwbl.07688		0	0	0	0	0

Figure 25: Top headings of datasheet for supercontigs

4 Annotating genomes

We begin the process with an un-annotated mitochondrion. Invertebrate mitochondria vary in size but are typically between 14,000 bps and 17,000 bps. Mitochondria of much lower length are likely to be just fragments whereas larger sequences could be chimeric or the result of large insertions (Figure 26). Only annotate genomes that are above 9000 bps.

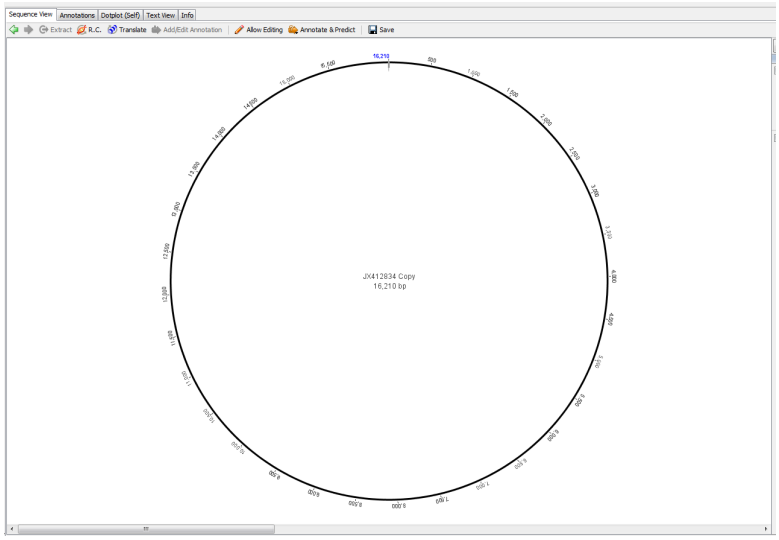


Figure 26: Un-annotated Mitochondrion

4.1 Finding a good reference mitochondria

Firstly we want to identify the region on the mitochondria which codes for the cytochrome c oxidase subunit 1 (COX1) as this is highly conserved. This means it is useful for determining relationships between Coleoptera taxa.

Click on the 'Live annotate and predict' pin (Figure 34) shown by the orange arrow to the right of the sequence in the quick tools menu. From here, it is possible to automatically look for similar sequences in mitochondria that have alcontigy been annotated and transfer them to our un-annotated trial sequence. There is a slider to determine the threshold similarity required for annotations to be shown as well as an option to select a specific folder in which to look for annotated mitochondria. The default is PlasMapper Features and should be changed to a user created folder containing a handful of well-annotated coleopteran mitochondrial genomes obtained from the NCBI database. In this case my folder is called 'Stock reference'.

Checking the box beside 'Find Annotations' will transpose the annotations onto the un-annotated sequence. COX1 is normally the second gene/CDS cluster moving clockwise from the red rRNA-annotations. By hovering your cursor over the annotations you can see their name, length and other useful information; such as translation exceptions and any user associated notes. Clicking on one of the COX1 annotations selects the base sequence of our trial sequence over the same region. This is shown in full colour whilst everything else becomes faded. Simply copy the selected region (Ctrl + C) (Figure 27).

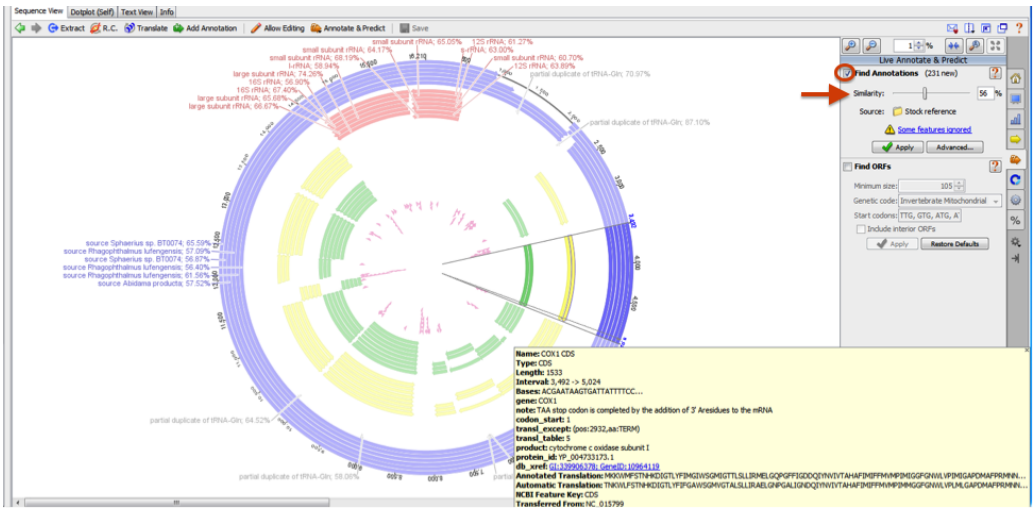


Figure 27: Selecting COX1

We shall now search this sequence within the NCBI database to find a good reference mitochondria with which to accurately transfer annotations that will require little editing afterwards. Go to the BLAST section of the NCBI website and choose 'nucleotide blast' from the given options. Paste the copied sequence (Ctrl+V) into the query box and click on the BLAST button at the bottom of the submission form (Figure 28).

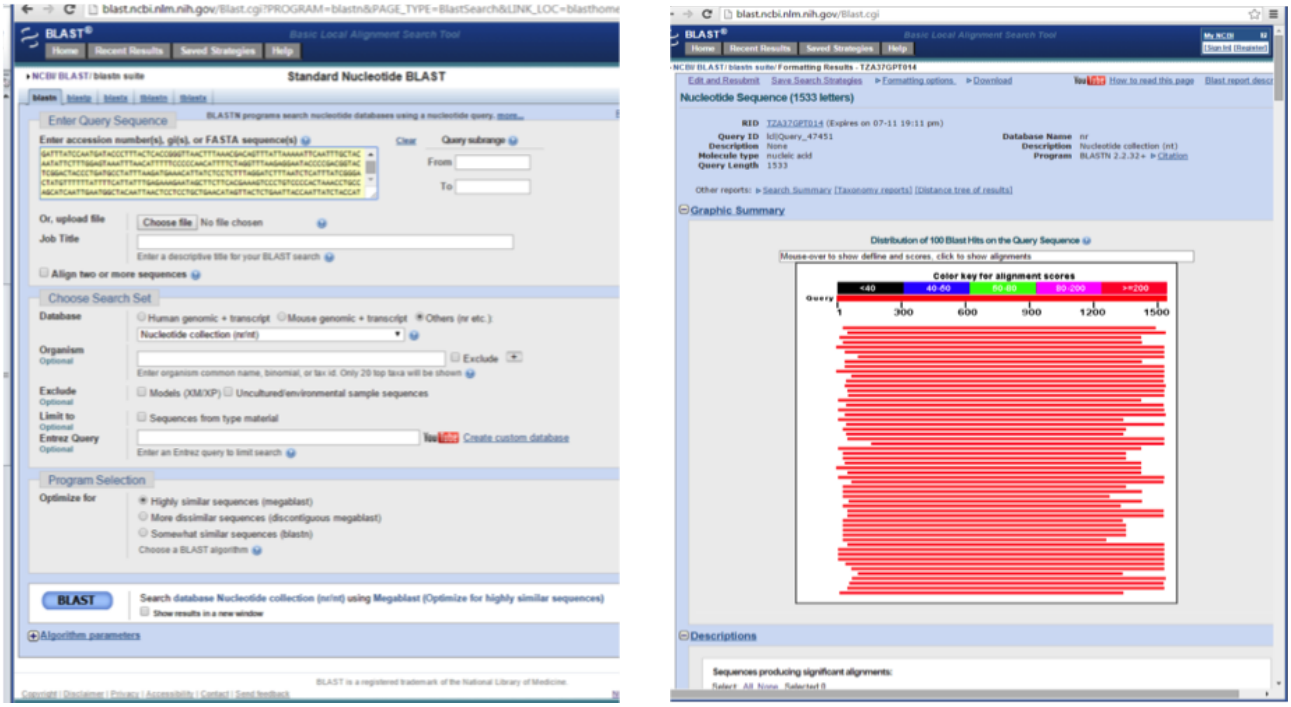
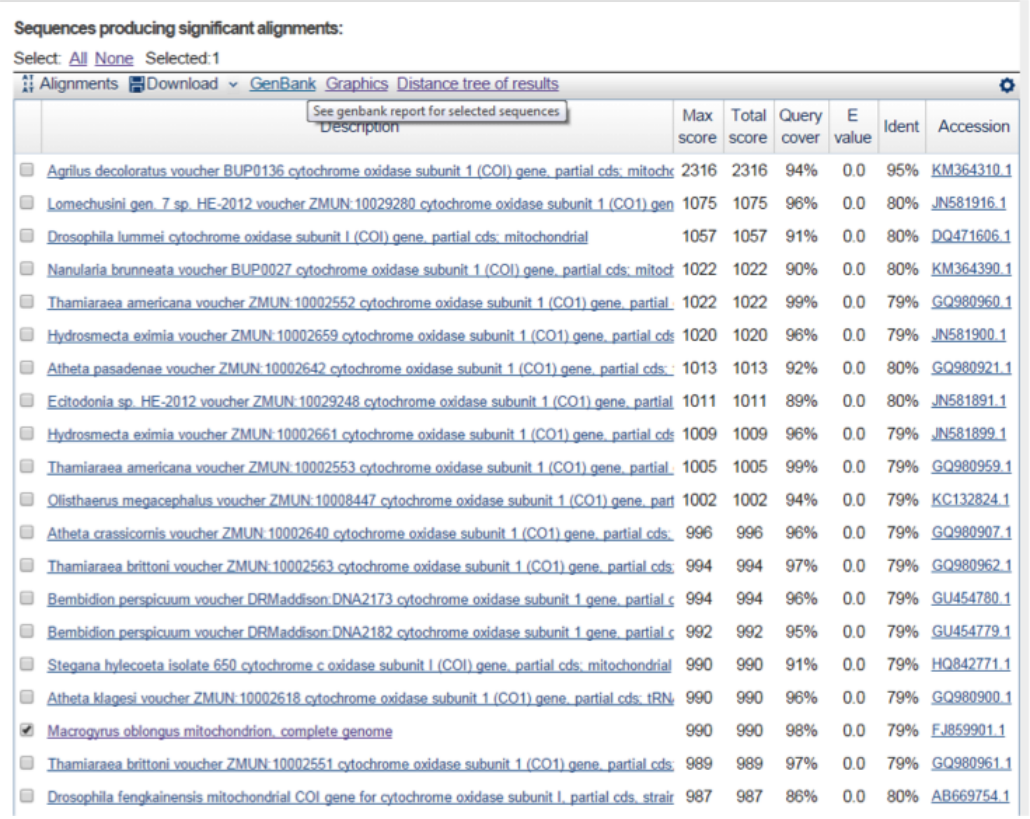


Figure 28: BLAST Search

The results are presented with similar database sequences aligned to our query sequence. You can click on each of the red bars to be taken directly to them or simply scroll down the page. The hits are given in order, with those with the highest score (most similar) at the top and least at the bottom (Figure 29). Many of the hits are single gene sequences, but we are interested in complete mitochondrial genomes so annotations for other genes can also be transferred.



Description	Max score	Total score	Query cover	E value	Ident	Accession
Agrilus decoloratus voucher BUP0136 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochc	2316	2316	94%	0.0	95%	KM364310.1
Lomechusini gen. 7 sp. HE-2012 voucher ZMUN:10029280 cytochrome oxidase subunit 1 (CO1) gen	1075	1075	96%	0.0	80%	JN581916.1
Drosophila lummei cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	1057	1057	91%	0.0	80%	DQ471606.1
Nanularia brunneata voucher BUP0027 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitoch	1022	1022	90%	0.0	80%	KM364390.1
Thamiaraea americana voucher ZMUN:10002552 cytochrome oxidase subunit 1 (CO1) gene, partial	1022	1022	99%	0.0	79%	GQ980960.1
Hydrosmecta eximia voucher ZMUN:10002659 cytochrome oxidase subunit 1 (CO1) gene, partial cds	1020	1020	96%	0.0	79%	JN581900.1
Atheta pasadenae voucher ZMUN:10002642 cytochrome oxidase subunit 1 (CO1) gene, partial cds;	1013	1013	92%	0.0	80%	GQ980921.1
Ecitodonia sp. HE-2012 voucher ZMUN:10029248 cytochrome oxidase subunit 1 (CO1) gene, partial	1011	1011	89%	0.0	80%	JN581891.1
Hydrosmecta eximia voucher ZMUN:10002661 cytochrome oxidase subunit 1 (CO1) gene, partial cds	1009	1009	96%	0.0	79%	JN581899.1
Thamiaraea americana voucher ZMUN:10002553 cytochrome oxidase subunit 1 (CO1) gene, partial	1005	1005	99%	0.0	79%	GQ980959.1
Ofisthaerus megacephalus voucher ZMUN:10008447 cytochrome oxidase subunit 1 (CO1) gene, part	1002	1002	94%	0.0	79%	KC132824.1
Atheta crassicornis voucher ZMUN:10002640 cytochrome oxidase subunit 1 (CO1) gene, partial cds;	996	996	96%	0.0	79%	GQ980907.1
Thamiaraea brittoni voucher ZMUN:10002563 cytochrome oxidase subunit 1 (CO1) gene, partial cds;	994	994	97%	0.0	79%	GQ980962.1
Bembidion perspicuum voucher DRMaddison:DNA2173 cytochrome oxidase subunit 1 gene, partial c	994	994	96%	0.0	79%	GU454780.1
Bembidion perspicuum voucher DRMaddison:DNA2182 cytochrome oxidase subunit 1 gene, partial c	992	992	95%	0.0	79%	GU454779.1
Stegana hylecoeta isolate 650 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial	990	990	91%	0.0	79%	HQ842771.1
Atheta klagesi voucher ZMUN:10002618 cytochrome oxidase subunit 1 (CO1) gene, partial cds; tRN	990	990	96%	0.0	79%	GQ980900.1
Macrogyrus oblongus mitochondrion, complete genome	990	990	98%	0.0	79%	FJ859901.1
Thamiaraea brittoni voucher ZMUN:10002551 cytochrome oxidase subunit 1 (CO1) gene, partial cds;	989	989	97%	0.0	79%	GQ980961.1
Drosophila fengkainensis mitochondrial COI gene for cytochrome oxidase subunit I, partial cds, strair	987	987	86%	0.0	80%	AB669754.1

Figure 29: BLAST Results

Scroll down the list until the description is not gene specific but gives a Latin name, followed by 'mitochondrion' and the tag 'complete genome' i.e *Macrogyrus oblongus* mitochondrion, complete genome. Select it using the checkbox and click on the link to GenBank at the top of the hit list. N.B. If there are multiple complete genomes with similar scores you can check their boxes and follow the GenBank link as normal.

If there are no related complete genomes found, record the superfamily of the first hit and record the supercontig as NG. If it is not a Coleoptera, record it as NA (Figure 39).

The GenBank format supplies a lot of information about the sequenced mitochondrion. Importantly, note down the ACCESSION which will be used to import the annotated mitochondrial genome into Geneious later i.e. FJ859901. You can see all of the annotated features are also given here such as genes, CDS and tRNAs - the base sequence is also given at the bottom of the page.

Click on the hyperlinked organism name to be taken to the Taxonomy Browser of the NCBI (Figure 30). Here the taxonomical lineage of the species is given all the way from super-kingdom to genus. For the purposes of this lab we are interested in the super-family and family levels. Record down the superfamily name of the reference genome. Note that the latin superfamily names have the suffix -dea and the families have the suffix -dae.

Figure 30: NCBI Taxonomy Browser

To find related species with annotated mitochondria click on one of the taxonomical rankings i.e. the super-family *Gyrinoidea*, select the Genome category at the top of the page and click Go (Figure 31). All the taxa with sequenced mitochondria will then be flagged which is useful for constructing a robust database of annotated reference genomes to speed up the annotation process for un-annotated sequences later.

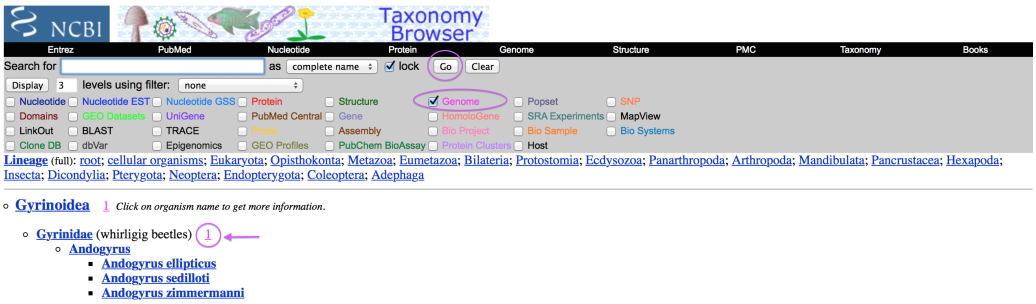


Figure 31: Fully annotated mitochondrial reference genome

Having determined the species lineage using the taxonomy browser, we can return to Geneious and import the annotated mitochondria. Select the NCBI logo in the sources pane on the lefthand side of Geneious and search using the ACCESSION key from GenBank that we acquired earlier. The fully annotated mitochondrial reference genome can now be seen (Figure 32).

Figure 32: Fully annotated mitochondrial reference genome

4.2 Transferring Annotations

Drag and drop (or alternatively copy and paste) a copy of this file to a new folder, called 'Trial' here, and place another copy into the simple taxonomic reference tree you are building using information from the NCBI taxonomy browser. For example, I have placed the annotated '*Macrogyrus oblongus* mitochondria' into a newly created folder to match its family classification of *Gyrinidae* (Figure 34). This will enable us to find it quickly in the future if we need to re-use this reference and will ultimately provide a reference database which can be given to other scientists working on similar projects.

Select the original un-annotated genome that we started with and use the tools on the right to find annotations (can be quickly mapped using Ctrl+Shift+A). This time select the 'Trial' folder as the source which contains only the reference mitochondrial genome found using the NCBI BLAST. Check that all 13 gene annotations are present and in the direction as seen in this example (Figure 33). If not all 13 gene annotations are present, reduce the Similarity percentage. However, when doing so, some smaller genes, such as ATP8 may appear at other places. Take note to delete these incorrect gene annotations after applying the annotations. Refer to Section 4.4.1 if not all 13 gene annotations can be found.

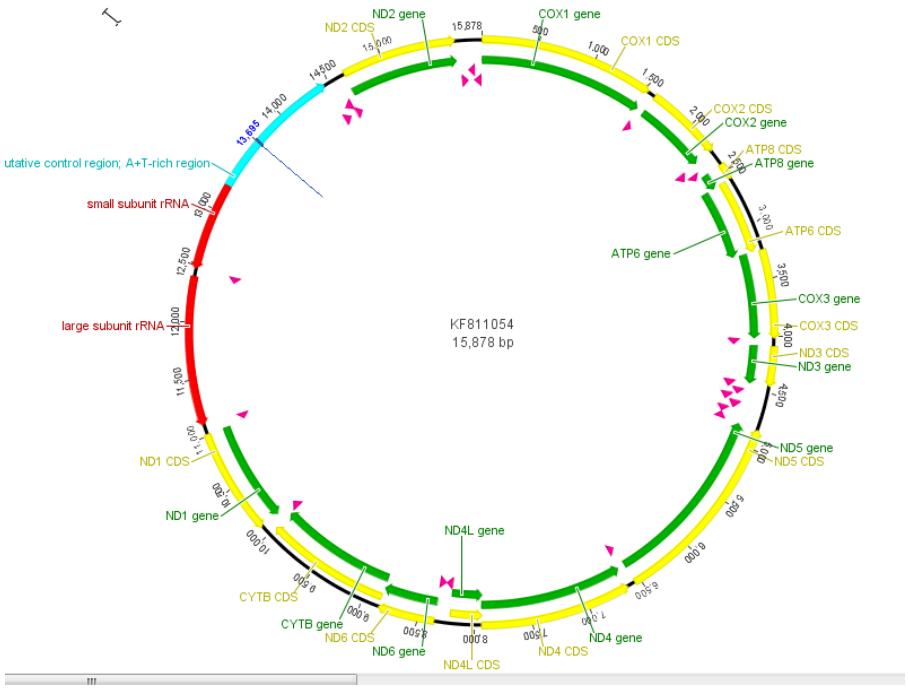


Figure 33: Example of a nicely annotated mitochondrial genome

Click the 'Apply' button to transfer these annotations to the un-annotated sample mitochondrion, at which point their colours will become solid rather than faded. You can now unselect the Find Annotations tick box and the annotations will remain.

On the same 'Live Annotate & Predict' toolbar (Figure 34), tick the 'Find ORFs' section to apply open contiging frame annotations to the sample sequence which allows us to align annotations correctly. **Make sure to change the genetic code from Standard (transl_table 1) to Invertebrate Mitochondrial (transl_table 5).**

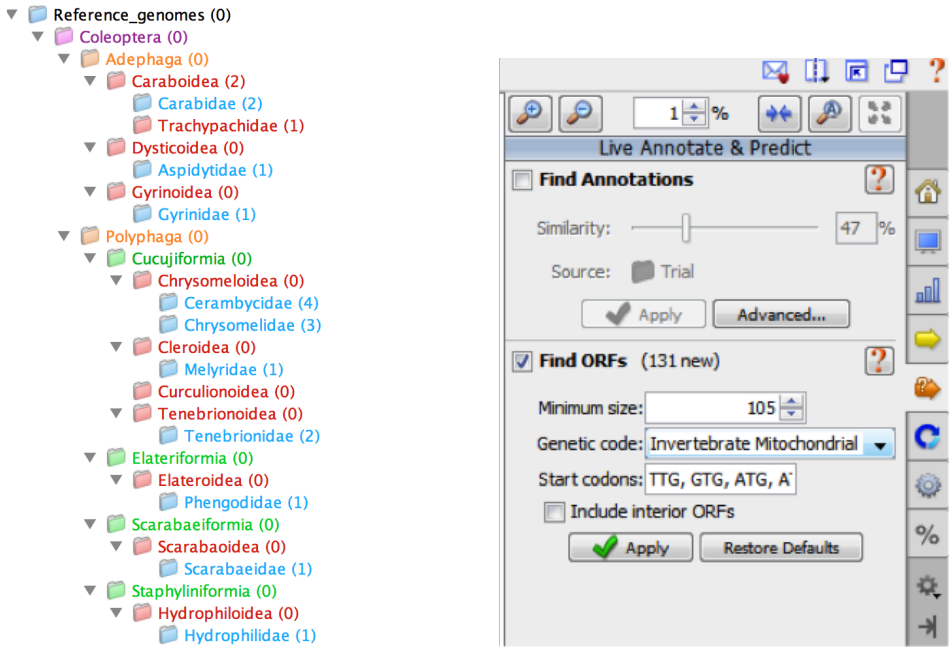


Figure 34: Folders and the 'Live Annotate & Predict' toolbar

Our previously un-annotated sample mitochondrion now looks very busy. To make alignment clearer we can choose which annotations we want to be displayed. Access the 'Annotation and Tracks' tool set by clicking on the yellow arrow above the 'Live Annotate & Predict' orange arrow pin.

For our purposes we only want to see the gene, CDS and ORF options so unselect the other annotation types. We now have a roughly annotated mitochondrial sequence showing the important information (Figure 35).

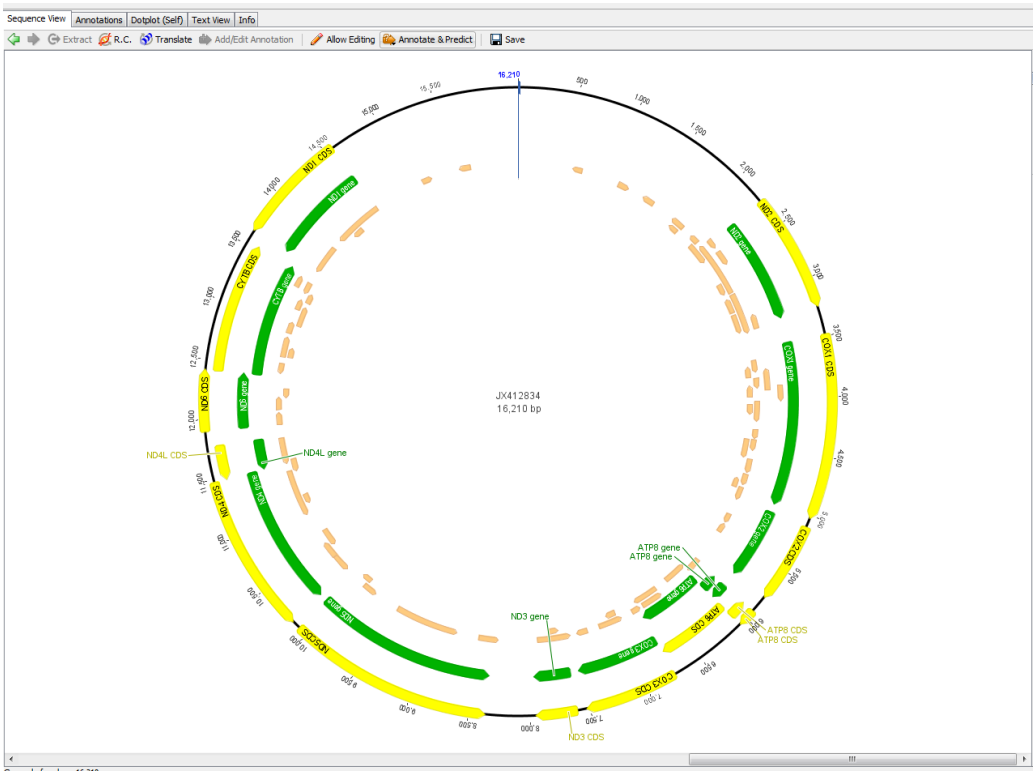


Figure 35: Roughly annotated mitochondrial sequence

4.3 Edit Annotations

Go to Annotations, select all and click Edit Annotations (Figure 38). Select transl_except and click Remove. Select note and click Remove. These are the translation exceptions from the previous references, which may not be applicable to our current supercontig.

The next step is to consider whether the annotations have been transferred accurately and that their position on the new mitochondrion makes biological sense i.e. within an ORF, starting with a methionine etc. In the 'Display' tool set on the right, tick the Translation option and alter the frame to 'By selection or annotation' and **set the genetic code to Invertebrate Mitochondrial: (transl_table 5)**. This adds a track which allows us to see what amino acids any selected sequence codes for, particularly useful for identifying start and stop codons. At this point it is also advisable to change the sequence view from circular to linear using the sequence tab at the top of the Geneious window or the mapped keyboard shortcut (Ctrl+Shift+C).

Click on the first gene ND2 CDS and zoom in (Zoom to selection: Ctrl+Shift+M). The imported annotation should start with a methionine. Edit each of the gene annotations such that they should start on a methionine (ATG, ATA, ATT, TTG, ATC) and end on a stop codon (TAA, TAG). Special cases for stop codons may occur (Refer to Section 4.4.3).

4.4 Case Studies

4.4.1 Missing Gene Annotations

In cases where even with 25% similarity, certain genes are missing, do the following steps. Create a folder with the missing gene name. For example, COX1. In many cases, despite blasting the COX1 gene to obtain the reference genome, the COX1 gene refuses to be transferred. We do not know why this is so. Nonetheless, click on the gene CDS from the reference genome and left click to Extract region. This will extract the COX1 gene into a new file. Transfer the new file into the COX1 folder. At the same time, extract COX1 gene regions from a multiple of other reference genomes that are found in the Reference genomes folder. Preferably, choose those that are closest in phylogeny to the Reference genome of the supercontig.

After transferring a number of COX1 genes into the COX1 folder, go back to the 'Live Annotate & Predict' Toolbar (Figure 34) and click Find Annotations, choosing the COX1 folder as the source. Out of the numerous genes extracted, usually one would be able to be transferred. If none have a high enough similarity, extract a few more COX1 genes from other reference genomes and repeat the process. Choose the one with the highest similarity and click Apply. You can do the same for other missing genes, using one folder in the Stock References for each gene.

4.4.2 Missing Start Codon

Sometimes, the ORF indicates that the gene starts at another methionine and not the one indicated by the transferred gene annotation. The gene annotation may also not start at a methionine. In such cases, where you don't know whether to extend the COX3 gene, run a translated blast (blastx) of the longer candidate sequence (Genetic code = Invertebrate Mitochondrial (5), within the non-redundant protein sequnces database in NCBI) and then look at the length of the hits. Also you can look at the hit flanking sequences and compare this directly to the sequence in Geneious i.e. WGS highly con-

served through the hits. This can then help you decide whether to extend or reduce the gene annotation.

Likewise, you can click on the longer candidate sequence and left click, selecting Translate... and click OK. Select the whole protein sequence and copy into blastp on NCBI website and run. If search times are becoming ridiculously long, under Organism in the BLAST options, type Coleoptera and select that taxa to make search timings much shorter. If no suitable start codon can be found, ensure that the starting amino acid is as similar in terms of properties to methionine as possible.

4.4.3 Missing Stop Codon

Similarly, if no Stop Codons are found, use the same method as that in Start Codons and blast the longer candidate sequence. Coleoptera genes can also end with a single T or TA. In such cases, double click the CDS which will produce a popup window (Figure 36). Ensure that you have deleted any transl_except and notes transferred over from the reference genome.

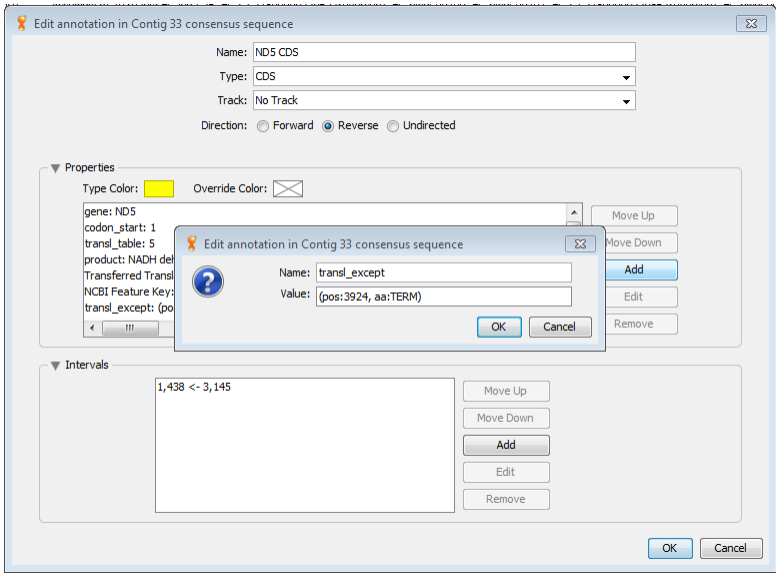


Figure 36: Adding Annotations

Click Add. Type the following for single T stop codons:

Name: **transl_except**

Value: **(pos:base position of T, aa:TERM)**

For example: Value: **(pos:3924, aa:TERM)**

If exception is on the reverse strand: Value: **(pos:complement(3924), aa:TERM)**

For TA stop codons:

Name: **transl_except**

Value: **(pos:[smaller base position]..[larger base position], aa:TERM)**

For example: Value: **(pos:3924..3925, aa:TERM)**

If exception is on the reverse strand: Value: **(pos:complement(3924..3925), aa:TERM)**

*When submitting to GenBank, we have changed all TA stop codons to T stop codons as it was giving us a submission error of 'Missing Stop Codon'. We have no idea why ☹.

Click OK. Add another annotation by clicking Add and type the following:

Name: **note**

Value: **TAA stop codon completed by the addition of 3'A residues to the mRNA**

If at anytime you are unsure of anything, select the sequence and click Add annotation. Key in the problem in the Name, for exampe, ND3 no stop, and click OK. This allows you to come back to be problem easily and also to delete all annotations of the misc_feature type after clarifying all problems (Figure 37).

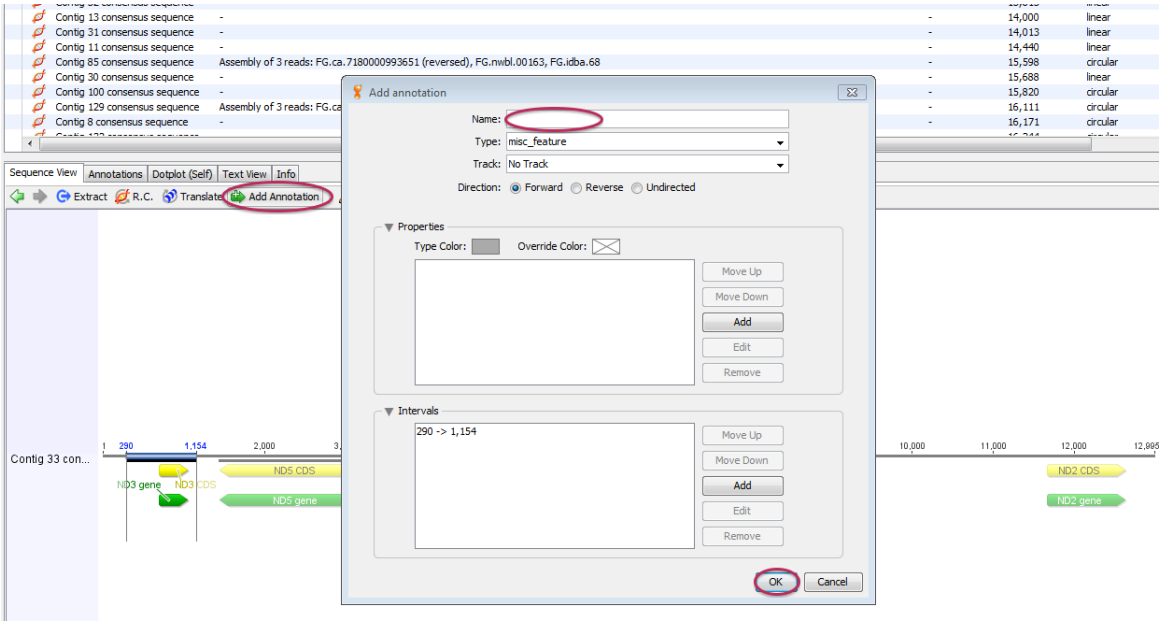


Figure 37: Adding misc_feature

4.5 Removing Annotations

Select all supercontigs and click Annotations. Select all annotations and click on Edit Annotation which is beside a green arrow (Figure 38). A pop-up window will appear. Select and click Remove to delete protein_id, db_xref, Transferred from, Transferred similarity, Translation. Also remove all Annotations that has Type: Source, misc_feature or tRNA. Delete also other random things that only appear for one supercontig and not the rest, for example Transferred Translation.

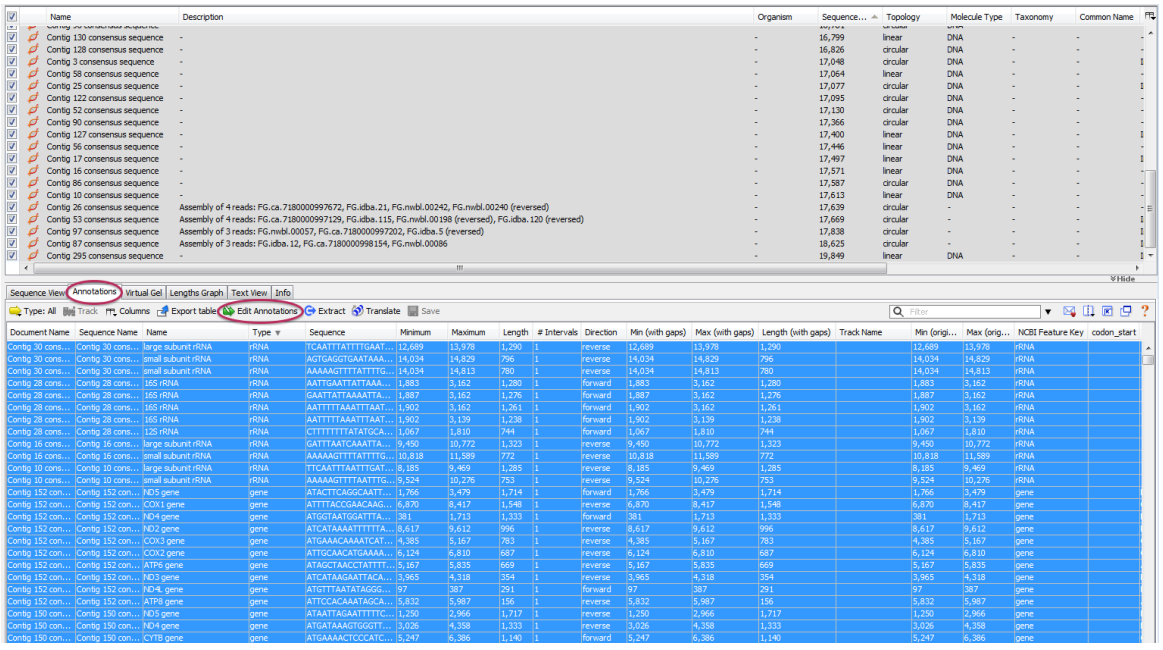


Figure 38: Removing Annotations

4.6 Keeping a datasheet

Keeping a datasheet of the annotation process will give us further insight into the genes. BlastGene referred to the gene from the supercontig that we used for Blast. ReferenceGenome refers to the genome from which the annotations were transferred. Superfamily referred to the superfamily of the ReferenceGenome (Figure 39).

Number of bases we extended or reduced the transferred annotations by were recorded under GeneBP by a positive number and a negative number respectively and the stop and start codons were recorded as well. An '!' was used when we could not find a start or stop codon and NA was used when the sequence is too short and hence did not have the indicated genes (Figure 40).

	A	B	F	G	H	I	J	K	L	M	N
1	ContigTitle	Finished	SequenceLength	AbleToCircularise	BlastGene	ReferenceAccession	Superfamily	COX1Recovered	COX1 recoverred using	LastGeneRecoverred	LastGeneSimilarity
2	1	0	16548	1	COX1	KP455510	Curculionoidea	1	NA	ND6	53.43
3	2	1	13674	0	COX1	KP455482	Curculionoidea	1	NA	ND6	57
4	3	0	17048	1	COX1	AM493668	Dytiscoidea	1	NA	ND6	54.8
5	4	1	16761	1	COX1	KM676219	Culicoidea	0	NC015799	ATP8	72.84
6	5	NA	16953	0	COX1	KJ947872	Ephydroidea				
7	6	1	16379	1	COX1	KP410324	Curculionoidea	0	NC015799	ATP8	59.39
8	7	NA	18084	0	COX1	DQ029097	Muscoidea				
9	8	0	16171	1	COX1	KJ947872	Ephydroidea	0	NC015799	ND2	56.07
10	9	NA	17831	1	COX1	AY518673	Ephydroidea				
11	10	0	17613	0	COX1	EU877950	Sphaeriusidae	0	NC015799	ND6	64.18
12	11	1	14440	0	COX1	FJ859903	Scarabaeoidea	1	NA	ATP8	53.41
13	12	NA	14251	0	COX1	KM200724	Muscoidea				
14	13	1	14000	0	COX1	EU877950	Sphaeriusidae	1	NA	ND2	58.59
15	14	NG	11262	0	COX1		Staphylinoidea				

Figure 39: Recording Annotation Edits

	A	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	ContigTitle	ND2StartBP	ND2StartCodon	ND2StopBP	ND2StopCodon	COX1StartBP	COX1StartCodon	COX1StopBP	COX1StopCodon	COX2StartBP	COX2StartCodon	COX2StopBP	COX2StopCodon
2	1	0	ATT	-11	TAA	0	ATT	8	TAA	0	ATT	0	TAA
3	2	0	ATC	-9	TAA	9	ATT	8	TAA	0	ATT	0	TAA
4	3	-4	ATA	-13	TAA	3	ATC	0	TAA	0	ATG	0	T
5	4	0	ATT	0	TAA	0	TTG	8	TAA	0	ATG	17	TAG
6	5												
7	6	0	ATT	0	TAA	6	ATT	8	TAA	0	ATC	18	TAA
8	7												
9	8	21	ATA	-4	TAA	6	ATT	5	TAA	0	ATG	0	T
10	9												
11	10	-3	ATT	0	TAA	6	ATT	5	ETA	0	ATG	3	T
12	11	NA				9	ATC	5	ETA	0	ATT	17	TAA
13	12												
14	13	1	ATA	4	TAA	9	ATC	5	ETA	0	ATG	31	TAA
15	14												
16	15												
17	16	-4	ATA	-12	TAA	6	ATT	8	TAA	0	ATT	0	TAA
18	17	-7	ATA	6	TAG	!	TTT	0	TAA	0	ATG	3	T

Figure 40: Recording Annotation Edits

Using these data, we could also create diagrams in R to show common trends (Figure 41, Figure 42, Figure 43, Figure 44).

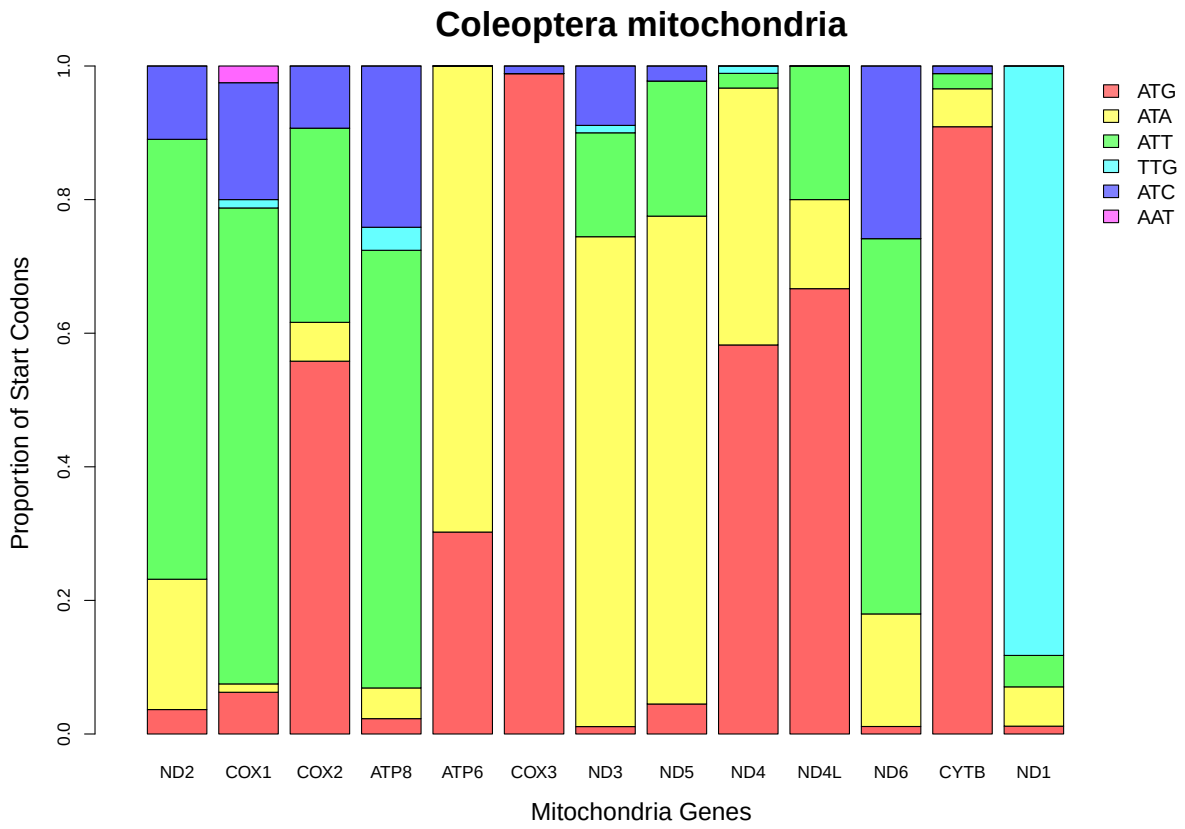


Figure 41: Start codons of mitochondrial genes.

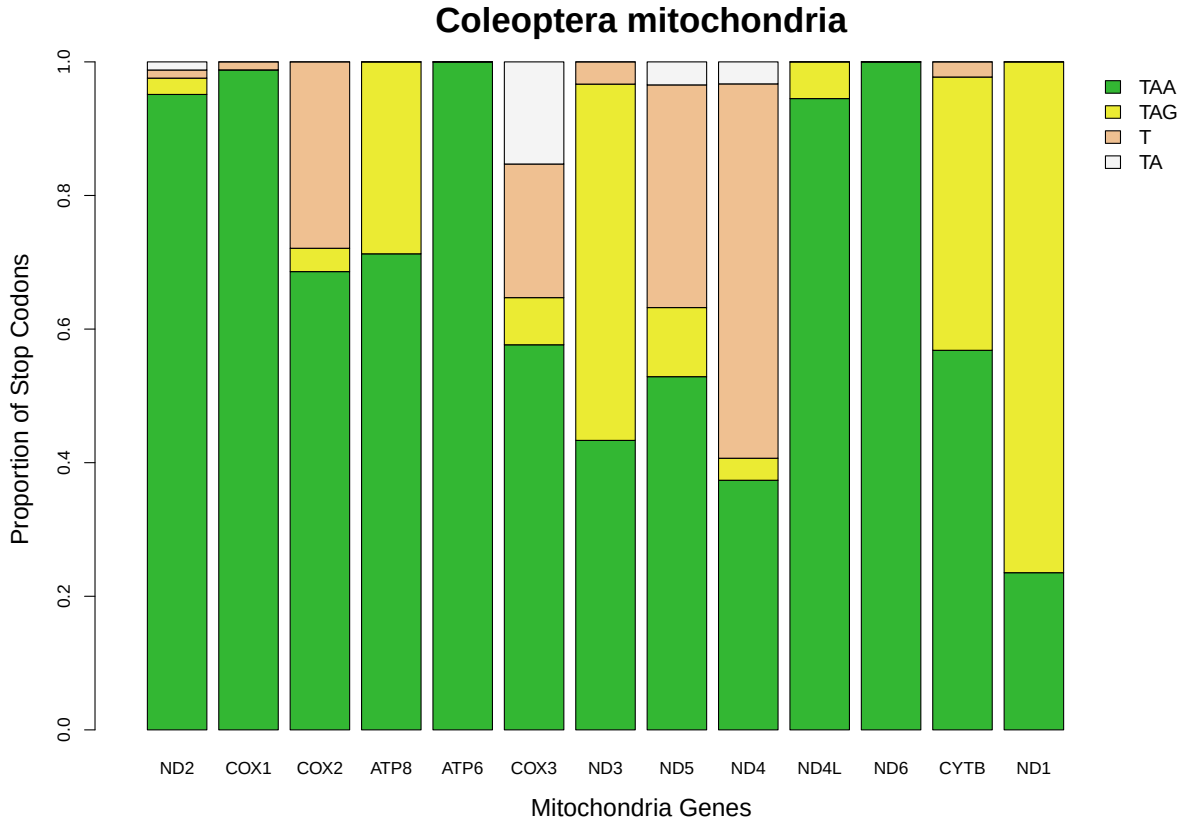


Figure 42: Stop codons of mitochondrial genes.

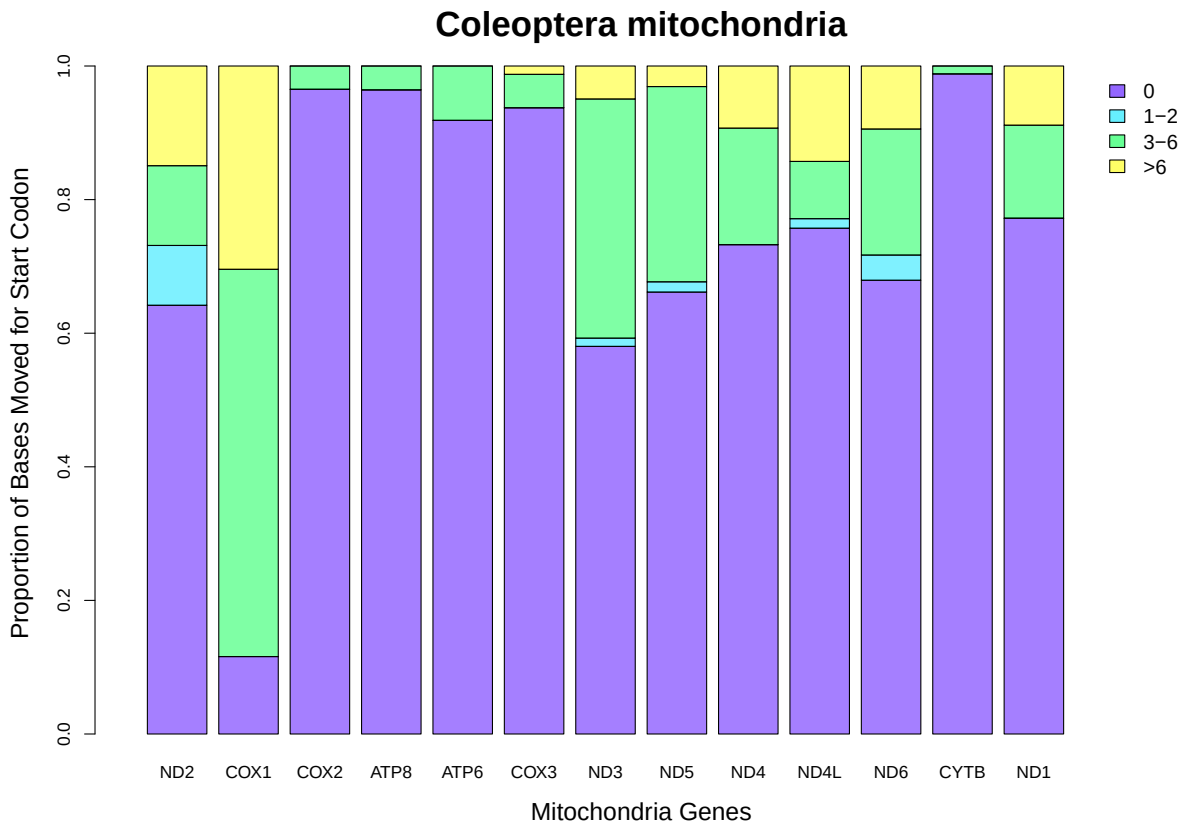


Figure 43: Number of bases moved in editing start codons of mitochondrial genes.

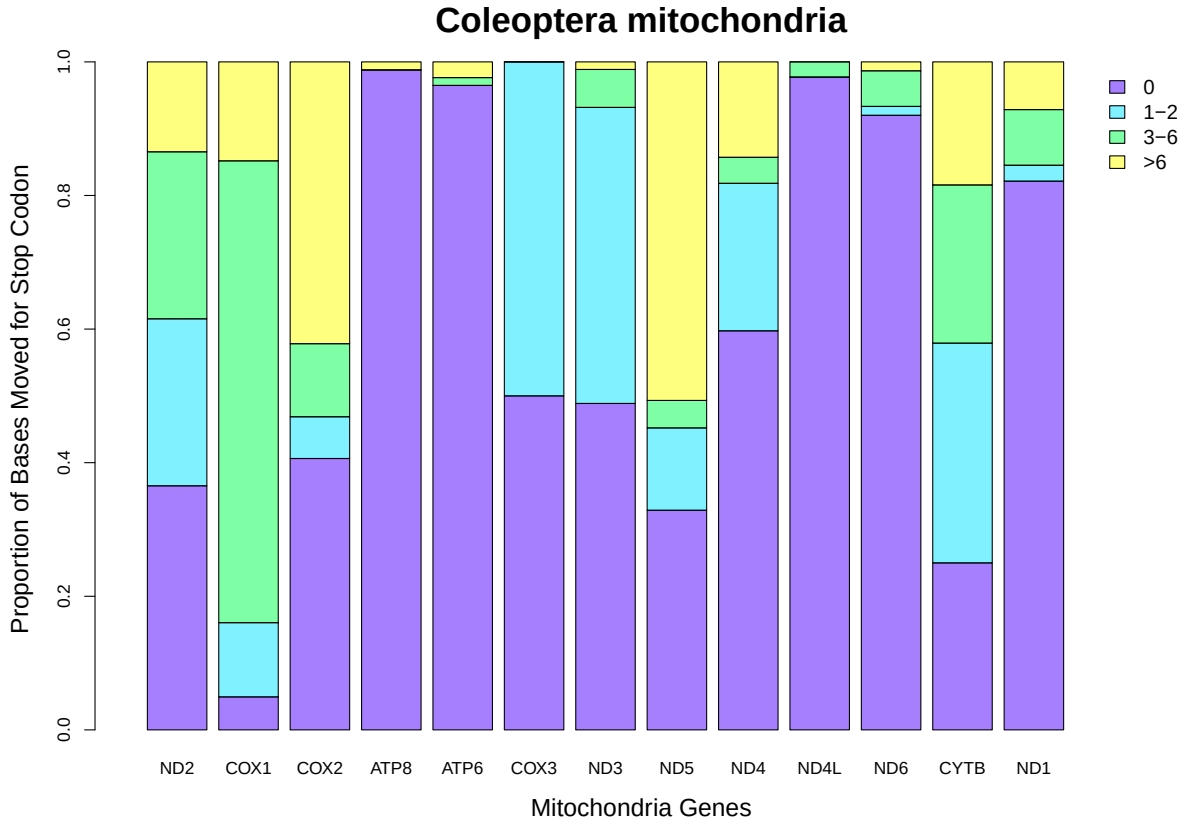


Figure 44: Number of bases moved in editing stop codons of mitochondrial genes.

5 Adding tRNAs

To add tRNA annotations, first delete all existing tRNA annotations if you have not already done so.

Refer to 'My time-saving COREALLA' guide. For Windows, download Filezilla portable and Putty. For Mac users, download Cyberduck. Filezilla and Cyberduck allow you to transfer files to and from the server easily. After obtaining your username and password from Benjamin, log on to ctag. Transfer the startup kit from ben's directory into your directory.

Export all the consensus sequences into one fasta file and upload it into the tRNA folder using Cyberduck or Filezilla. Set your directory into the tRNA stuff folder by using the cd command which should look somewhat like this:

cd starter\kit/tRNA\stuff/

Run the following commands:

perl all2many.pl [filename] 1000

sh covel_wrap_multithread.sh

*You have to allocate memory (15 or 20)

perl covels_wrap.pl.

*This script will take a long time to run. To run it overnight, refer to the 'My time-saving COREALLA'.

perl cove_output20130710.pl

*This will generate a folder called GB and in it is the .gb file which you can transfer back into Geneious.

Select all of your original supercontigs and the ones with tRNAs on it. Go to Tools, Align/Assemble, De Novo Assemble... If your files all start with the same format, Tick Assembly by 1st or 2nd part of name to make the process easier and leave other options as default. It is recommended to assemble by name.

If the names are separated by different things, for example the one by tRNA is by _ and the original supercontigs by a space, this will not work. If you do not assemble by name, Geneious will reorder the supercontigs and it becomes hard to track which supercontig comes from where. Otherwise, you can chnage the supercontig names manually after assembly.

There should be no unused reads. If there are, something somewhere has gone wrong. Recheck your sequences and tRNA annotations.

After assembly is done, place the Assembly report and Consensus Sequences into another folder. Select all the assembled new supercontigs and go to Annotate & Predict at the top menu bar, Transfer Annotations and follow through. As the sequences are all identical, you do not have to do it manually. Thereafter, you should see that the consensus sequence contains both the annotations that you have manually edited as well as the new tRNA annotations (Figure 45).

Check through all the consensus sequences and ensure that they all have the transferred annotations. Remove any generated unneeded annotations (Refer to subsection 4.5).

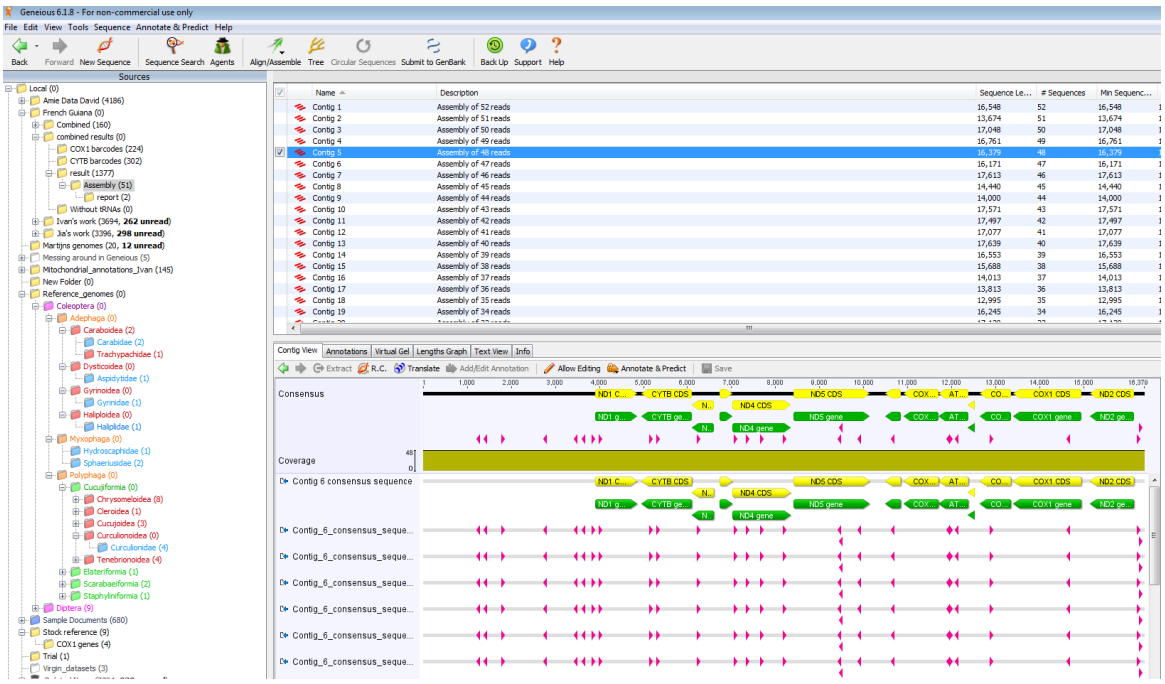


Figure 45: Transferring tRNA annotations.

A complete circular mitochondrion should ideally have 22 tRNAs as beautifully demonstrated in Example 46.

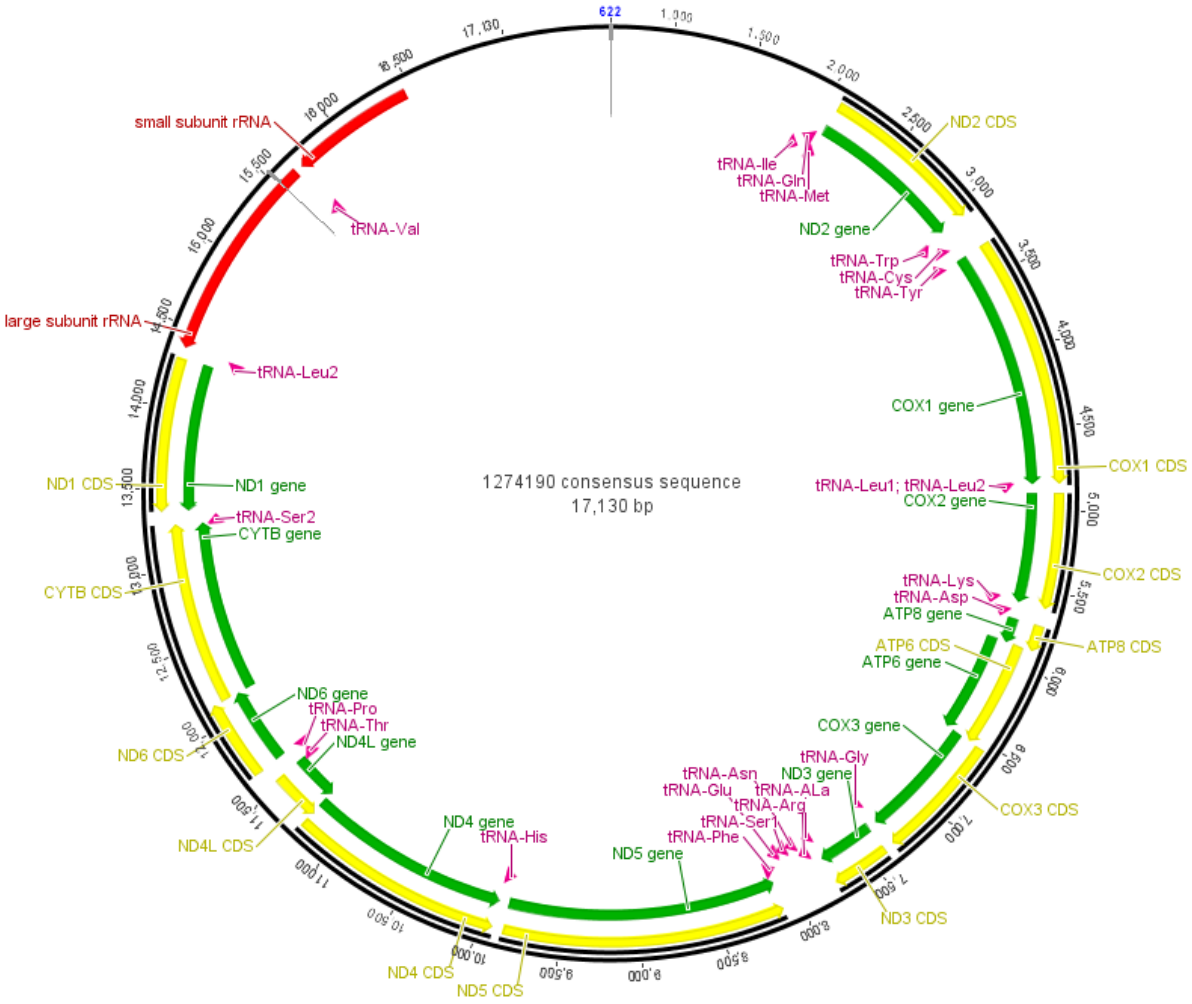


Figure 46: Consensus sequence with all 22 tRNA annotations.

If any tRNA or rRNA annotations are missing, create a folder called tRNA annotations. Create a copy of one of the reference genomes with 22 tRNA annotations downloaded. Click on Annotations of the copy of reference genome. Select all non-tRNA and non-rRNA annotations and press Delete. You should be left with all tRNA and rRNA annotations as seen in Example 47. This step is similar to that in Section 4.5.

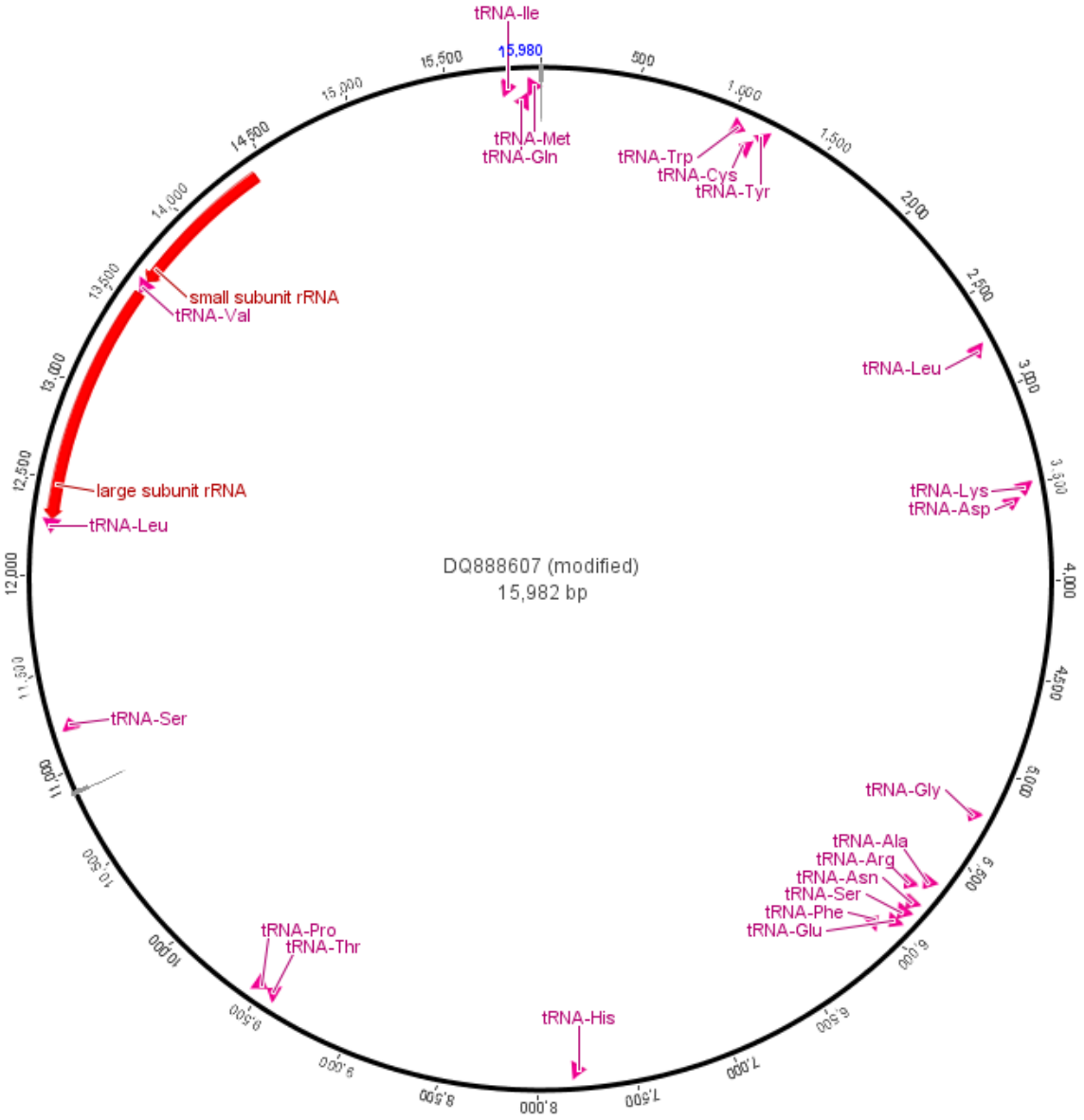


Figure 47: Reference genome with only tRNA and rRNA annotations.

Place the file in the tRNA annotations folder. Go back to your consensus sequences with missing tRNAs. Select one of the consensus sequences. Click on Find Annotations in the 'Live Annotate & Predict' toolbar (Figure 34) but choose your tRNA annotations folder under Source. Select a high similarity of around 70%. Ensure that all tRNA or rRNA annotations that are missing are present in the right location by comparing it to Example 46. Click Apply. Remove any duplicated tRNA or rRNA annotations that have been transferred over. Your Sequence View should now look similar to Example 46 and have 22 tRNAs. This step is similar to that in Section 4.2.

You can now prepare your files for submission to GenBank.

6 Submitting to GenBank

After barcoding, the process is not yet over. There are a few ways of submitting sequences. Geneious is one of the methods. However, it provides little flexibility and is a rather tedious process. The metadata of the project, such as the organism name and taxa, has to be keyed in for every consensus sequence individually. The other method that we have chosen to use is through tbl2asn.

6.1 Submitting using Geneious.

Go to Tools, Submit to GenBank in Geneious to reach the window below (Figure 48).

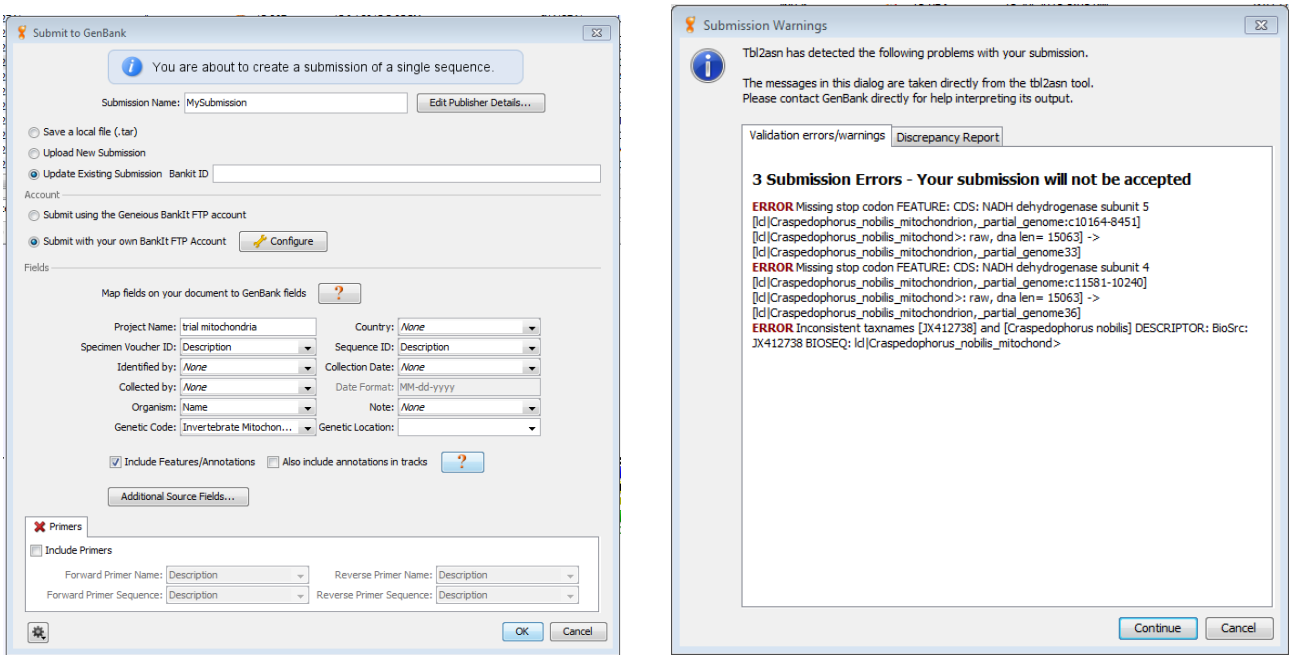


Figure 48: Submitting to Genbank in Geneious

Fill in the appropriate information. Select remove when asked if it is alright to remove unneccessary annotations. Usually, if not always, errors will appear (Figure 48). Below are some of the errors and methods of removing them (Table 1).

Table 1: Errors and solutions to Submit to Genbank in Geneious.

Errors	Solutions
'given protein length does not match translation length'	Delete annotation of type 'translation'
'code-break location not in coding region'	Delete incorrect annotations with feature 'transl_except'
'anti-codon location.../unparsed anti-codon'	Delete annotation of type 'anticodon'
'Missing stop codon'	Change transl_except of TA to only T For ex. (pos:9559..9560, aa:TERM) to (pos:9559, aa:TERM).
'Inconsistent taxnames'	Change name of file to organism name such that the description is the same as name.

6.2 Submitting using tbl2asn.

In our internship, we decided to use tbl2asn instead, as it allows us to submit many sequences at once. You require three files for tbl2asn: the template file, fasta file and the feature table file. For more information about tbl2asn and the three files needed to generate a .sqn file, go to http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/. A summary of steps is found below (Refer to PDF diagram).

6.2.1 Generating the template file.

You require the following information for the template file: the project title, sequence authors and publication status. Go to http://www.ncbi.nlm.nih.gov/WebSub/template.cgi and key in the required details and click Create template. This should generate the template file with suffix .sbt. Take note that special characters such as Ä and ß are unfortunately not accepted in author names.

6.2.2 Generating the fasta file.

Ensure that the name of sequences is BMNH number_(anything you like ☺). For ex. 1234567_consensus_sequence. Choose all sequences to be submitted and go to File, Export... in Geneious. Select FASTA(*.fasta) file type before saving the file.

Check that the title of every sequence in the fasta file shows >BMNH number_(anything you like ☺) and that the BMNH number corresponds with the title in the feature table file: >Feature_BMNH number. For example, >1234567_consensus_sequence in the fasta file should correspond with >Feature_1234567 in the feature table file. If it does not match, the script to generate the complete fasta file will not work. Ensure that title is separated by a _ and not by spaces or any other character. If they are, simply do a search for the character and replace with '_' in the fasta file before running it with Benjamin's Script.

Before running Benjamin's Script, compile a excel sheet with the following information: BMNH, the species name, the topology of the sequence(circular or linear), country and isolation_source(method of collecting organism) (Table 2). Save the file as a Comma Separated Values(*.csv) file. It is crucial that under organism: Family name, sp., BMNH and number are separated by exactly one space.

Table 2: CSV file to generate Fasta file.

BMNH	organism	topology	country	isolation_source
1234567	Staphylinidae sp. BMNH 1234567	circular	French Guiana	light trap

Benjamin's script assumes that the following is true:

environmental_sample = TRUE
location = mitochondrion
tech = wgs
mgcode = 5
metagenomic = TRUE

Here are some of the problems that Benjamin realised while running the script:

if topology=circular, you need to add in completeness=complete.
specimen-voucher=BMNH(space, not colon)number
if environmental_sample=TRUE, you need to add in isolation_source.

With the CSV file and the Fasta file, you can now run Benjamin's Script. Select your files and acquire the complete fasta file.

6.2.3 Generating the feature table file.

Select all your sequences and go to File, Export... Select file type GenBank Flat File Format (*.gb). Place the file in the same folder as the convertToFeatureTable.py script. Open Terminal on Mac or Windows Command Line(CMD). You need to have python on your computer to run this script. Change directory to folder with script and file. Run the following code: python convertToFeatureTable.py It will ask for your file name. Input your file name and the output file with your file name_output.tbl will be found in the same folder. This is your feature table file.

6.2.4 Generating the .sqn file.

After obtaining all three types of files, download tbl2asn. Now, you can run it in terminal with the following command: /usr/local/src/NCBI_toolkit/linux.tbl2asn -t template.sbt -p ./ -V v -a s
To check for errors, run: check output.val If errors are present, go to http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/errmsg/valid.msg. Finally, you can go to http://www.ncbi.nlm.nih.gov/projects/LargeDirSubs/dir_submit.cgi to submit the .sqn file generated.

YAY! That's it! THE END ☺

Template file

Metadata: Project Title, Sequence authors, Publication status
Suffix of file: **.sbt**

Nucleotide sequence data

in Fasta format
Suffix of file: **.fsa**

Feature Table

Suffix of file: **.tbl**

Create Submission Template
http://www.ncbi.nlm.nih.gov/WebSub/template.cgi

Export from Geneious and run Benjamin's Script

Export flat file from Geneious and run convertToFeatureTable.py script

tbl2asn

Generate in Terminal(Mac) or Windows Command Line(CMD)

File in .sqn format

Load onto Sequin

Submit to Genbank !

Summary of Steps

1. Assembling sequences. Select all raw data and select De Novo Assemble using the correct configurations.

2. Blasting against Barcodes. Sequence search using the barcodes provided and change the supercontig names accordingly.

3. Editing supercontigs. Attempt to resolve mismatches and try to circularise sequences that are around 18,000 bps in length.

4. Annotating genomes. Find a reference complete mitochondria by blasting the COX1 gene on your supercontig. Transfer the gene annotations onto your supercontig. Manually edit the start and stop codons of all 13 genes.

5. Removing Annotations. Remove unneeded and irrelevant annotations transferred from the reference genome.

6. Adding tRNAs. Use Putty and log on to ctag. Thereafter, assemble the results with your supercontigs and transfer annotations.

7. Remove other Annotations. Check again to see if all unwanted annotations have been removed.

8. Submitting to GenBank. Generate three different files: Template file(suffix .sbt), Fasta file(suffix .fsa) and the Feature table file(suffix .tbl). Use tbl2asn to generate a .sqn file to be submitted to GenBank.

9. YAY. Pass the Geneious files and any other relevant documents to your supervisor. Good job! You deserve a beer! Cheers! ☺ ☺ ☺.

References

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. and Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics (Oxford, England) *28*, 1647–9.