GLUE and MADDOG

Using GLUE

GLUE is a great tool for initial analysis and finding relevant sequences and studies!

It can be accessed at: http://rabv-glue.cvr.gla.ac.uk/#/home

Task 1

Use the `nig-af2-seqs.fasta` file in the `Home/RAGE-workshop-2024/day4/Phylogenetics/Data` folder.

Go to RABV-GLUE (http://rabv-glue.cvr.gla.ac.uk/#/home) and click Analysis -> Genotyping and Interpretation

Click Add Files and select the fasta file you just downloaded

Click Submit and wait while the sequences are analysed

This might take a few minutes - especially with everyone using it at once! While you're waiting, open up another RABV-GLUE tab and start task 2.

When the analysis is complete, explore the 3 sections (summary, genome visualisation, and phylogenetic placement) to answer the following questions:

1. What clade do these sequences belong to?
2. Are these whole genome sequences? How can you tell?

Compare the G gene of sequence 3 to the master reference. Hint: you'll need to press 'update' each time you change any settings!

3. What is the effect of the A->T substitution in position 4?
4. Does sequence 10 also have this substitution?
5. Now compare the sequences to the Africa-2 reference instead. What happens to this substitution? What do you think this means?

Task 2

In RABV-GLUE, click Sequence Data -> NCBI Sequences by Clade

Select 'Africa-2' from the list at the top.

Using the 'Filters' tab, try to answer the following questions about the Africa-2 clade sequences:

1. How many sequences are there in total?
2. How many sequences have at least 90% coverage of the whole genome? What percentage is this of all the sequences?
3. How many sequences are from dogs (Canis familiaris)? What percentage is this of all the sequences?
4. How many sequences are from the Nigeria?
5. Combine all these filters to find all the whole genome dog sequences from the Nigeria.

This is how we can find all relevant sequences to our study! If we wanted to download these, we could click Download -> Download sequences and Download -> Download metadata (we don't need to do this now).

6. What is the title of the study sequence KC196743 is part of?

This is how GLUE also lets us find papers relevant to our dataset!

Using MADDOG

Locate the MADDOG folder within the SSD/storage device provided. This should be in `Home/workshop_dir/backup/MADDOG`

Access the terminal.

Navigate inside the MADDOG directory using the terminal.

Enter the MADDOG conda environment using `conda activate MADDOG`

Make a new folder **in the MADDOG directory**, and name this with your first name. Note: There can be no spaces or special characters in this name! Hyphens and underscores are allowed

Copy the `nig-af2-seqs.fasta` and `MADDOG.csv` files from `Home/RAGE-workshop-2024/day4/Phylogenetics/Data` into this folder.

Assignment

We're first going to run assignment! This is a quick look at the sequences prior to any further analysis. This only needs the fasta file, but we've also copied in the metadata for our next steps. Assignment will compare our sequences to a references of sequences, and use this to identify which lineages sequences belong to.

Ensure you're in the MADDOG folder, and the MADDOG conda environment.

Type `sh assignment.sh`

It will ask you if you've pulled the repository. Type `Y` and press enter.

When prompted, enter the name of the folder you just made.

The sequences are now being assigned! When this is complete, you'll find a file called (FOLDER_NAME)_assignment.csv in the folder you made. Open this!

1. How many lineages are present in the dataset we analysed? What is the most common lineage here?
2. How does this compare to our analysis in RABV-GLUE?
3. Where have these lineages been previously seen?
4. Checking for new lineages is appropriate if there are at least 10 sequences from the same lineage. Should we check for new lineages here?

Now delete the _assignment.csv file. We are going to progress to lineage designation!

Designation

As we have lots of sequences from the same lineage, we want to check if this means our dataset contains new, undocumented lineages! This involves lineage designation. This is a much more lengthy procedure, as we need to check that sequences fit all of the rules of lineage designation (genome coverage, number of sequences, shared mutations, statistical support). Therefore, this involves comparing the sequences to existing sequences, building a tree, and testing for all of these rules.

First, have a look at the `MADDOG.csv` file. This is how our metadata needs to be formatted.

1. What are the 4 column headers we need?
2. Where might we find the information to put in the 'assignment' column for each sequence?

Ensure you're in the MADDOG folder, and the MADDOG conda environment.

Type `sh designation.sh`

When prompted, enter the name of the folder you just made.

Lineage designation is now taking place!

It would take too long to run (maybe 30 mins - 1 hour!), so we'll stop the run by pressing ctrl + c and instead look at a run I completed earlier on this dataset in the next session.