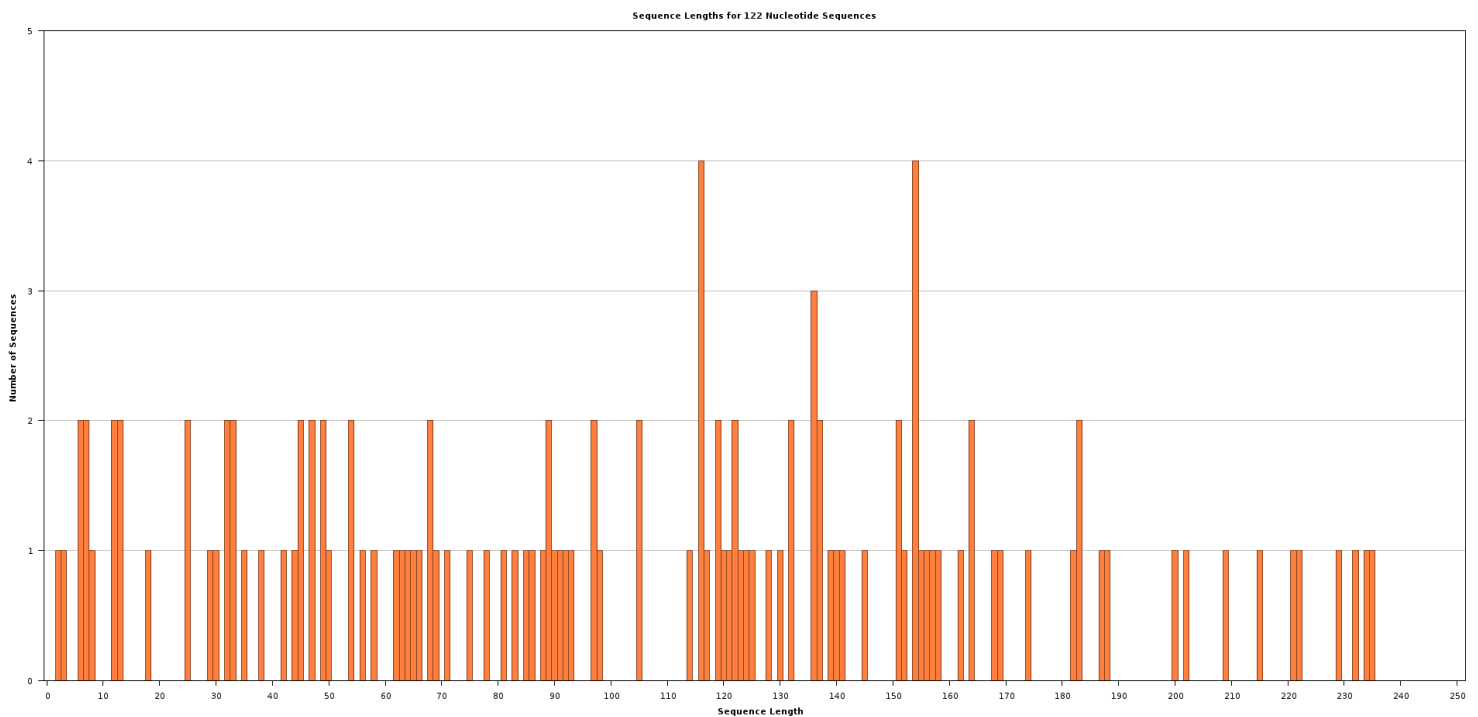


- **Species: Cinnamomum capparucoronde**
- **Data size: 1GB**
- **Comparison between ITSx output obtained using vsearch output as input and ITSxpress output obtained using contigs as input**

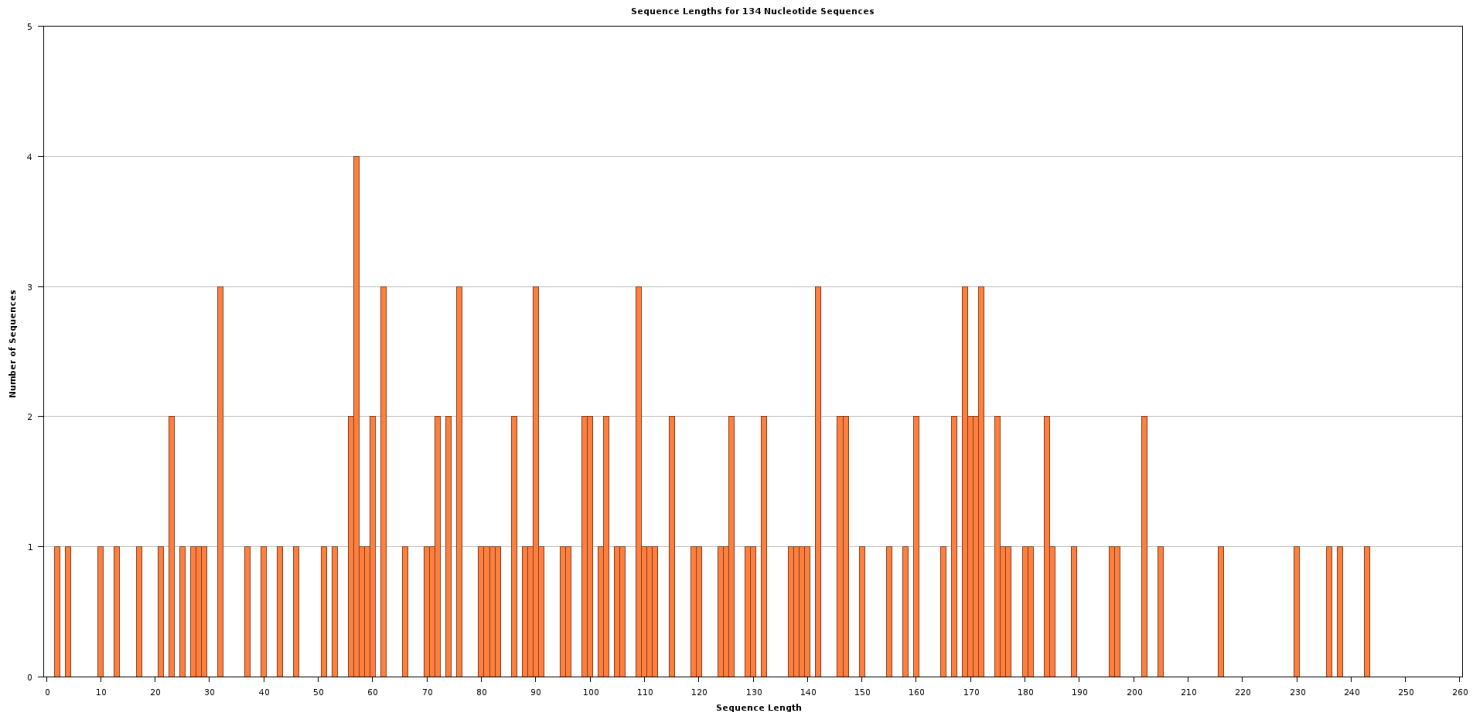
ITSx- Output

- creates separate sequence files for ITS1 and ITS2 (one FASTA file for the ITS1 and ITS2 regions)
- There are multiple sequences identified as candidate ITS1 and ITS2 sequences (considering all the merged reads)
- **“ITSx is not designed to make accurate predictions on organism groups, and no double-checking of this prediction is performed. Therefore, the ITS sequences extracted should be further examined using e.g. BLAST searches.” - Taken from the ITSx user guide**
- 122 candidate sequences for ITS2 and 134 candidate sequences for ITS1
- Length ranges of candidate sequences

❖ ITS2



❖ ITS1



Determine average lengths of ITS2 and ITS1 in plants

1. Yao, Hui & Song, Jingyuan & Liu, Chang & Luo, Kun & Jianping, Han & Li, Ying & Pang, Xiaohui & Xu, Hongxi & Zhu, Yingjie & Xiao, Pei-Gen & Chen, Shilin. (2010). Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals. *PloS one*. 5. 10.1371/journal.pone.0013102.
 2. Liston, Aaron & Robinson, William & Oliphant, James & Alvarez-Buylla, Elena. (1996). Length Variation in the Nuclear Ribosomal DNA Internal Transcribed Spacer Region of Non-Flowering Seed Plants. *Systematic Botany*. 21. 109. 10.2307/2419742.
 3. Feng, Shangguo & Mengying, Jiang & Shi, Yujun & Jiao, Kaili & Shen, Chenjia & Lu, JiangJie & Ying, Qicai & Wang, Huizhong. (2016). Application of the Ribosomal DNA ITS2 Region of *Physalis* (Solanaceae): DNA Barcoding and Phylogenetic Study. *Frontiers in Plant Science*. 7. 10.3389/fpls.2016.01047.
 4. Wang, Xin-Cun & Liu, Chang & Huang, Liang & Bengtsson-Palme, Johan & Chen, Haimei & Zhang, Jian-Hui & Cai, Dayong & Li, Jian-Qin. (2015). ITS1: a DNA barcode better than ITS2 in eukaryotes?. *Molecular Ecology Resources*. 15. 573–586. 10.1111/1755-0998.12325.
- Considering the details in the above publications and Cinnamon is a dicot plant, the sequences with 100bp - 681bp and 101bp-522bp were taken as the candidate ITS1 and ITS2 regions respectively.
 - Accordingly, 62 sequences for ITS2 and 74 sequences for ITS1 were retained for further analyses.

Test1: Align all the sequences taken as ITS2 and ITS1 regions separately

ITS2

- The multiple alignment for the selected 62 sequences was done using the MAFFT algorithm
- Base-pair difference ranges from 0bp to 183bp
- The distance matrix was divided into 4 sections to have a deep evaluation

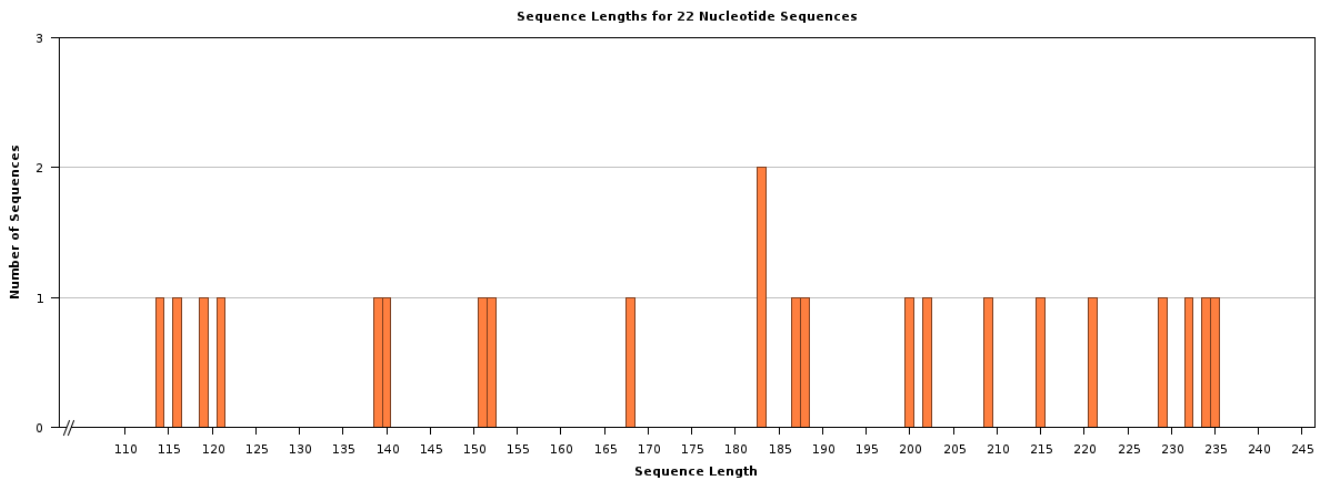
Section 1:

	E00526:96:HWJYCCXY:5:2105:25915:9255[T]ITS2)052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	
E00526:96:HWJYCC...	0	0	0	0	0	1	1	0	3	0	0	0	2	2	1	1	1	1	1	3	2
E00526:96:HWJYCC...	0	0	0	0	0	1	1	0	3	0	0	0	2	2	1	1	1	1	1	3	2
E00526:96:HWJYCC...	0	0	0	0	0	1	1	0	3	0	0	0	2	2	1	1	1	1	1	3	2
E00526:96:HWJYCC...	0	0	0	0	0	1	1	0	3	0	0	0	2	2	1	1	1	1	1	3	2
E00526:96:HWJYCC...	0	0	0	0	0	1	1	0	3	0	0	0	1	1	1	1	1	1	1	3	2
E00526:96:HWJYCC...	0	0	0	0	0	1	1	0	3	0	0	0	1	1	1	1	1	1	1	3	2
E00526:96:HWJYCC...	1	1	1	1	1	1	2	0	4	0	0	0	3	3	2	1	1	1	1	3	3
E00526:96:HWJYCC...	1	1	1	1	1	1	2	1	2	1	1	1	3	3	2	2	2	2	2	4	3
E00526:96:HWJYCC...	0	0	0	0	0	0	1	1	0	0	0	1	1	1	1	1	1	1	1	3	2
E00526:96:HWJYCC...	3	3	3	3	3	3	4	2	1	1	1	5	5	4	3	2	2	2	4	5	5
E00526:96:HWJYCC...	0	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1	1	1	1	3	2
E00526:96:HWJYCC...	0	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1	1	1	1	3	2
E00526:96:HWJYCC...	0	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1	1	1	1	3	2
E00526:96:HWJYCC...	2	2	2	2	1	1	3	3	1	5	1	1	1	1	0	0	0	0	0	2	1
E00526:96:HWJYCC...	2	2	2	2	1	1	3	3	1	5	1	1	1	1	0	0	0	0	0	2	1
E00526:96:HWJYCC...	1	1	1	1	1	1	2	2	1	4	1	1	1	1	0	0	0	0	0	2	1
E00526:96:HWJYCC...	1	1	1	1	1	1	2	1	3	1	1	1	1	0	0	0	0	0	0	2	1
E00526:96:HWJYCC...	1	1	1	1	1	1	2	1	2	1	1	1	1	0	0	0	0	0	0	2	1
E00526:96:HWJYCC...	1	1	1	1	1	1	2	1	2	1	1	1	1	0	0	0	0	0	0	2	1
E00526:96:HWJYCC...	1	1	1	1	1	1	2	1	2	1	1	1	1	0	0	0	0	0	0	2	1
E00526:96:HWJYCC...	1	1	1	1	1	1	2	1	2	1	1	1	1	0	0	0	0	0	0	2	1
E00526:96:HWJYCC...	3	3	3	3	3	3	4	3	4	3	3	3	2	2	2	2	2	2	2	3	3
E00526:96:HWJYCC...	2	2	2	2	2	2	3	3	2	5	2	2	2	1	1	1	1	1	1	3	3

- This section contains the most similar sequences - 22 sequences
- Randomly selected sequences were blasted against the NCBI nr/nt database (02/16/2021; 1.30 pm - 2.00 pm)
- The blast results contained ITS2 regions of *Cinnamomum wightii*, *Cinnamomum chemungianum*, *Cinnamomum perrottetii*, *Cinnamomum travancoricum*, *Cinnamomum mathewianum*, *Cinnamomum dubium*, *Cinnamomum tamala*, *Cinnamomum litseifolium*, *Cinnamomum keralaense*, *Cinnamomum verum*, *Cinnamomum sulphuratum*, *Cinnamomum goaense*, *Cinnamomum agasthyamalayanum*, *Cinnamomum verum*, *Cinnamomum macrocarpum*.
- Thus, these sequences can be considered for further evaluation.

GC content : 73.0% - 76.0%

Length distribution



- The blast results contained ITS2 regions of fungi species (*Neopestalotiopsis saprophytica*, *Neopestalotiopsis formicarum*, *Neopestalotiopsis foedans*, *Pestalotiopsis scorpina*, *Pestalotiopsis maculans*, *Diaporthe searlei*, *Diaporthe drengthii*, *Neofusicoccum parvum*, *Neofusicoccum yunnanense*).

Section 4:

E00526:96:HWJYCCC...	X	4	4	3	3	2	2	2	17	16
E00526:96:HWJYCCC...	4	X	0	1	1	1	1	0	13	12
E00526:96:HWJYCCC...	4	0	X	1	1	1	1	0	13	12
E00526:96:HWJYCCC...	3	1	1	X	0	0	0	0	14	13
E00526:96:HWJYCCC...	3	1	1	0	X	0	0	0	14	13
E00526:96:HWJYCCC...	2	1	1	0	0	X	0	0	14	13
E00526:96:HWJYCCC...	2	1	1	0	0	0	X	0	14	13
E00526:96:HWJYCCC...	2	0	0	0	0	0	0	X	13	12
E00526:96:HWJYCCC...	17	13	13	14	14	14	14	13	X	1
E00526:96:HWJYCCC...	16	12	12	13	13	13	13	12	1	X

- Randomly selected sequences were blasted against the NCBI nr/nt database (02/16/2021; 3.14 pm - 3.20 pm)
- The blast results contained ITS2 regions of fungi species (*Colletotrichum acutatum*, *Colletotrichum nymphaeae*, *Colletotrichum fructicola*, *Colletotrichum gloeosporioides*, *Colletotrichum fragariae*, *Colletotrichum siamense*).

- **Considering the above results, sequences considered in sections 2-4 can be discarded from the analysis.**

❖ **Test1.1: All these 62 sequences were mapped to the *Cin. capparu-coronde* ITS (ITS1-5.8s-ITS2-26s) sequence which I (sandali) constructed.**

❖ **Interestingly, only 22 sequences in Section 1 were mapped.**



ITS1

- Same as above, all the 74 sequences were aligned using the MAFFT algorithm and the base-pair difference in the resulting distance matrix was 0bp -196bp.
- The read alignment was evaluated by dividing the distance matrix into two sections.

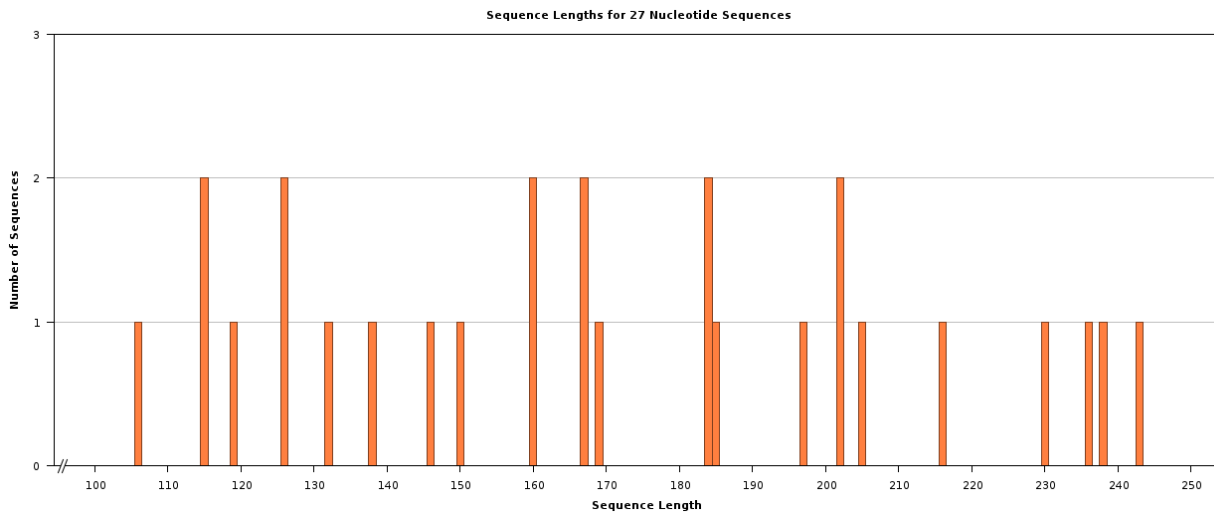
Section 1

E00526:96:HWYCCXXY:5:1101:18081:16252(T)ITS1:52...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...	E0052...		
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	2	1	1	1	1	2	2	1	0	0	1	0	1	0	0	0	0	0	0	0	0
0	1	1	1	0	0	0	0	1	1	0	1	1	0	1	0	1	0	0	0	0	0	0	0
1	2	1	1	1	1	1	3	3	1	2	2	1	3	2	2	2	1	1	1	1	1	1	1
0	1	0	1	0	0	0	0	2	2	0	1	1	0	2	1	1	1	0	0	0	0	0	0
0	1	0	1	0	0	0	0	2	2	0	1	1	0	2	1	1	1	0	0	0	0	0	0
0	1	0	1	0	0	0	0	2	2	0	1	1	0	2	1	1	1	0	0	0	0	0	0
0	1	0	1	0	0	0	0	2	2	0	1	1	0	2	1	1	1	0	0	0	0	0	0
1	2	1	3	2	2	2	2	0	2	3	3	1	2	1	2	2	1	1	1	1	1	1	1
1	2	1	3	2	2	2	2	0	2	3	3	1	2	1	2	2	1	1	1	1	1	1	1
0	1	0	1	0	0	0	0	2	2	0	1	1	0	2	1	1	1	0	0	0	0	0	0
0	0	1	2	1	1	1	1	3	3	1	0	1	1	2	1	0	0	0	0	0	0	0	0
0	0	1	2	1	1	1	1	3	3	1	0	1	1	2	0	0	0	0	0	0	0	0	0
0	1	0	1	0	0	0	0	1	1	0	1	1	1	1	0	1	1	0	0	0	0	0	0
0	0	1	3	2	2	2	2	2	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0
0	1	0	2	1	1	1	1	1	2	2	0	1	1	1	1	0	0	0	0	0	0	0	0
0	0	1	2	1	1	1	1	2	2	1	0	0	1	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- This section contains the most similar sequences - 27 sequences
- Randomly selected sequences were blasted against the NCBI nr/nt database (02/16/2021; 3.47 pm - 4.00 pm)
- The blast results contained ITS2 regions of *Cinnamomum tonkinense*, *Cinnamomum verum*, *Cinnamomum multiflorum*, *Cinnamomum iners*, *Cinnamomum dictyoneuron*, *Cinnamomum bejolghota*, *Cinnamomum kotoense*, *Cinnamomum rhynchophyllum*, *Cinnamomum javanicum*, *Cinnamomum cordatum*, *Cinnamomum tamala*, *Cinnamomum grandifolium*)
- Thus, these sequences can be considered for further evaluation.

GC content : 71.3% - 75.5%

Length distribution

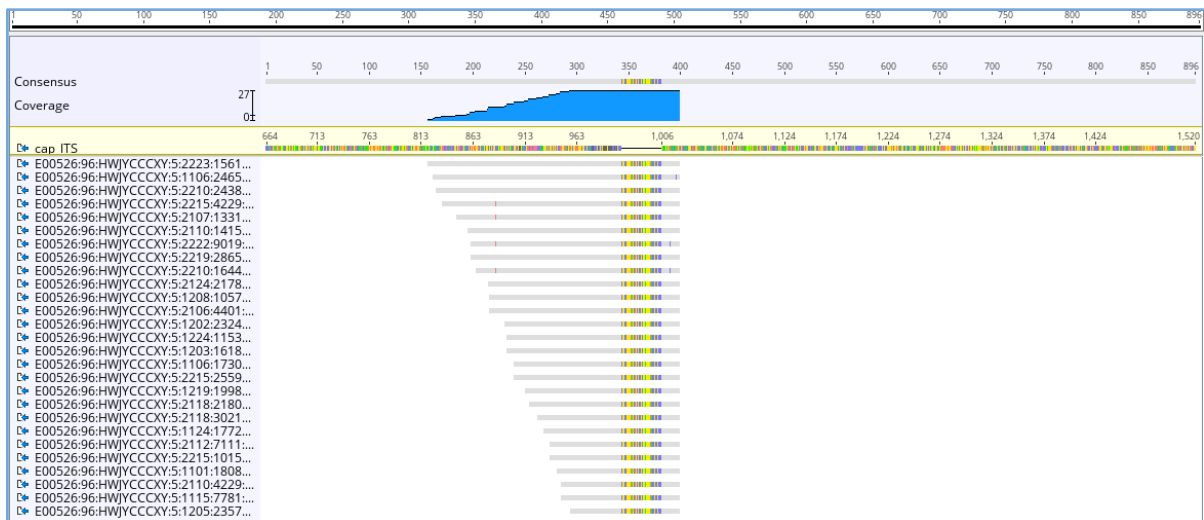


Section 2:

- Randomly selected sequences were blasted against the NCBI nr/nt database (02/16/2021; 4.05 pm - 4.20 pm)
- The blast results contained ITS2 regions of fungi species (*Ceramothyrium longivolcaniforme*, *Cyphellophora pruni*, *Neofusicoccum parvum*, *Diaporthe guangxiensis*, *Diaporthe searlei*, *Diaporthe lithocarpus*, *Phomopsis phyllanthicola*, *Colletotrichum scovillei*, *Colletotrichum fioriniae*).
- Thus, these sequences can be discarded from the analysis.

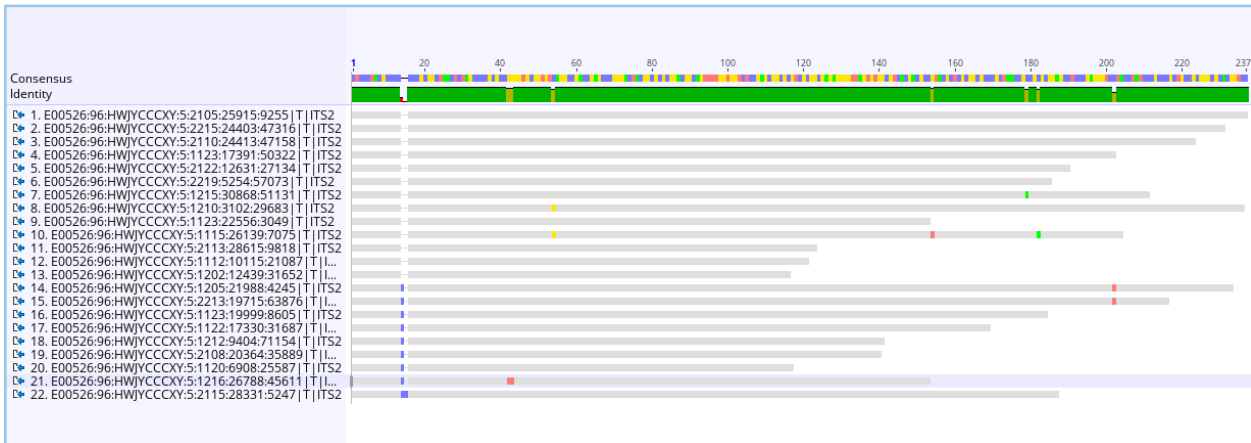
❖ **Test1.2: All these 74 sequences were mapped to the *Cin. capparu-coronde* ITS (ITS1-5.8s-ITS2-26s) sequence which I (sandali) constructed.**

❖ **Interestingly, only 27 sequences in Section 1 were mapped.**

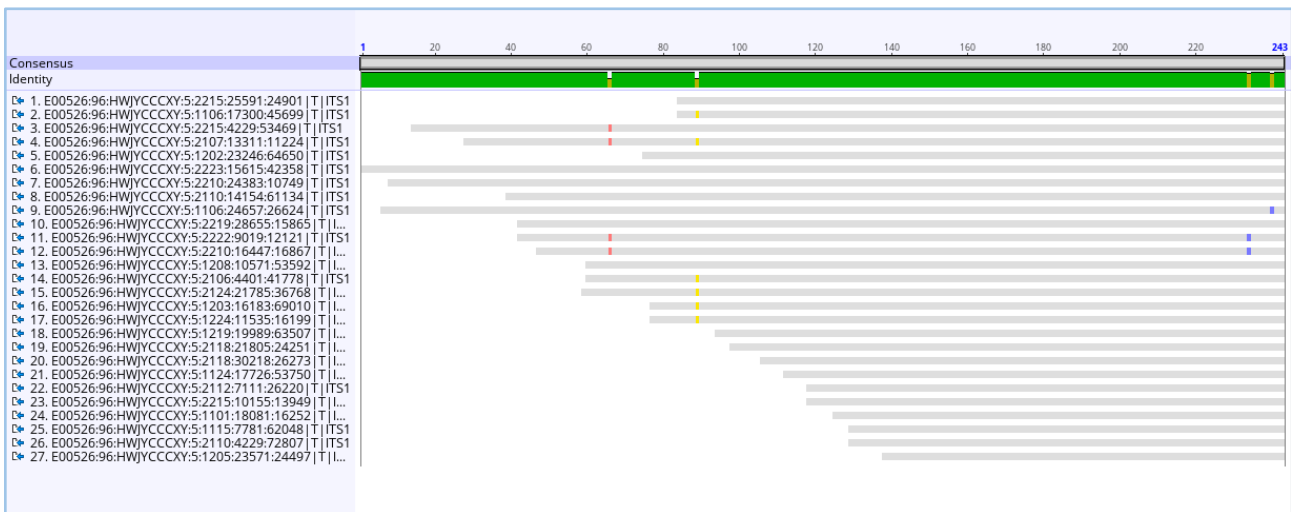


- **Since there are 22 sequences for ITS2 and 27 for ITS1, the consensus sequences taken from the multiple alignments are considered to compare ITSx and ITSxpress results. (Taking consensus doesn't negatively impact on further comparisons- there are a maximum of 5 base-pair difference**

ITS2 alignment view



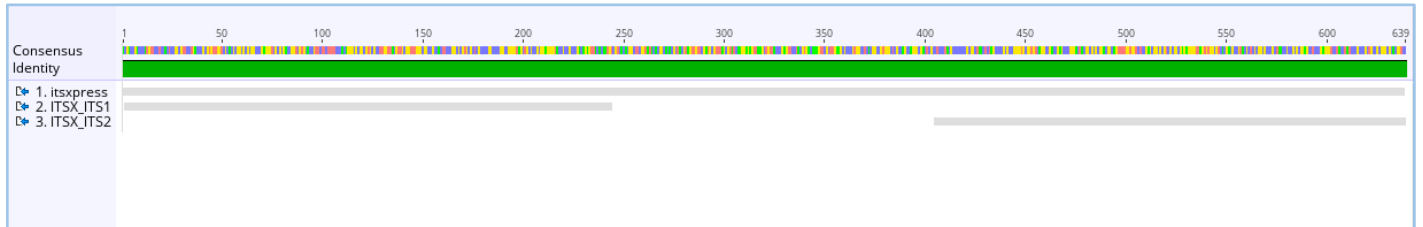
ITS1 alignment view



Test2: Compare the results from ITSx and ITSxpress

- ITS1 (consensus - ITSX_ITS1) and ITS2 (consensus - ITSX_ITS2) sequences taken from ITSx and the output sequence of ITSxpress (itsxpress) were aligned using the MAFFT algorithm.

Alignment view



Distance matrix

	itsxpress	ITSX_ITS1	ITSX_ITS2
itsxpress		0	0
ITSX_ITS1	0		0
ITSX_ITS2	0	0	

- According to the above results, there is no difference in the results taken from ITSx and ITSxpress