Identifying cryptic species within Mola Mola

- ## Working with the COI-5P gene

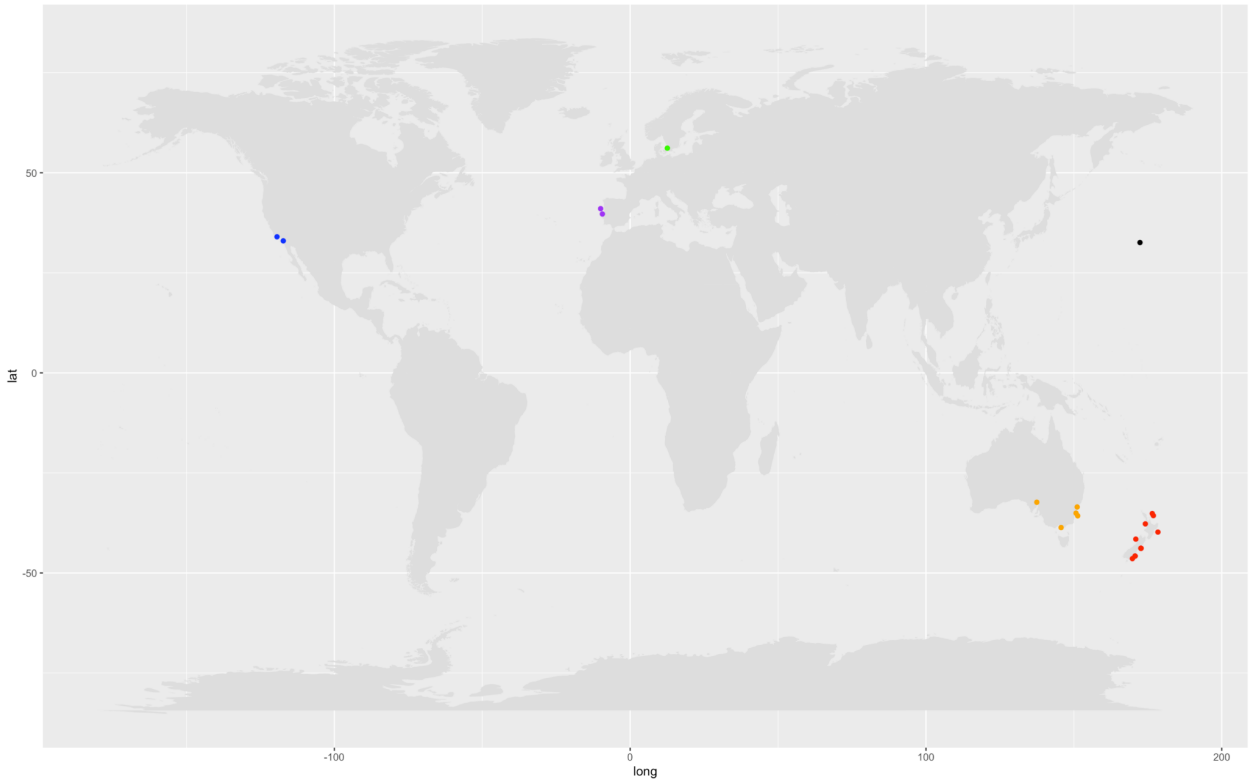
1. Extracting COI sequences from BOLD:

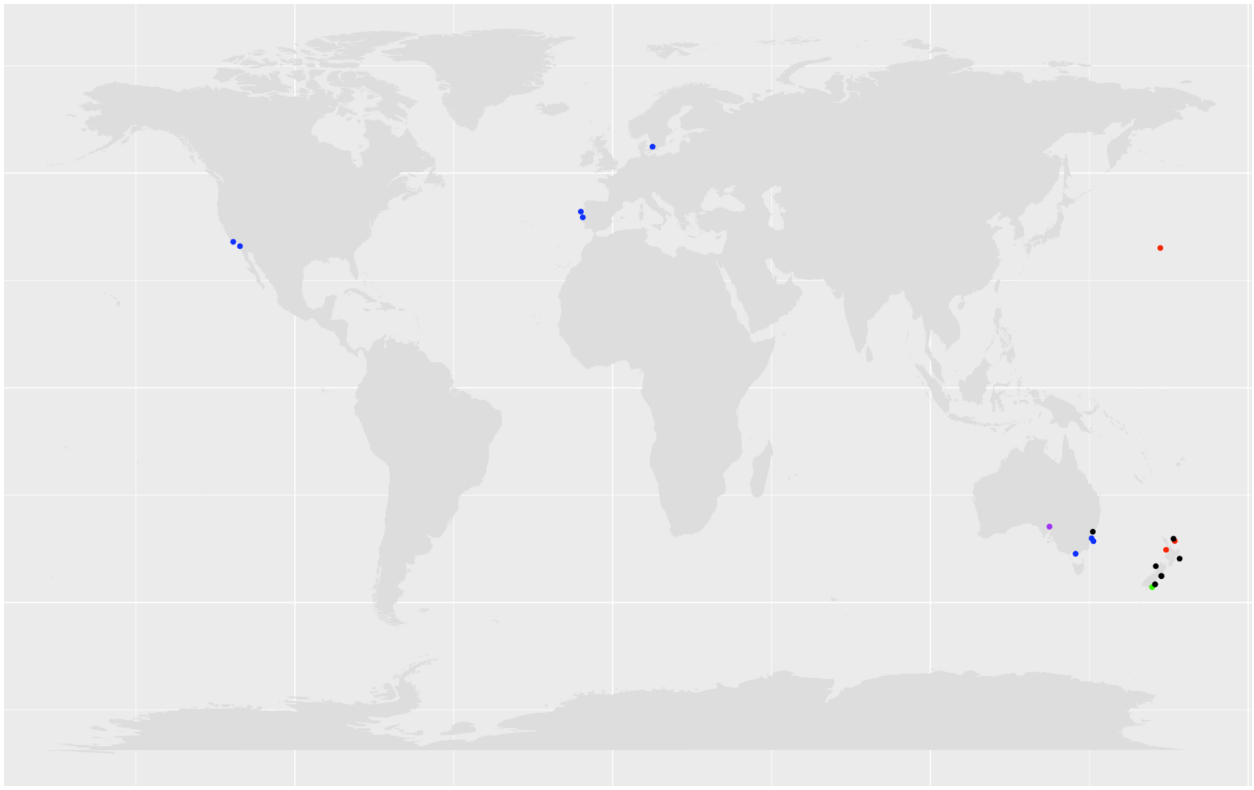
```
Seq_Mola <- bold_seqspec(taxon="Mola", marker="COI-5P")
Seq_Mola<-Seq_Mola%>%filter(!is.na(lat))
```

2. Divide sequences into different regions:

```
Seq_Mola$country[5]<-"Northwest Pacific"
Seq_Mola$country[which(Seq_Mola$country == "Pacific Ocean")] <- "New Zealand"
Seq_Mola$country[which(Seq_Mola$country == "United States")] <- "US West Coast"
```

3. Making a map:





4. Export sequences data:

```
>AMS170-08_Australia
CCTATACCTAATTTTCGGTGCCTGAGCCGGAATAGTAGGCACAGCACTAAGCCTTCTTATTCGAGCCGAACTAAGCCAACCCGGCGCTCT
TCTGGGCGACGACCAGATTTACAATGTAATCGTTACAGCCCACGCATTCGTAATGATTTTCTTTATAGTAATACCAATCATGATTGGAGG
CTTTGGAAACTGATTAATTCCCCTTATGATCGGAGCCCCTGATATAGCCTTCCCCCGAATGAATAACATAAGCTTTTGACTTCTTCCCCC
CTCTTTCCTCCTTCTACTCGCTTCTTCTGGAGTAGAGGCCGGAGCAGGAACGGGCTGAACTGTGTACCCCCCACTAGCAGGAAACCTTGC
CCACGCAGGAGCTTCTGTAGACCTAACTATCTTCTCTCTACACCTTGCAGGTGTCTCATCCATCCTAGGGGCTATCAACTTCATTACAAC
AATTATTAACATAAAACCACCTGCGATCTCACAATACCAAACACCTCTCTTCGTATGGGCAGTTCTAATTACTGCTGTGCTTCTTCTTCT
CTCCCTGCCAGTCCTCGCAGCGGGAATCACAATGCTTCTCACTGACCGAAACCTTAACACAACCTTCTTTGATCCCGCAGGCGGAGGGGA
CCCCATCCTTTACCAGCACCTA
>AMS174-08_Australia
CCTTTATTTAGTATTCGGTGCATGAGCCGGGATAGTGGGGACGGCCTTAAGCCTACTCATTCGAGCGGAGCTAAGTCAACCTGGCGCTCT
TCTTGGAGACGACCAAATTTACAATGTCATCGTCACAGCACATGCATTTGTAATAATTTTCTTTATAGTAATACCAATTATGATCGGGGG
CTTCGGAAACTGACTGATCCCTCTTATGATTGGGGCCCCCGATATGGCCTTCCCCCGAATGAATAATATGAGCTTTTGACTCTTGCCCCC
CTCTTTTCTTCTCCTTCTTGCCTCCTCAGGCGTCGAAGCAGGTGCCGGAACAGGATGAACTGTATACCCCCCTTTAGCCGGAAACTTAGC
CCACGCAGGCGCCTCTGTTGATTTAACAATCTTTTCCCTTCACCTGGCCGGTATCTCCTCAATTCTAGGGGCCATTAATTTTATCACAAC
AATTATTAACATGAAACCGCCTGCAATTTCGCAATACCAAACCCCCCTATTTGTATGGGCAGTCCTCATTACGGCAGTACTTCTTCTCCT
CTCGCTTCCAGTTCTTGCAGCCGGAATCACGATGCTTCTTACGGATCGAAACCTTAACACTACCTTCTTCGACCCGGCAGGCGGAGGAGA
CCCAATCCTGTATCAACATCTC
>ANGBF29549-19_New_Zealand
TTCGGTGCATGAGCCGGGATAGTGGGGACGGCCTTAAGCCTACTCATTCGAGCGGAGCTAAGTCAACCTGGCGCTCTTCTTGGAGACGAC
CAAATTTACAATGTCATCGTCACAGCACATGCATTTGTAATAATTTTCTTTATAGTAATACCAATTATGATCGGGGGCTTCGGAAACTGA
CTGATCCCTCTTATGATTGGGGCCCCCGATATGGCCTTCCCCCGAATGAATAATATGAGCTTTTGACTCTTGCCCCCCTCTTTTCTTCTC
CTTCTTGCCTCCTCAGGCGTCGAAGCAGGTGCCGGAACAGGATGAACTGTATACCCCCCTTTAGCCGGAAACTTAGCCCACGCAGGCGCC
TCTGTTGATTTAACAATCTTTTCCCTTCACCTGGCCGGTATCTCCTCAATTCTAGGGGCCATTAATTTTATCACAACAATTATTAACATG
AAACCGCCTGCAATTTCGCAATACCAAACCCCCCTATTTGTATGGGCAGTCCTCATTACGGCAGTACTTCTTCTCCTCTCGCTTCCAGTT
CTTGCAGCCGGAATCACGATGCTTCTTACGGATCGAAACCTTAACACTACCTTCTTCGACCCGGCAGGCGGAGGAGACCCAATCCTGTAT
CAACATCTCTTCTGATTCT
```

5. **Align sequences using ClustalW** (https://www.genome.jp/tools-bin/clustalw)

```
CLUSTAL 2.1 multiple sequence alignment


GBMIN97864-17_Australia             ----------------------------------------TTCGGTGCATG
GBMIN97865-17_New_Zealand           ----------------------------------------TTCGGTGCATG
GBMIN127327-17_New_Zealand          ----------------------------------------TTCGGTGCATG
GBMIN133155-17_New_Zealand          ----------------------------------------TTCGGTGCATG
GBMIN127325-17_New_Zealand          ----------------------------------------TTCGGTGCATG
GBMIN127326-17_New_Zealand          ----------------------------------------TTCGGTGCATG
ANGBF29549-19_New_Zealand           ----------------------------------------TTCGGTGCATG
GBMIN97866-17_New_Zealand           ----------------------------------------TTCGGTGCATG
GBMIN97867-17_New_Zealand           ----------------------------------------TTCGGTGCATG
AMS174-08_Australia                 --------------------------CCTTTATTTAGTATTCGGTGCATG
FOAO2277-20_Australia               ------------AAGATATCGGCACCCTTTATTTAGTATTCGGTGCATG
ANGBF46642-19_Northwest_Pacifi      ----------------------------------------TTCGGTGCATG
ANGBF46643-19_New_Zealand           ----------------------------------------TTCGGTGCATG
GBMIN127324-17_New_Zealand          ----------------------------------------TTCGGTGCATG
GBMIN122614-17_New_Zealand          ----------------------------------------TTCGGTGCATG
FCFPW158-06_Portugal                --------------------------CCTTTATTTAGTATTCGGTGCATG
FCFPW216-06_Portugal                --------------------------CCTTTATTTAGTATTCGGTGCATG
FMVIC396-08_Australia               --------------------------CCTTTATTTAGTATTCGGTGCATG
GBGCA8530-15_Sweden                 ------------------------------TTAGTATTCGGTGCATG
ANGBF46640-19_US_West_Coast         TCAACCAACCACAAAGACATTGGCACCCTTTATTTAGTATTCGGTGCATG
ANGBF46644-19_New_Zealand           ----------------------------------------TTCGGTGCATG
TCHE024-12_US_West_Coast            ----------------------------------------------------
AMS170-08_Australia                 --------------------------CCTATACCTAATTTTCGGTGCCTG
```

6. Make phylogenetic trees:

- Calculate DNA distance: Mola_dist <- dist.dna(Mola_Aligned)
- Make tip color based on regions:

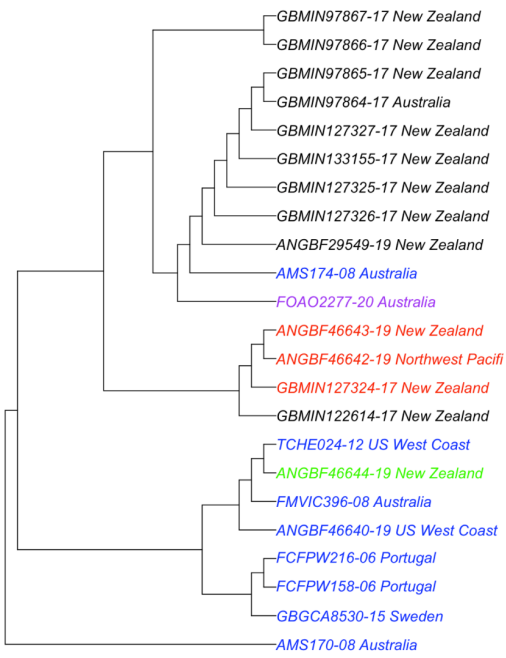
```
tipcol <- rep('black', length(Mola_UPGMA$tip.label))
Area<-c("Australia" ,"New_Zealand" ,"Northwest_Pacifi","Portugal","Sweden","US_West_Coast")
Color<-c("blue","red","green","purple","orange","black")
for (i in 1:6){
  tipcol[grep(Area[i], Mola_UPGMA$tip.label)] <- Color[i]
}
plot(Mola_UPGMA, main="UPGMA",tip.color=tipcol,use.edge.length = FALSE,cex = 1)
```

- Make tip color based on identified species:

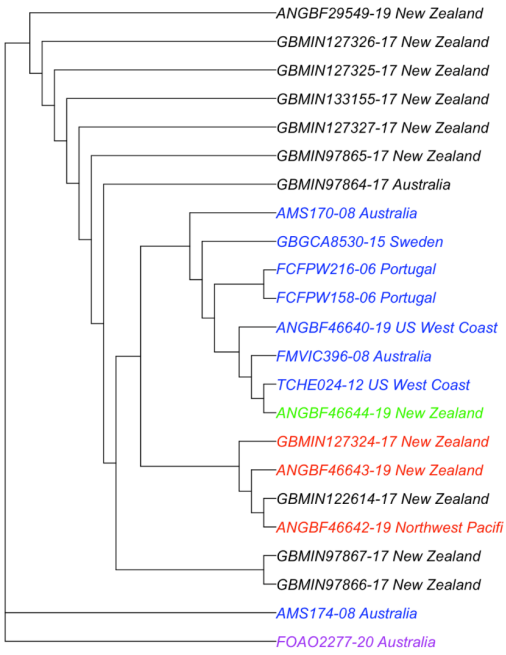
```
tipcol_species <- rep('black', length(Mola_UPGMA$tip.label))
Mola_mola<-Seq_Mola$processid[which(Seq_Mola$species_name == "Mola mola")]
Mola_spA<-Seq_Mola$processid[which(Seq_Mola$species_name == "Mola sp. A MN-2017")]
Mola_spB<-Seq_Mola$processid[which(Seq_Mola$species_name == "Mola sp. B")]
Mola_tecta<-Seq_Mola$processid[which(Seq_Mola$species_name == "Mola tecta")]

for (i in 1:8){tipcol_species[grep(Mola_mola[i], Mola_UPGMA$tip.label)] <- "blue"}
for (i in 1:3){tipcol_species[grep(Mola_spA[i], Mola_UPGMA$tip.label)] <- "red"}
tipcol_species[grep(Mola_spB, Mola_UPGMA$tip.label)] <- "green"
tipcol_species[grep(Mola_tecta, Mola_UPGMA$tip.label)] <- "purple"
```

- Draw UPGMA and NJ trees:

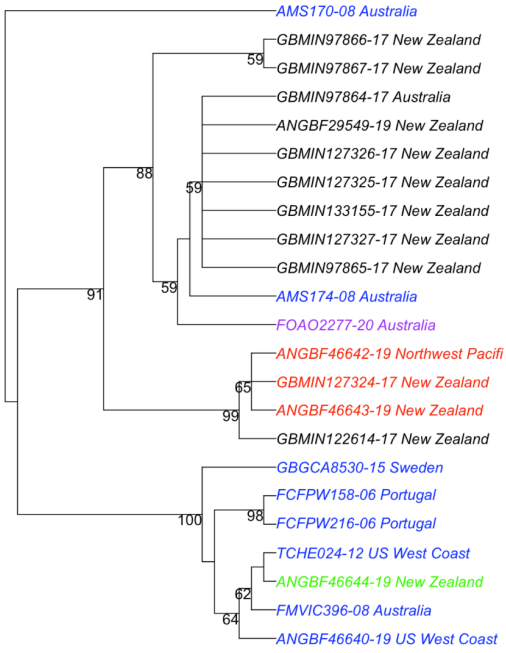
UPGMA

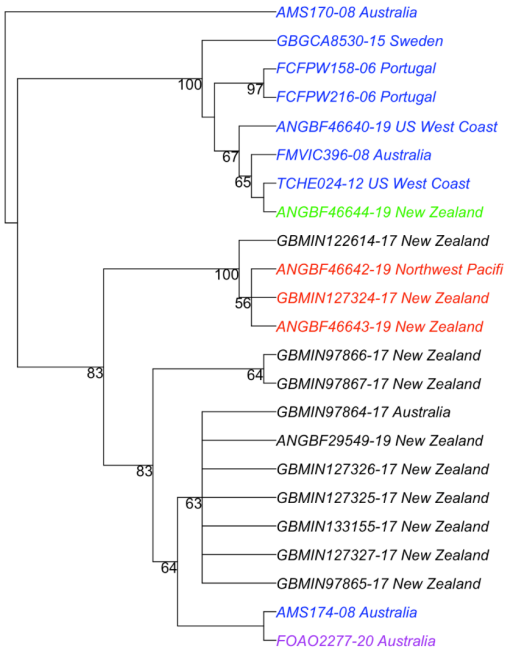


Neighbor Joining



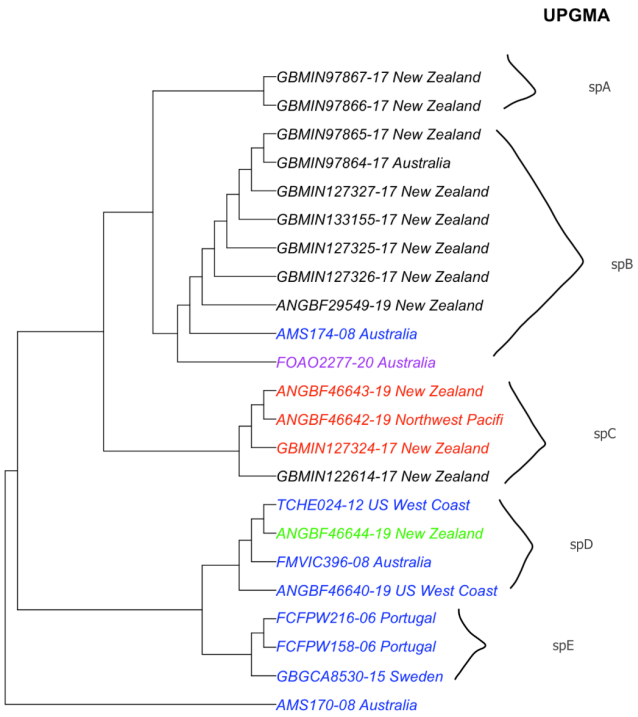
- Make trees with maximum likelihood:





7. Identifying cryptic species:

- Group sequences to find their group distances:



- Find the distances between groups of organisms by averaging the pairwise distances:

```
dif_spA_B<-mean(mola_dif$distance[c(intersect(mola_spA_c1,mola_spB_c2), intersect(mola_spA_c2,mola_spB_c1))])
dif_spA_C<-mean(mola_dif$distance[c(intersect(mola_spA_c1,mola_spC_c2), intersect(mola_spA_c2,mola_spC_c1))])
dif_spA_D<-mean(mola_dif$distance[c(intersect(mola_spA_c1,mola_spD_c2), intersect(mola_spA_c2,mola_spD_c1))])
dif_spA_E<-mean(mola_dif$distance[c(intersect(mola_spA_c1,mola_spE_c2), intersect(mola_spA_c2,mola_spE_c1))])
dif_spB_C<-mean(mola_dif$distance[c(intersect(mola_spB_c1,mola_spC_c2), intersect(mola_spB_c2,mola_spC_c1))])
dif_spB_D<-mean(mola_dif$distance[c(intersect(mola_spB_c1,mola_spD_c2), intersect(mola_spB_c2,mola_spD_c1))])
dif_spB_E<-mean(mola_dif$distance[c(intersect(mola_spB_c1,mola_spE_c2), intersect(mola_spB_c2,mola_spE_c1))])
dif_spC_D<-mean(mola_dif$distance[c(intersect(mola_spC_c1,mola_spD_c2), intersect(mola_spC_c2,mola_spD_c1))])
dif_spC_E<-mean(mola_dif$distance[c(intersect(mola_spC_c1,mola_spE_c2), intersect(mola_spC_c2,mola_spE_c1))])
dif_spD_E<-mean(mola_dif$distance[c(intersect(mola_spD_c1,mola_spE_c2), intersect(mola_spD_c2,mola_spE_c1))])
```

	dif_Mola
dif_spA_B	0.04204993
dif_spA_C	0.09747940
dif_spA_D	0.09346111
dif_spA_E	0.05213972
dif_spB_C	0.01634060
dif_spB_D	0.02842873
dif_spB_E	0.03934232
dif_spC_D	0.06226652
dif_spC_E	0.08026428
dif_spD_E	0.04423285

- **Work with Dloop gene:**

1. Get the list of Genbank accession number and location from Yukiko Yoshita et al(2009) (https://link.springer.com/article/10.1007/s10228-008-0089-3)

Mola_Dloop

Location	Sample code (collection number)	Sampling date	TL (cm)	Accession number	Group	lon	lat
Pacific_northeastern_Japan	MA-1	25 Oct 2002	52	AB191719	B	141.23	41.07
Pacific_northeastern_Japan	MA-2	25 Oct 2002	58	AB191718	B	141.23	41.07
Pacific_northeastern_Japan	MA-4	26 Oct 2002	53	AB191717	B	141.23	41.07
Pacific_northeastern_Japan	MA-5	27 Oct 2002	38	AB191716	B	141.23	41.07
Pacific_northeastern_Japan	YI-1a	17 Jun 2005	230	AB439011	B	142.04	39.29
Pacific_northeastern_Japan	YI-5a	19 July 2005	227	AB439012	B	142.04	39.29
Pacific_northeastern_Japan	YI-6	20 July 2005	>220	AB439013	A	142.04	39.29
Pacific_northeastern_Japan	YI-7	21 July 2005	ND	AB439014	A	142.04	39.29
Pacific_northeastern_Japan	YI-8a	23 July 2005	256	AB439015	B	142.04	39.29
Pacific_northeastern_Japan	YI-9a	25 July 2005	269	AB439016	A	142.04	39.29
Pacific_northeastern_Japan	YI-11	4 Aug 2005	94	AB439017	B	142.04	39.29
Pacific_northeastern_Japan	YI-12	4 Aug 2005	155	AB439018	B	142.04	39.29
Pacific_northeastern_Japan	YI-13	21 Sep 2005	38	AB439019	B	142.04	39.29
Pacific_northeastern_Japan	YI-14	27 Sep 2005	36	AB439020	B	142.04	39.29
Pacific_northeastern_Japan	YI-15	27 Sep 2005	38	AB439021	B	142.04	39.29
Pacific_northeastern_Japan	YI-16a	21 Jun 2006	277	AB439022	B	142.04	39.29
Pacific_northeastern_Japan	YI-17a	24 Jun 2006	242	AB439023	B	142.04	39.29
Pacific_northeastern_Japan	YI-18a	26 Jun 2006	194	AB439024	B	142.04	39.29
Pacific_northeastern_Japan	YI-19a	29 Jun 2006	224	AB439025	B	142.04	39.29
Pacific_northeastern_Japan	YI-20a	30 Jun 2006	226	AB439026	B	142.04	39.29
Pacific_northeastern_Japan	YI-21a	22 July 2006	219	AB439027	B	142.04	39.29

2. Extracted dna sequences from genBank with ape:

```
seq_Mola_Dloop<-read.GenBank(Mola_Dloop$`Accession number`)
write.fasta(sequences = as.character.DNAbin(seq_Mola_Dloop), names =
as.list(paste(Mola_Dloop$`Accession number`,Mola_Dloop$Location,sep = "_")),file.out =
"Seq_Mola_Dloop.fasta" )
```

3. **Align sequences using ClustalW** (https://www.genome.jp/tools-bin/clustalw)

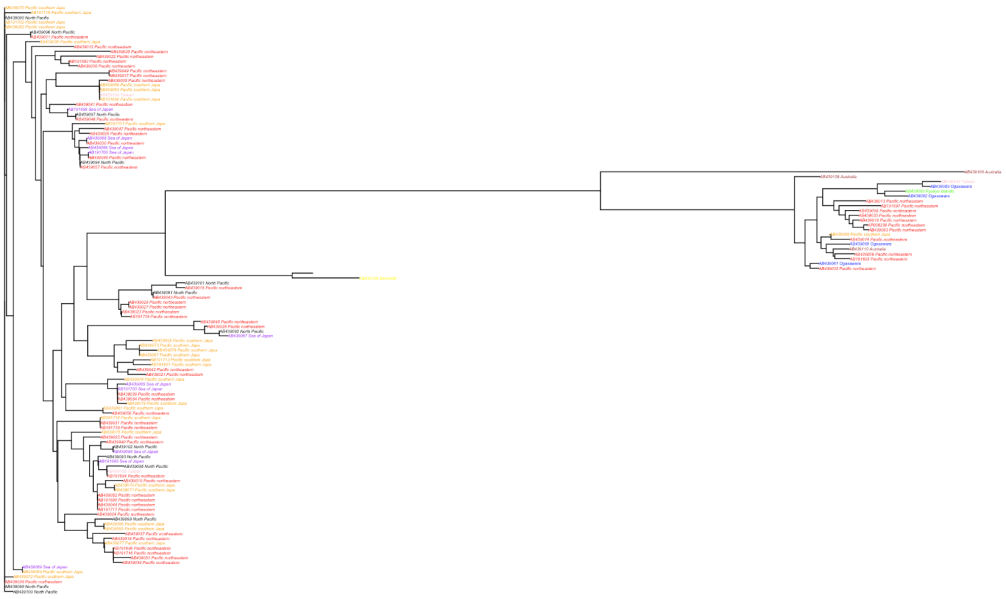
```
AB191719_Pacific_northeastern_      CACCATATATATATATGCACCATTAATGAATATTATCTCTAGGACAATAA
AB439023_Pacific_northeastern_      CACCATATATATATATGCACCATTAATGAATATTATCTCTAGGACAATAA
AB439027_Pacific_northeastern_      CACCATATATATATATGCACCATTAATGAATATTATCTCTAGGACAATAA
AB439029_Pacific_northeastern_      CACCATATATATATATGCACCATTAATGAATATTATCTCTAGGACAATAA
AB439018_Pacific_northeastern_      CACCATATATATATATGCACCATTAATGAATATTATCTCCAGGACAATAA
AB439101_North_Pacific              CACCATATATATATATGCACCATTAATGAATATTATCTCCAGGACAATAA
AB439043_Pacific_northeastern_      CACCATATATATATATGCACCATTAATGAATATTATCTCTAGGACAATAA
AB439091_North_Pacific              CACCATATATATATATGCACCATTAATGAATATTATCTCTAGGACAATAA
AB439082_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439090_North_Pacific              CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB191702_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439070_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439038_Pacific_northeastern_      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439098_North_Pacific              CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439072_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439100_North_Pacific              CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439011_Pacific_northeastern_      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439096_North_Pacific              CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB191715_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAACAA
AB439064_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439089_Sea_of_Japan_              CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439080_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439017_Pacific_northeastern_      CACCATATATATATATGTACCATTAATGAATACTATCTCTAGGACAATAA
AB439049_Pacific_northeastern_      CACCATATATATATATGTACCATTAATGAATACTATCTCTAGGACAATAA
AB439035_Pacific_northeastern_      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439063_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB439069_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
AB191690_Pacific_southern_Japa      CACCATATATATATATGTACCATTAATGAATATTATCTCTAGGACAATAA
```

4. Make phylogenetic trees:

- Calculate DNA distance: Mola_Dloop_dist <- dist.dna(Mola_Dloop_Aligned)
- Make tip color based on regions:

```
#Color tip based on location
tipcol <- rep('black', length(Mola_Dloop_UPGMA$tip.label))
Area<-c("Pacific_northeastern" ,"Ogasawara" ,"Pacific_southern","Ryukyu","Sea_of_Japan",
        "North_Pacific","Taiwan","Denmark","Australia","UK")
Color<-c("red","blue","orange","green","purple","black","pink","yellow","brown","white")
for (i in 1:10){
  tipcol[grep(Area[i], Mola_Dloop_UPGMA$tip.label)] <- Color[i]
}
```

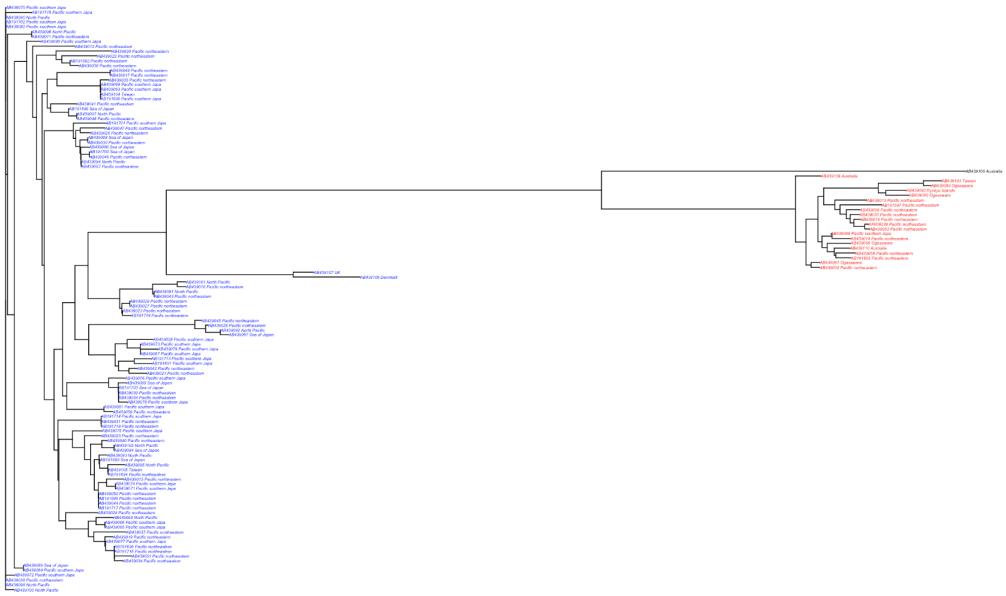
Neighbor Joining



- Make tip color based on groups:

```
#Color tip based on groups
tipcol_group <- rep('black', length(Mola_Dloop_UPGMA$tip.label))
groupA<-Mola_Dloop$`Accession number`[which(Mola_Dloop$Group == "A")]
for (i in 1:20){
  tipcol_group[grep(groupA[i], Mola_Dloop_UPGMA$tip.label)] <- "red"
}

groupB<-Mola_Dloop$`Accession number`[which(Mola_Dloop$Group == "B")]
for (i in 1:101){
  tipcol_group[grep(groupB[i], Mola_Dloop_UPGMA$tip.label)] <- "blue"
}
```

Neighbor Joining



Red: group A, Blue: Group B