CONVENTIONAL BARCODING

ONTbarcoder (available from https://github.com/asrivathsan/ONTbarcoder) is installed by unzipping the folder with the version of the program that supports the operating system on your computer. The folders are available from releases in Github. For MacOS, an .app bundle has been created.

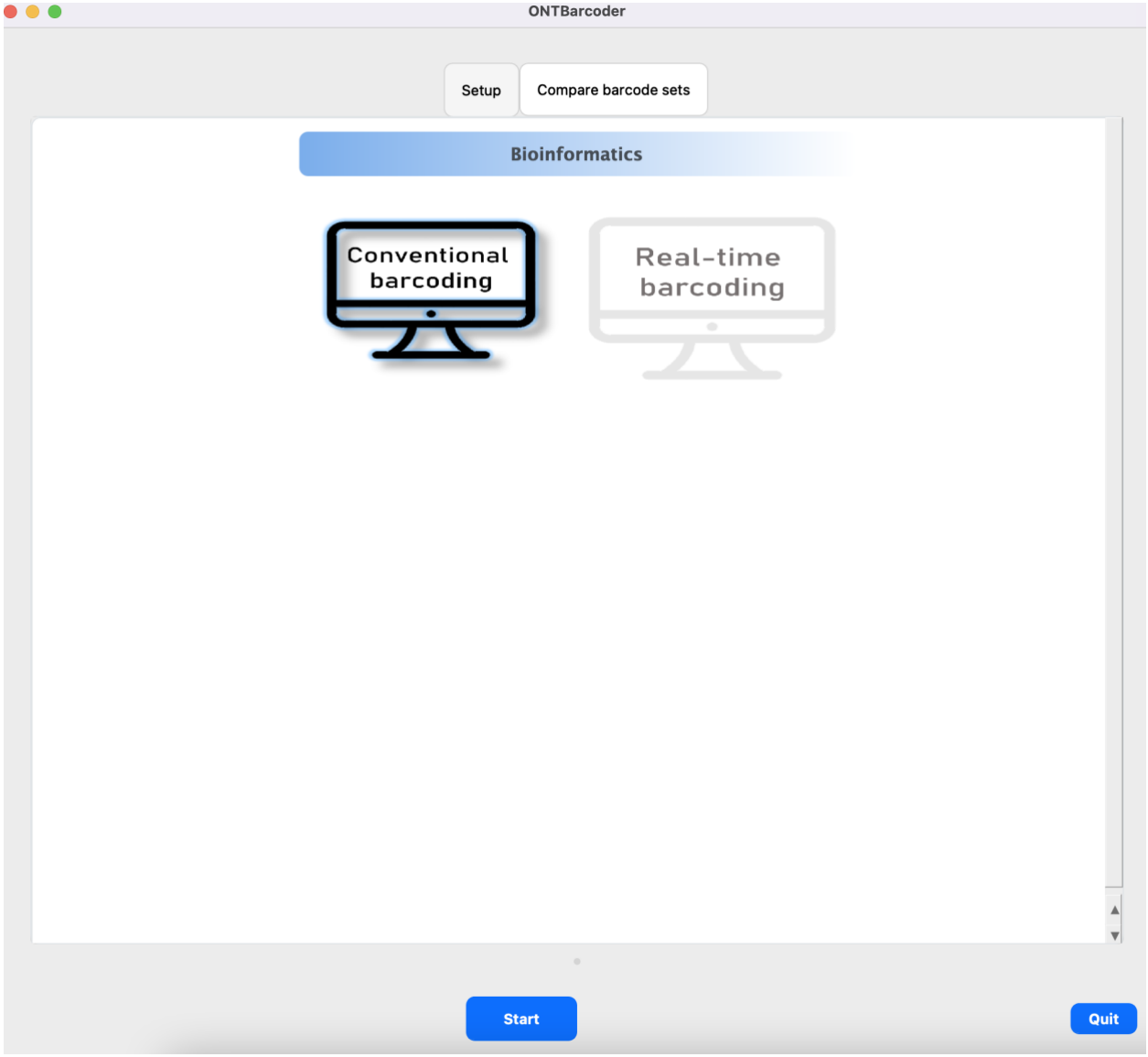
ONTbarcoder is designed to handle dual tagged amplicon pools. It performs demultiplexing and consensus calling If the software is used for the whole process, the format specifications are relatively simple and you can skip to the next section. If however, the demultiplexing is carried out independently, please follow the guidelines in Format Specifications, Point 2 in Page 10 of this documentation. It would be important to ensure that consensus calling is done under the assumption that primers and tags are removed. At least, tags should be removed if primers do not have ambiguous bases. The length parameters should be modified to expected consensus length.

ONTbarcoder has been optimized for **protein coding gene like COI**. While there are ways to obtain consensus for length variable non coding genes, this has not been extensively tested. *Please refer to last page of the manual to see which outputs to use if length variable and noncoding genes are being barcoded.*

For **conventional barcoding** recommend that you try the pipeline using our DatasetA containing 257 amplicons generated from Flongle. It is available from here. Information to download the various datasets is available here. Prior to running any dataset, ensure you have at least 1.5 times the space required for the input FASTQ file.

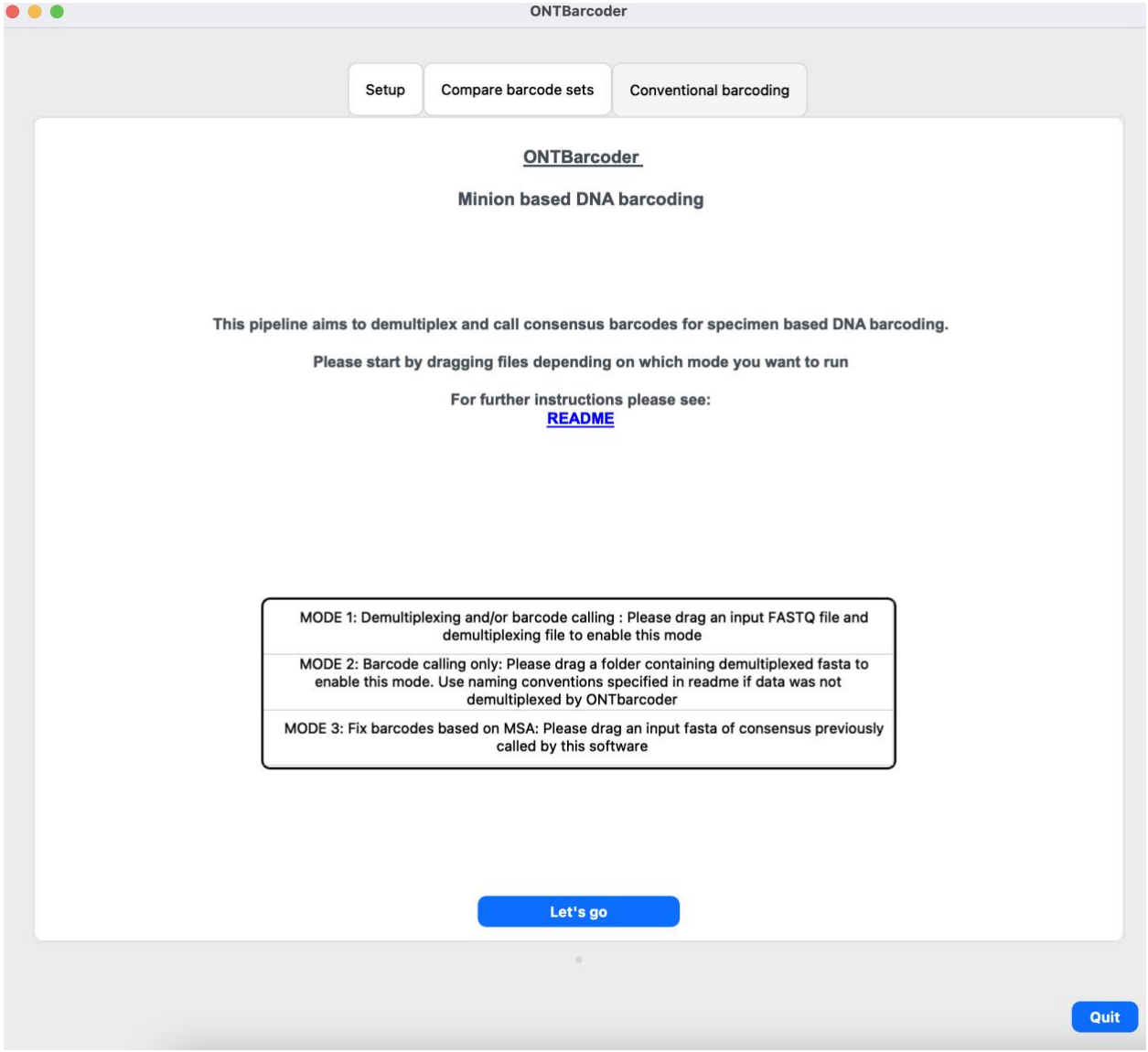
A simple video tutorial is available. This tutorial is for the older ONTbarcoder (0.1.9) which describes Conventional barcoding. You can use “Conventional barcoding” of ONTbarcoder2, you don’t need to download the older ONTbarcoder.

1. Open ONTbarcoder, select Conventional barcoding



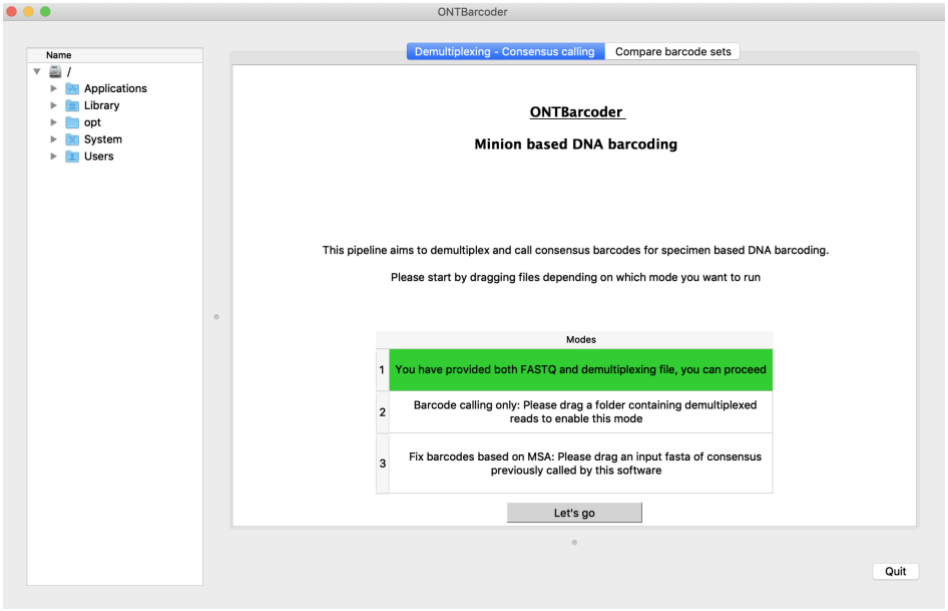
This should lead to the following menu. The rest of the steps are identical to ONTbarcoder1 shown below.

Only modification in ONTbarcoder2 is the option for selecting tag errors



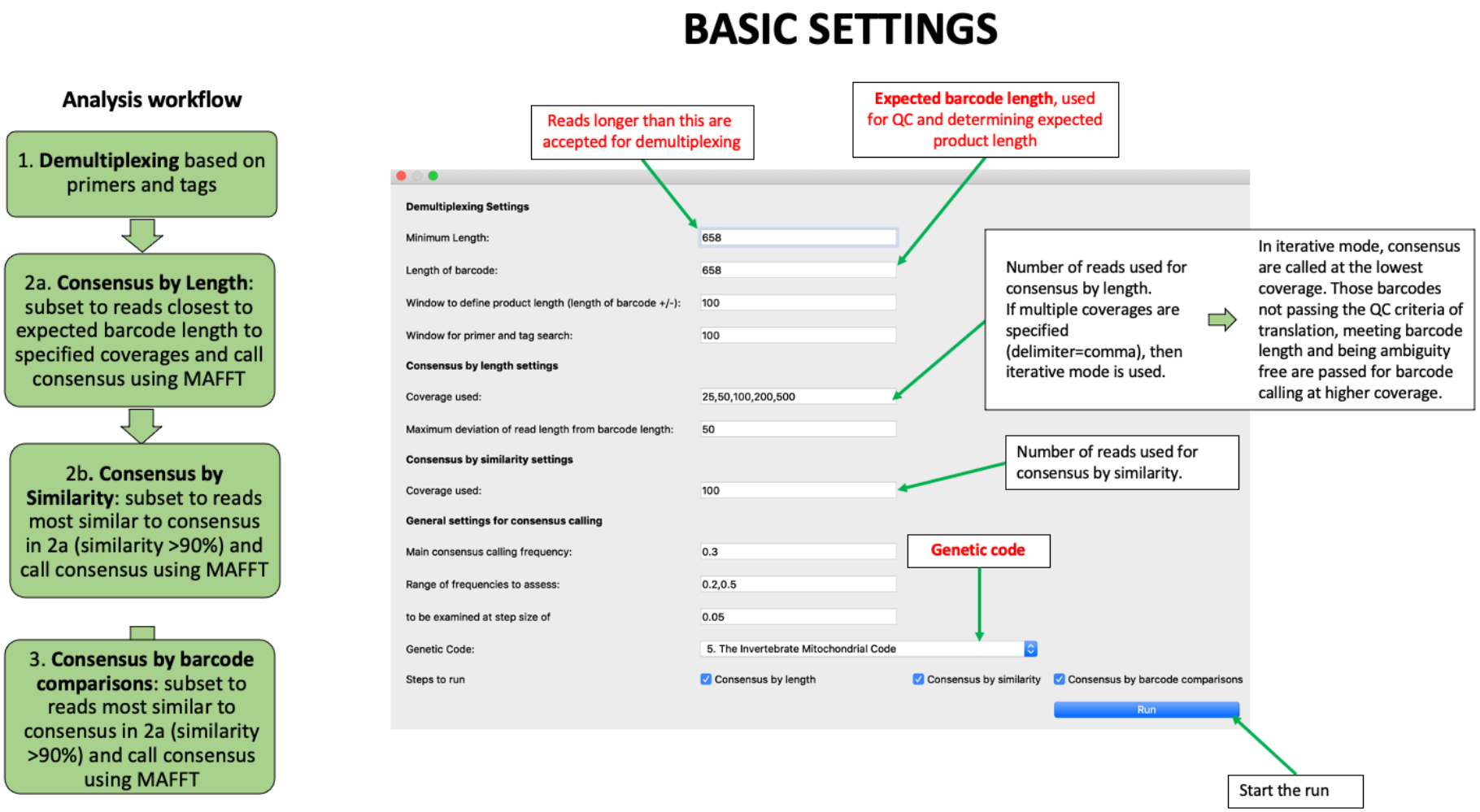
Demultiplexing and barcode calling using the “Demultiplexing-Consensus calling tab”

2. Drag in Input Fastq and demultiplexing file, either simultaneously or one by one

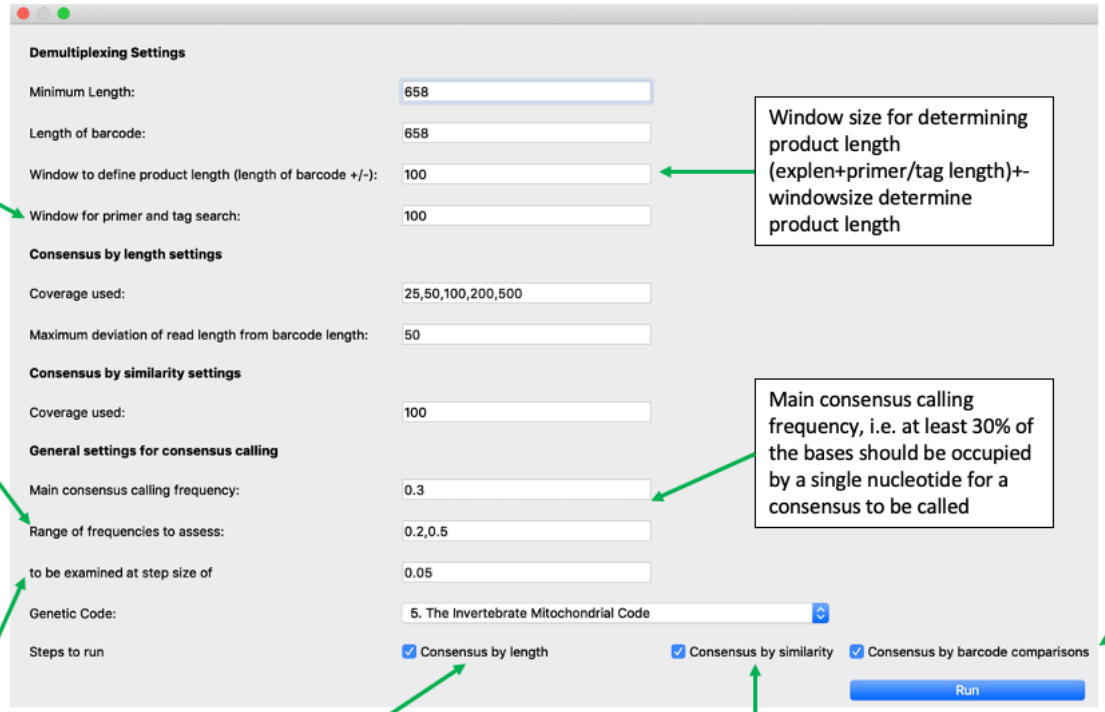


3. Press “Let’s go”

4. Configure your run based on the settings described below. DatasetA does not require any changes to the default. The default assumes that you are working on invertebrate barcodes that are 658 bp in length. The remaining settings define which reads are demultiplexed, where the primers are found, how many reads are used in the different steps, and how the consensus barcodes are determined.



ADVANCED SETTINGS



Primers & tags are searched within the first/last specified number of bp of a read

Window size for determining product length (explen+primer/tag length)+- windowsize determine product length

Other consensus calling frequencies to test

Main consensus calling frequency, i.e. at least 30% of the bases should be occupied by a single nucleotide for a consensus to be called

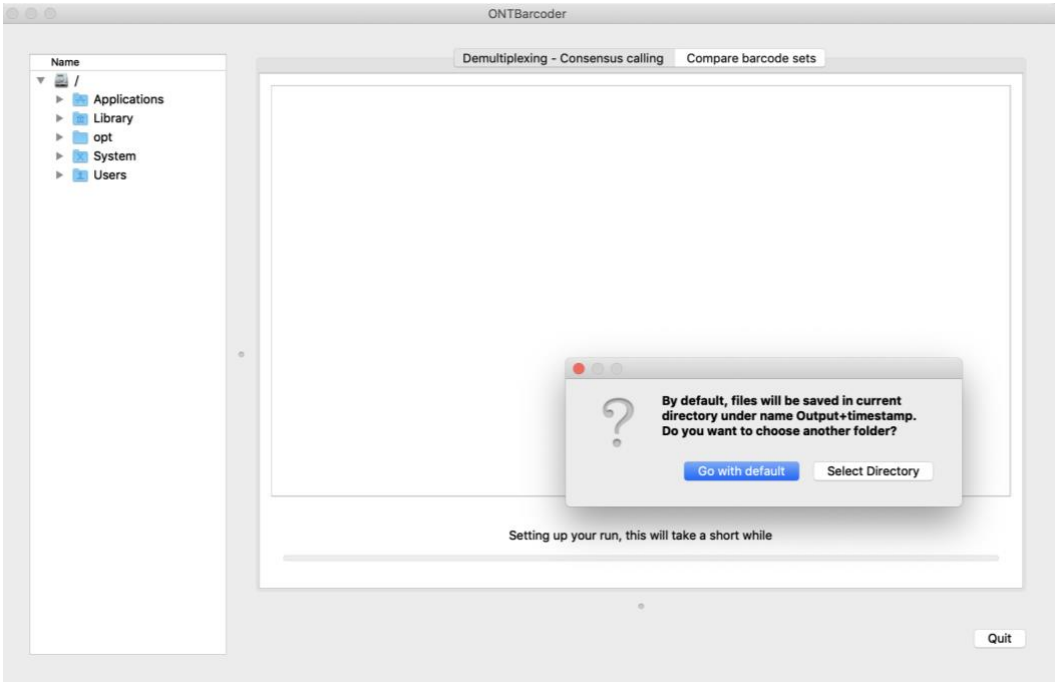
Specifies the steps for the range. In current settings 0.2,0.25,0.3,0.35,0.4,0.45,0.5 are tested

Enables/disables **Consensus by barcode comparisons,** i.e. barcode fixing that fills/ eliminates gaps in barcodes by comparison to other barcodes in the run

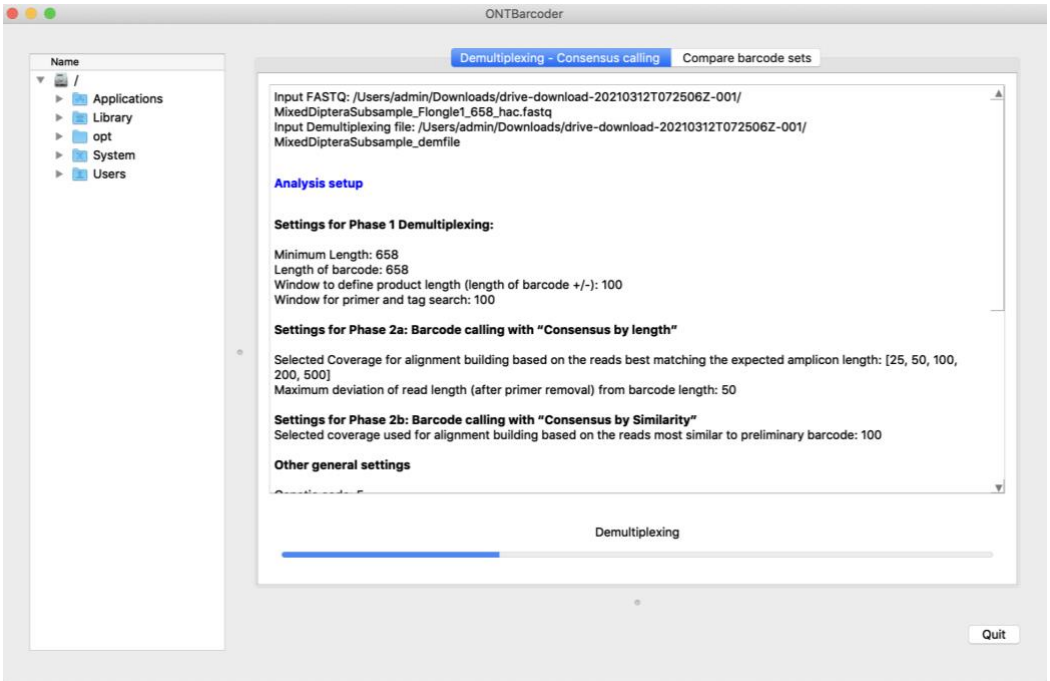
Enables/disables **Consensus by length**: Step that uses reads closest to the known length of the barcode are used for consensus calling.

Enables/disables **Consensus by similarity**: Step that improves consensus in previous step by taking reads most similar to it.

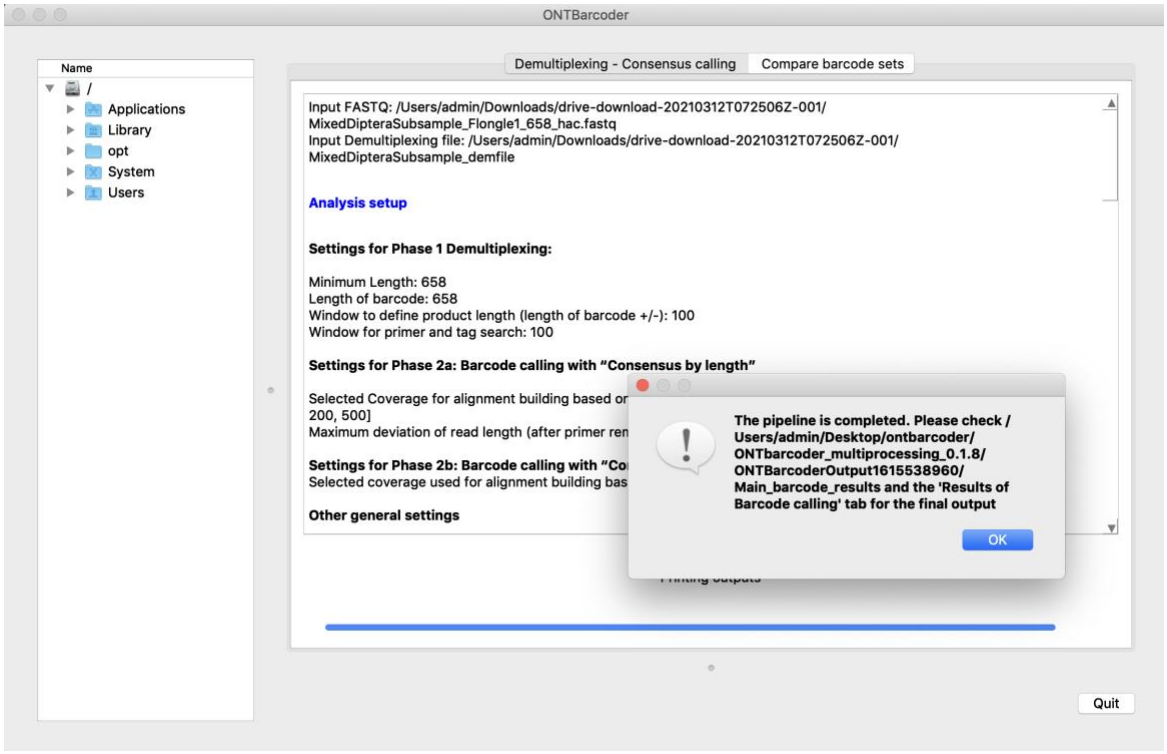
5. Select an output folder. If you select “default”, it will create a folder within the directory containing the software with a unique name starting with ONTbarcoderOutput. If you choose “Select Directory” ensure that the directory name has no empty space and is empty. Generally avoid complex characters in your names (spaces, brackets, slashes, colons)



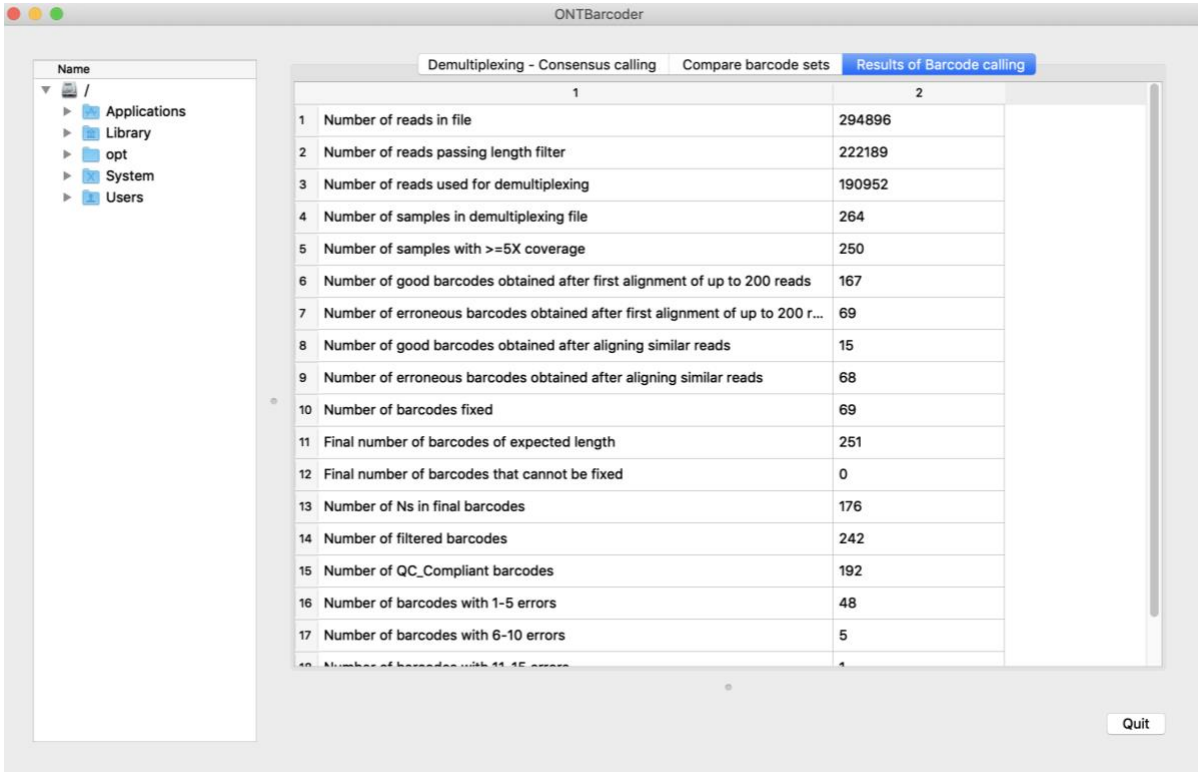
6. The analysis starts, a progress bar appears, and results are generated in respective folders as the analysis advances.



7. The analysis is complete and all output files have been created.

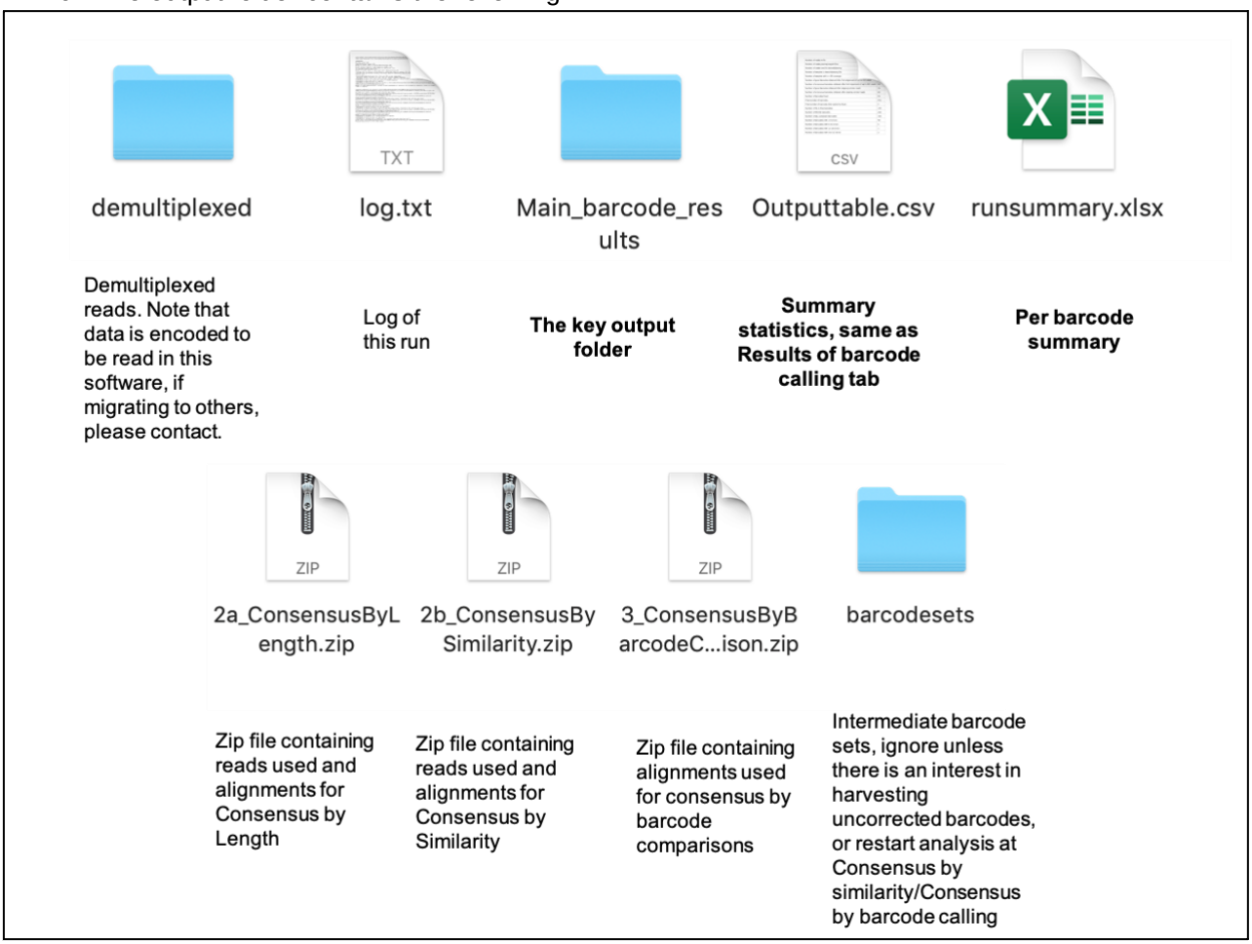


8. Results Table: The table reports that 242 "Filtered" barcodes were obtained of which 192 are QC-Compliant. This means they have no ambiguities, are translatable, have expected barcode length and are not having any gaps in an internal Multiple Sequence Alignment (MSA) check. The filtered sets include addition 50 barcodes that have <1% ambiguous bases, are translatable, have expected length, and may have been corrected for up to 5 indels. The table can be copied but it also stored in the "Outputtable.csv" file in the output folder.

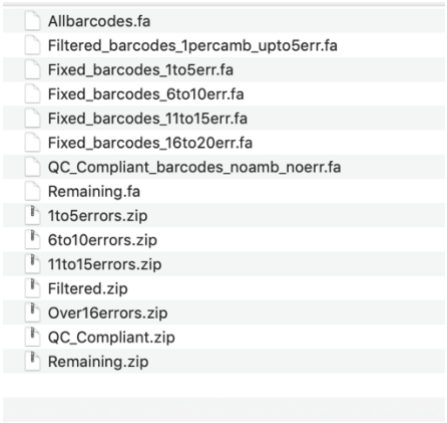


	1	2
1	Number of reads in file	294896
2	Number of reads passing length filter	222189
3	Number of reads used for demultiplexing	190952
4	Number of samples in demultiplexing file	264
5	Number of samples with >=5X coverage	250
6	Number of good barcodes obtained after first alignment of up to 200 reads	167
7	Number of erroneous barcodes obtained after first alignment of up to 200 r...	69
8	Number of good barcodes obtained after aligning similar reads	15
9	Number of erroneous barcodes obtained after aligning similar reads	68
10	Number of barcodes fixed	69
11	Final number of barcodes of expected length	251
12	Final number of barcodes that cannot be fixed	0
13	Number of Ns in final barcodes	176
14	Number of filtered barcodes	242
15	Number of QC_Compliant barcodes	192
16	Number of barcodes with 1-5 errors	48
17	Number of barcodes with 6-10 errors	5

9. The output folder contains the following



10. The Main Barcode Results folder contains the main barcode output file **"Filtered_barcodes"** fasta file. The barcodes without ambiguous bases are in the QC_Compliant_barcodes file. The zip folders contain demultiplexed reads per dataset divided into each of the categories. All barcodes should be checked via BLAST for contamination and/or verification that the barcode obtained belongs to the expected taxon.



Summary

Expected input and output

1. Conventional barcoding	Input required	Results
Mode 1: Demultiplexing + barcode calling	• FASTQ file obtained after base-calling • Demultiplexing file	• Demultiplexed reads in "demultiplexed" folder • Overall summary in "Outputtable.csv" • Per barcode summary in "runsummary.xlsx
Mode 2: Barcode calling only: Barcodes are derived from reads in specimen-specific bins	• Folder containing Demultiplexed FASTA files. Format specified under "Directory containing Demultiplexed FASTA files"	• Barcode sets in "Main_barcode_results" folder • Overall summary in "Outputtable.csv" • Per barcode summary in "runsummary.xlsx"
Mode 3: Improving barcodes using "Consensus by similarity" and "Consensus by barcode comparisons"	• Sequences in "barcodesets" folder	• Barcode sets available from "Main_barcode_results" folder • Overall summary in "Outputtable.csv" • Per barcode summary in "runsummary.xlsx"

Format Specifications

File	
1. FASTQ file	Standard fastq, generated after basecalling with ONT software
2. Demultiplexing File	A 5-column csv file with the following headers: SpecimenID, TagFsequence, TagRsequence, PrimerF, PrimerR You can only demultiplex one one primer pair at a time. FASTQ files with data for multiple pairs, have to processed sequentially. **Please avoid unusual characters in Specimen ID (e.g. characters like "(){}[]\/.,;*$" will lead to crashes)**
3. Directory containing Demultiplexed FASTA files	Input for Mode 2 should be a directory that contains only FASTA files. Directory should not be empty nor should it contain other files. The fasta files must have names as "sampleID_all.fa" i.e. the suffix _all.fa is critical
4. Input file for improvement (Step 4)	ONTbarcoder's pipeline can be started at different points. Demultiplexing files can be used to only carry out barcode calling. Consensus by barcode can be started with files that have the format specified in the hyperlink

LENGTH VARIABLE AND NON-CODING GENES

Accurate barcodes have been obtained by preselection of reads to the specified length criteria of the gene. Barring this issue, ONTbarcoder can be used for non-coding, length variable genes. For this, please disable consensus by similarity and consensus by barcode fixing step in the menu. The barcode calling can be conducted as per normal. The resulting barcode file can be found in barcodesets/consensus_all_step1.fa

This can be parsed by the user to retain barcodes with few ambiguities (for e.g. <1%)