Viral Bioinformatics 2024
Topics: MSA, Phylogeny, Virus Typing & Recombination detection
Hands-on exercises

Multiple Sequence Alignments:

1. **Module Developers & Assistants:** Dr. Urmila Kulkarni-Kale & Mr. Sanket Limaye

2. **Learning outcomes:**
   At the end of the session, participants are expected to understand –
   - sequence alignment concepts, pairwise sequence alignment (local and global), multiple sequence alignment algorithms, and their applications in virus bioinformatics.
   - fundamentals of sequence analysis, the basis of scoring matrices, the importance of gap penalties, etc.
   - the steps in the MAFFT algorithm, familiarity with sequence input/output formats, how to carry out MSA, interpretation of results, etc.

3. **Background (MSA):**
   Multiple sequence alignment (MSA) is one of the most frequently used bioinformatics techniques for aligning nucleotide and protein sequences. MSA often serves as a starting point for analyzing viral sequences to identify regions of similarity/differences and indicate potential functional, structural, and evolutionary relationships. Several algorithms and software tools are available for MSA, of which MAFFT will be discussed in detail using a case study.

4. **Key resources:**
   MAFFT publication: Katoh et al., 2009; DOI: 10.1007/978-1-59745-251-9_3
   Phylogeny review: Chao et al., 2022; DOI: 10.3390/biom12040546
   Lecture on MSA overview & MAFFT: To be shared.
   Software: MAFFT for Linux - a multiple sequence alignment program (cbrc.jp)
   MAFFT documentation & help:
   https://mafft.cbrc.jp/alignment/software/manual/manual.html

5. **Method:**
   Case study: Perform multiple genome alignment of Dengue Virus serotype 4 (DENV4) isolates (DENV4_seq.fas) using MAFFT. What is the percentage of identical sites in the alignment? Save the genome alignment in fasta and Clustal formats.

   Input dataset:
   MSA_phylogeny_RDP\1_mafft\input\ DENV4_seq.fas

   Output files:
   MSA_phylogeny_RDP\1_mafft\output\ DENV4_aln.fas and DENV4_clustal_aln.fas

   Commands:

   For fasta format:

   ```
   mafft DENV4_seq.fas > DENV4_aln.fas
   ```

   For clustal format

```
mafft --clustalout DENV4_seq.fas > DENV4_clustal_aln.fas
```

For identifying the conserved nucleotides:

```
grep -o “*” DENV4_aln.fas | wc -l
```

Molecular Phylogeny and virus typing:

1. **Module Developers & Assistants:** Dr. Urmila Kulkarni-Kale & Mr. Sanket Limaye

2. **Learning outcomes:**
   At the end of this session, participants are expected to develop an understanding of –
   - concepts in phylogeny and evolution.
   - various methods for phylogenetic analysis.
   - tree visualization and interpretation.
   - Various approaches for serotyping and genotyping of virus isolates.

3. **Background (Phylogeny and typing):**
   Molecular phylogeny analyses are performed not only to carry out basic research, such as understanding the genetic diversity of viruses, tracking the spread of viral diseases, and studying the evolution and diversification of viruses, but also for translation research to design drugs and vaccines.
   Typing virus isolates is an essential prerequisite for many downstream analyses. Computational approaches for typing involve phylogenetic analysis of the isolates with unknown types and the known ones (reference set). The type is assigned based on clustering proximity of the unknown with the known, which demands human interventions. On the other hand, alignment-free approaches automate the typing process and circumvent a few limitations of alignment-based MPA. An overview of typing methods and a few case studies will be presented.

   Phylogeny
   Phylogeny publication: Link to PDF
   Lecture on Phylogeny and typing: To be shared
   Software:
   - IQ-Tree: Getting Started (iqtree.org);
   - Figtree: Releases · rambaut/figtree (github.com)

   IQ-Tree documentation & help: http://www.iqtree.org/doc/Quickstart

   Typing
   Webserver: Genome Detective (DENGUE VIRUS TYPING TOOL):
   https://www.genomedetective.com/app/typingtool/dengue/
   RTD-based Dengue Typer Server: The URL is to be shared.

4. **Method:**
   A. Use the genome alignment of DENV4 isolates and generate a whole-genome phylogenetic tree with the help of IQTREE. Select the best nucleotide substitution model that fits the data using ModelSelector. Reconstruct maximum likelihood-based phylogeny with 1000 bootstrap replicates using the Ultrafast method available in IQTREE. What are the number of invariant sites and parsimoniously informative sites? Which nucleotide model best fits the data provided? Using the consensus tree, find the clusters where Indian isolates are observed. Also, check if the isolates of the same genotype cluster together.

Input dataset: \MSA_phylogeny_RDP\2_iqtree\input\ DENV4_aln.fas
Output files: \MSA_phylogeny_RDP\2_iqtree\output\

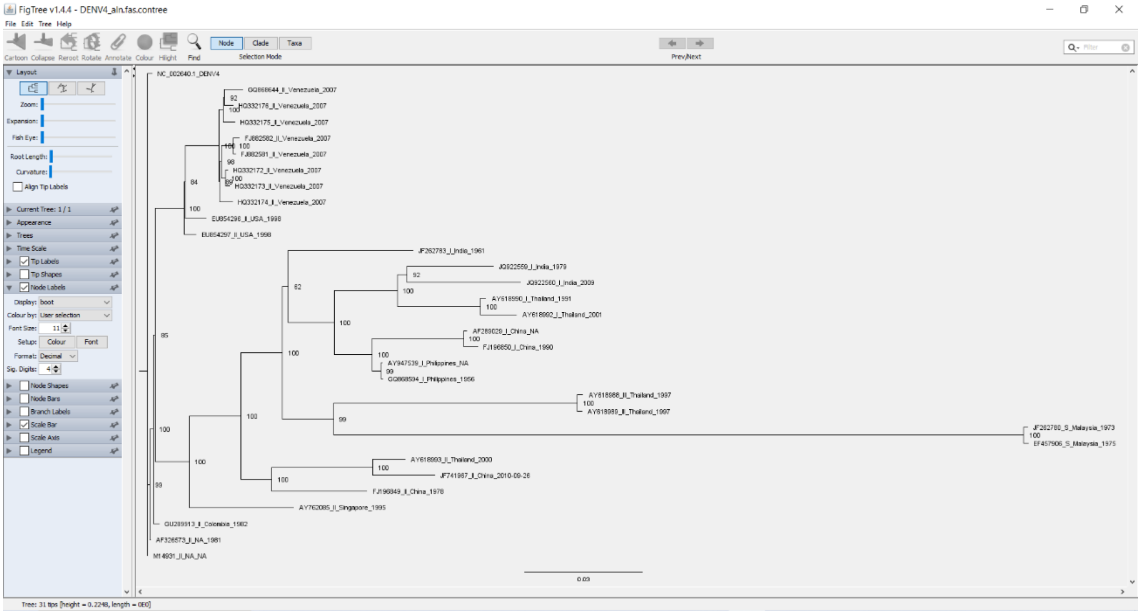
Commands:

<u>IQTREE</u>

```
iqtree -s DENV4_aln.fas -bb 1000
```

<u>Figtree</u>

Open ".contree" file > select node labels > select "Display: bootstrap"



FigTree v1.4.4 - DENV4_aln.fas.contree

☐ Align Tip Labels
Current Tree: 1 / 1
☑ Tip Labels
☐ Tip Shapes
☑ Node Labels
Display: boot
Colour by: User selection
Font Size: 11
Format: Decimal
Sig. Digits: 4
☐ Node Shapes
☐ Node Bars
☐ Branch Labels
☑ Scale Bar
☐ Scale Axis
☐ Legend

Tree: 31 tips [height = 0.2248, length = 0E0]

B. Perform genotyping of the unknown DENV4 isolates using the methods mentioned below. Observe if the methods assign different genotypes to both the isolates.

i. View the "DENV4_genotyping_aln.contree" file in Figtree software and assign genotypes to the two isolates (DENV4_A and DENV4_B) based on the clustering pattern.

Input dataset:
\MSA_phylogeny_RDP\4_test\ DENV4_genotyping_aln.contree

Use the file "DENV4_unknown.fas" as input to the DENGUE VIRUS TYPING TOOL available at the GENOME DETECTIVE web server and assign genotypes to the two isolates (DENV4_A and DENV4_B).

URL: https://www.genomedetective.com/app/typingtool/dengue/

Input dataset: \MSA_phylogeny_RDP\4_test\ DENV4_unknown.fasta

Recombination Detection:

1. **Module Developers & Assistants:** Dr. Urmila Kulkarni-Kale & Mr. Sanket Limaye

2. **Learning outcomes:**
   At the end of the session, participants are expected to develop an understanding of –
   - various methods available for recombination detection
   - why and how to carry out recombination analysis

3. **Background (Recombination Detection):**
   A recombination event occurs when genetic material from two or more viruses mixes during replication. It is one of the major driving forces behind virus diversification and the emergence of novel strains.

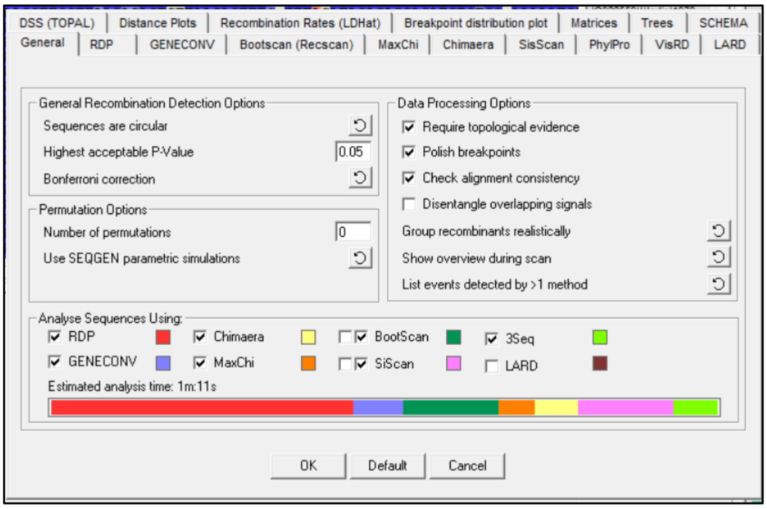
4. **Key resources (Recombination Detection):**
   RDP4 publication: https://academic.oup.com/ve/article/1/1/vev003/2568683
   Lecture on Recombination: To be shared.
   Software: **Recombination (RDP4 software):** RDP home page (uct.ac.za)
   a. Download and install **RDP4.101** (we use and prefer the **Windows** version; the link for the Linux version is available on the same page)
   b. Version 4.101 is more stable than RDP5.55

   RDP4 documentation and help: http://web.cbio.uct.ac.za/~darren/RDP4Manual.pdf

5. **Method:**
   A. Use the genome alignment of DENV4 isolates and detect the recombination events using the RDP4 program. Select the p-value cutoff and threshold of the minimum number of methods that positively detect a recombination event. List the number of recombination events that were detected. How many of them were considered significant? From the significant recombination hits obtained, report the major parent and minor parent along with the breakpoints of the sequences.

   Input dataset: \MSA_phylogeny_RDP\3_RDP4\input\ DENV4_aln.fas
   Output files: \MSA_phylogeny_RDP\3_RDP4\output\ DENV4_RDP.csv

   Commands:
   RDP4
   Open the alignment file > Set the "Highest acceptable P-value" > Select " List events detected by > 3 " > Select the different scanning methods > RUN



General Recombination Detection Options

Sequences are circular

Highest acceptable P-Value 0.05

Bonferroni correction

Permutation Options

Number of permutations 0

Use SEQGEN parametric simulations

Data Processing Options

☑ Require topological evidence

☑ Polish breakpoints

☑ Check alignment consistency

☐ Disentangle overlapping signals

Group recombinants realistically

Show overview during scan

List events detected by >1 method

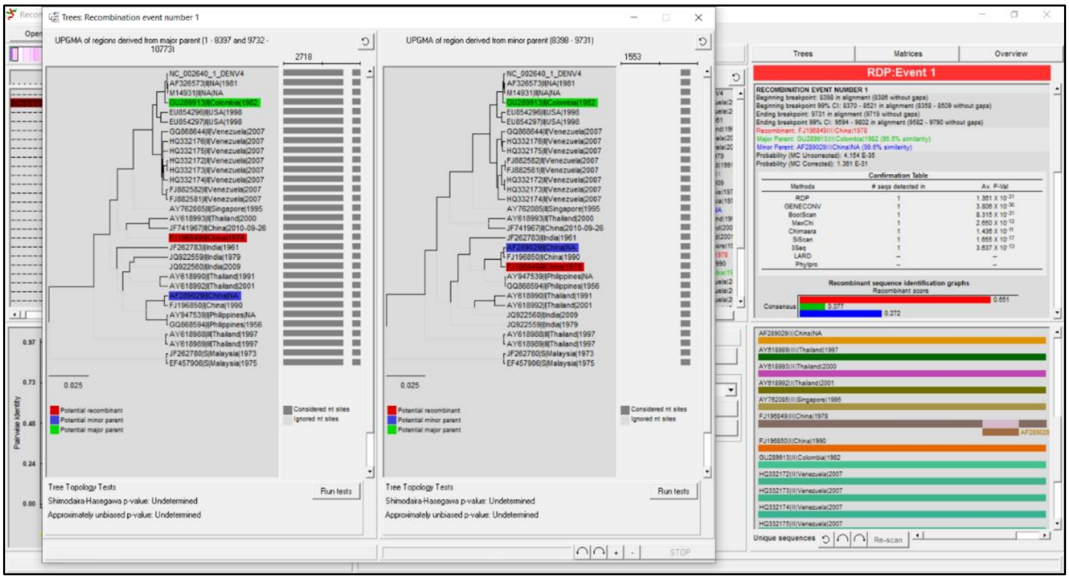
Analyse Sequences Using:

☑ RDP 🟥 ☑ Chimaera 🟨 ☐☑ BootScan 🟩 ☑ 3Seq 🟩

☑ GENECONV 🟦 ☑ MaxChi 🟧 ☐☑ SiScan 🟪 ☐ LARD 🟫

Estimated analysis time: 1m:11s

Select the “Overview” in the top right corner > Select an event > Select the “Recombination info” tab.



Trees: Recombination event number 1
UPGMA of regions derived from major parent (1 - 8397 and 9732 - 10773)
UPGMA of region derived from minor parent (8398 - 9731)
Potential recombinant
Potential minor parent
Potential major parent
Considered nt sites
Ignored nt sites
Run tests
Trees
Matrices
Overview
RDP:Event 1

Tree Topology Tests

Shimodaira-Hasegawa p-value: Undetermined

Approximately unbiased p-value: Undetermined

RECOMBINATION EVENT NUMBER 1

Beginning breakpoint: 8398 in alignment (8386 without gaps)

Beginning breakpoint 99% CI: 8370 - 8521 in alignment (8358 - 8509 without gaps)

Ending breakpoint: 9731 in alignment (9719 without gaps)

Ending breakpoint 99% CI: 9594 - 9802 in alignment (9582 - 9790 without gaps)

Recombinant: FJ196849|||China|1978

Major Parent: GU289913|||Colombia|1982 (95.5% similarity)

Minor Parent: AF289029|||China|NA (99.6% similarity)

Probability (MC Uncorrected): 4.154 E-35

Probability (MC Corrected): 1.361 E-31

Confirmation Table

Methods	# seqs detected in	Av. P-Val
RDP	1	1.361×10^{-31}
GENECONV	1	3.806×10^{-30}
BootScan	1	8.315×10^{-31}
MaxChi	1	2.650×10^{-12}
Chimaera	1	1.436×10^{-11}
SiScan	1	1.655×10^{-17}
3Seq	1	3.637×10^{-13}
LARD	–	–
Phylpro	–	–

Recombinant sequence identification graphs

Save the results as “.rdp” and .csv file.

B. File containing genome sequences of Bovine Coronavirus (BCoV_genome.fas) is provided (\MSA_phylogeny_RDP\3_RDP4\Exercise). Perform alignment and detect recombination events using the RDP4 program. Report various statistics such as significant recombination hits obtained, the major parent and minor parent of identified recombinants along with the breakpoints of the sequences.