Tutorial

Phylogenetic Trees and Metadata

November 21, 2017

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

Phylogenetic Trees and Metadata

This tutorial briefly introduces the reconstruction of phylogenetic trees and visualization using the tree viewer. The main focus in this tutorial is the visualization of metadata as a powerful tool for analyzing your data.

The features demonstrated in this tutorial include:

- Alignment of sequences
- Reconstruction of phylogenetic trees
- Introduction to the tree viewer
- Metadata and visualization of these
- Grouping of nodes
- Labeling of subtrees

Visualizing metadata on phylogenetic trees is an easy and flexible way to view different types of metadata in context. The phylogenetic tree viewer provides many features for customizing tree visualization such as:

- Circular and radial layouts
- Adjustment of node size and branch width
- Curved branches
- Grouping of nodes and visualization of these
- Visualization of metadata

Example Dataset

This tutorial involves a dataset that contains sequences and metadata from the viral hemorrhagic septicemia virus (VHSV). Ths virus is a fish novirhabdovirus (negative stranded RNA virus) with an unusually broad host spectra: it has been isolated from more than 80 fish species in locations around the Northern hemisphere. This dataset contains two files. One contains sequences encoding the viral surface glycoprotein of VHSV in fasta format. The other contains key metadata information in an excel spreadsheet.

Sequence data:
http://resources.qiagenbioinformatics.com/testdata/phylogeny_module_tutorial_data/Phylogeny_module_example_data.fa
Metadata:
http://resources.qiagenbioinformatics.com/testdata/phylogeny_module_tutorial_data/Phylogeny_module_tutorial_meta_data.xls

The metadata includes the following categories, which are used later in this tutorial.

- **Sequence** - The nucleotide sequence of a surface glycoprotein.

- **Strain** - The virus strain.
- **Host** - The fish from which the sample was obtained.
- **Water** - The type of water the fish lives in.
- **Country** - The country where the fish was caught.
- **ACCNo** - The accession number of the virus genome.
- **Year** - The year in which the sample was obtained.

Importing the Sequence Data

To download the example data, click the two links above. The sequence data can then be imported into the CLC Workbench by following these steps:

1. Start up the Workbench and use the import tool at:

   File | **Import (**  **)** | **Standard Import (**  **)**

2. Choose the file "phylogeny_module_example_data.fa". Ensure the import type under **Options** is set to **Automatic import** (figure 1). Click on the button labeled **Next**.

3. Select the location where you want to store the imported sequences and click on the button labeled **Finish**.

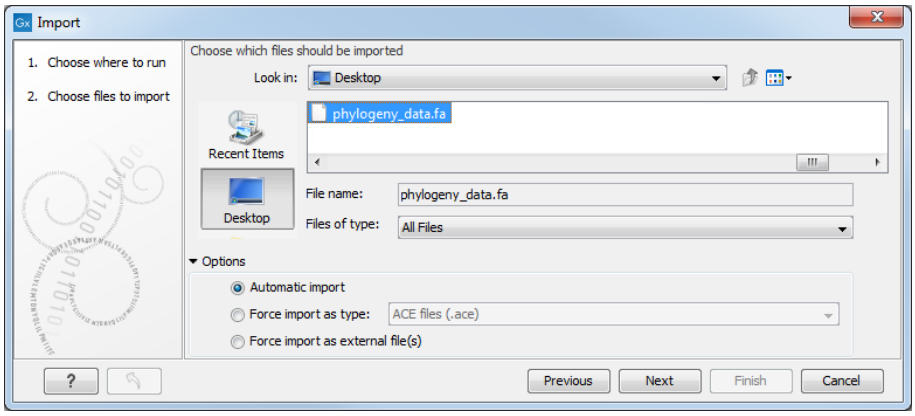

Figure 1: *Import of the virus sequence data.*

Aligning Sequences

Most phylogenetic reconstruction methods require a multiple alignment of the input sequences, which is used to reconstruct the corresponding phylogenetic tree.

It is, however, possible to reconstruct a phylogenetic tree without first creating a multiple alignment of the input sequences. For example, the **K-mer Based Tree Construction** tool available in the Workbench utilizes a kmer- based approach to estimate pair-wise distances between the input sequences. This distance estimate can then be used to reconstruct the tree using either the UPGMA or Neighbor-Joining algorithms. If the input dataset is large and/or

contain very long sequences, it can be an advantage to use this tool for reconstructing trees to avoid the time consuming task of creating a multiple alignment.

In this tutorial, we will create a multiple alignment, on which the tree wil be based. To create an alignment of the imported sequences:

1. Start the **Create alignment** tool by going to:

   Toolbox | **Classical Sequence Analysis** ( ) | **Alignments and Trees** ( ) | **Create Alignment** ( )

2. Select the imported sequences **Phylogeny_module_example_data** and click on the button labeled **Next**.

3. Use the default gap cost settings, but select the **Less accurate (fast)** option in the Alignment section (figure 2). Click on the button labeled **Next**.

4. Choose the option **Save** and select the location where you wish to store the alignment before clicking on the button labeled **Finish**.

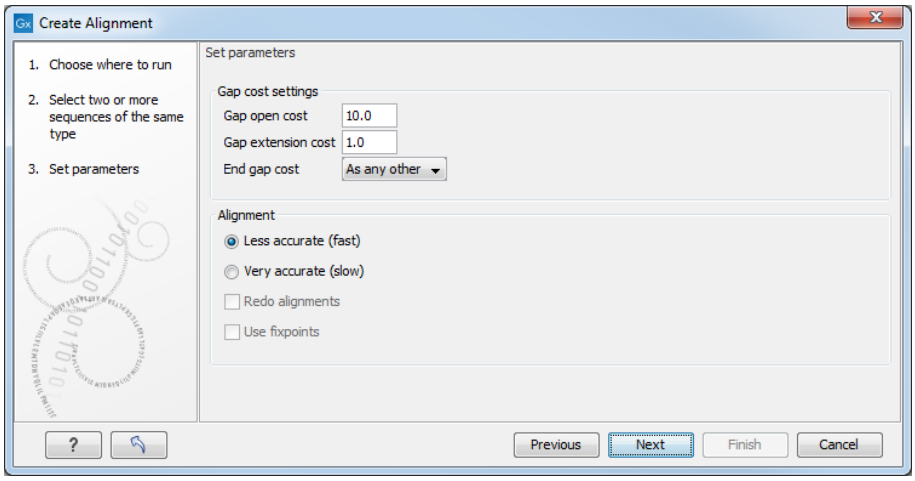

Figure 2: *Creating a multiple alignment of the sequence data.*

To check your alignment, open the generated output from the destination where you chose to save it. Alternatively, the output can be opened by clicking on the small arrow next to the process bar and choosing **Find Results** as shown in figure 3):

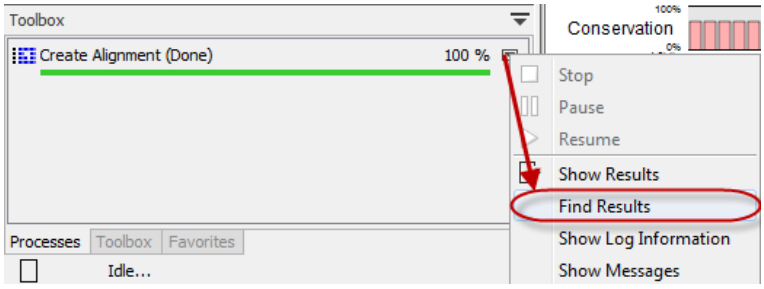

Figure 3: *Find and open the output that was generated.*

Reconstructing the Tree

A phylogenetic tree can now be reconstructed Using the multiple sequence alignment created in the previous step. There are two tools that can be used for this in the Workbench. Both are found under the Alignments and Trees section of the Toolbox. They are:

- **Create Tree** - This tool constructs trees using one of two distance based methods:
  - UPGMA
  - Neighbor-Joining
- **Maximum Likelihood Phylogeny** - This tool reconstructs phylogenetic trees using a maximum likelihood approach.

In this tutorial we will use the Neighbor-Joining method. This is a fast and fairly accurate method for phylogenetic reconstruction.

1. Start the **Create Tree** tool:

   Toolbox | **Classical Sequence Analysis** ( ) | **Alignments and Trees** ( ) | **Create Tree** ( )
2. Select the multiple alignment made in the previous section.
3. Work through the Wizard, leaving all options set to the default values as in figure 4.
4. Choose to save the tree.

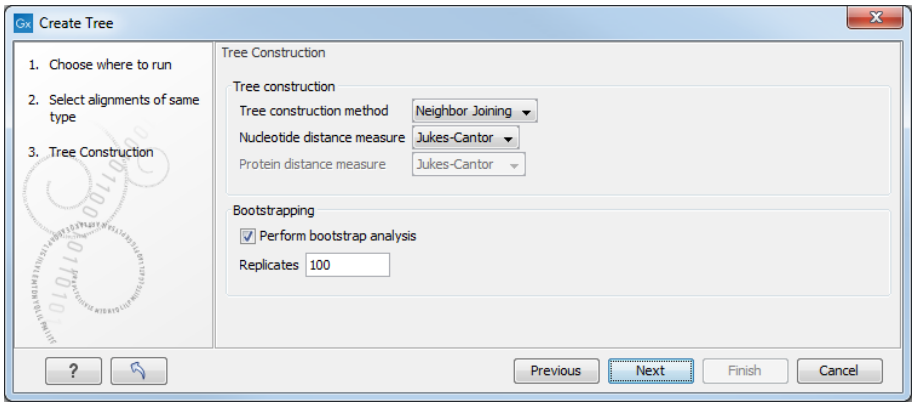

Figure 4: *Reconstruct a neighbor-joining tree with 100 bootstrap replicates.*

It can take some time to reconstruct the tree because the default parameters include bootstrapping with 100 replicates. That is, using this option, the tree is reconstructed 101 times and the results are used to test the confidence of each internal node in the resulting tree.

Visualizing the Tree

To open a tree in the tree viewer, double click the tree object in the **Navigation Area**. The viewer can display the tree in five different layouts:

- **Phylogram** - The tree is displayed as a rooted tree where branch lengths correspond to the computed lengths.
- **Cladogram** - The tree is displayed as a rooted tree where branch lengths are ignored and all leaves are aligned to the right.
- **Circular Phylogram** - Same as a phylogram but with the leaves in a circular layout.
- **Circular Cladogram** - Same as a cladogram but with the leaves in a circular layout.
- **Radial** - The tree is displayed as an unrooted tree where branch lengths correspond to the computed lengths.

In this tutorial we will primarily use the phylogram layout, but we would encourage you to try the different layouts while going through this tutorial.

Zooming and scrolling in the tree view work in the same way as other viewers/editors in the Workbench. When zooming in on a large tree, one can easily lose orientation. Hence, in order to reduce the need for zooming out again, the new tree viewer includes a small minimap in the right side panel (figure 5). The minimap shows the full tree, with a grey rectangle highlighting the area of the tree that currently is shown in the View Area. The grey rectangle can be dragged around in the minimap with the mouse, and as this is done, the location highlighted in the minimap and the part of the tree shown in the View Area are synchronized. The minimap is particularly useful to understand what section of a large tree is visible the View Area.

Try clicking on the minimap and dragging the grey rectangle around to see how the tree is displayed in the View Area.

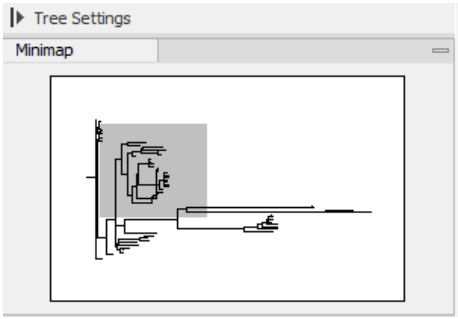

Figure 5: *The minimap. The grey rectangle represents the area of the tree that is visible in the View Area.*

The right side panel contains options for visualization of the tree and associated information, such as adjusting tree colors, node shape, node size and the shape of branches. For more details on these options, please refer to the manual section: `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Tree_Settings.html`.

Bootstrap Values

In the right side panel, under **Bootstrap settings**, enable the option **Show bootstrap values** (figure 6). To help get an overview of the confidence values we can highlight all branches that lead to a node with a bootstrap value of e.g. $\geq$95%.

Now each internal node is labeled with a value between 0 and 100 (figure 7). These values represent confidence levels, where a high confidence indicates a clade strongly supported by the

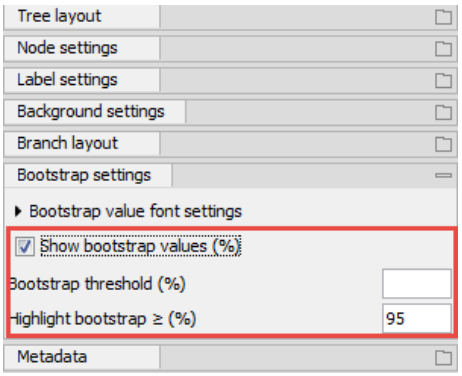

Figure 6: *Enabling bootstrap values.*

data from which the tree was constructed. Bootstrap values are useful for identifying clades in the tree where the topology (and branch lengths) should not be trusted.

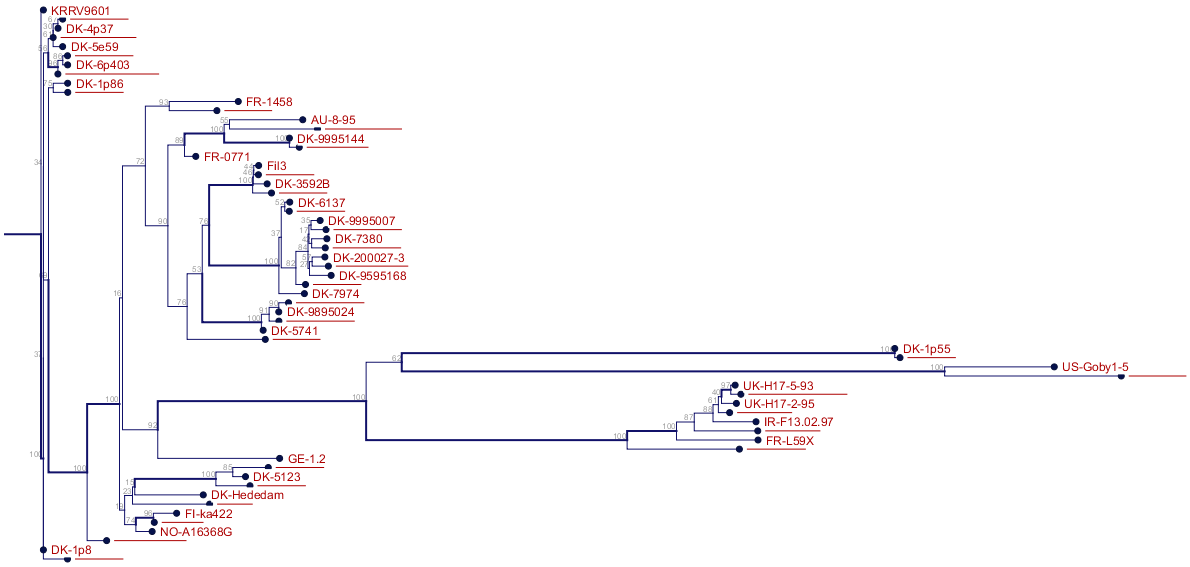

Figure 7: *The tree where edges corresponding to nodes with bootstrap values ≥95% are highlighted.*

Another way to visualize bootstrap values is to collapse internal nodes with bootstrap values under a certain threshold. After removing a node the child nodes will be connected to the parent of the collapsed node. This creates a multifurcating tree containing only high confidence nodes. To enable this visualization enter a threshold in the field labeled **Bootstrap threshold** under **Bootstrap settings** in the right side panel. Figure 8 shows an example of this.

Metadata

When a tree is reconstructed, some mandatory metadata will be added to nodes in the tree. These metadata are special in the sense that the tree viewer has specialized features for visualizing the data and some of them cannot be edited. Examples of mandatory metadata fields are:

- **Name** - The node name.
- **Branch length** - The length of the branch which connects a node to the parent node.
- **Bootstrap value** - The bootstrap value for internal nodes.

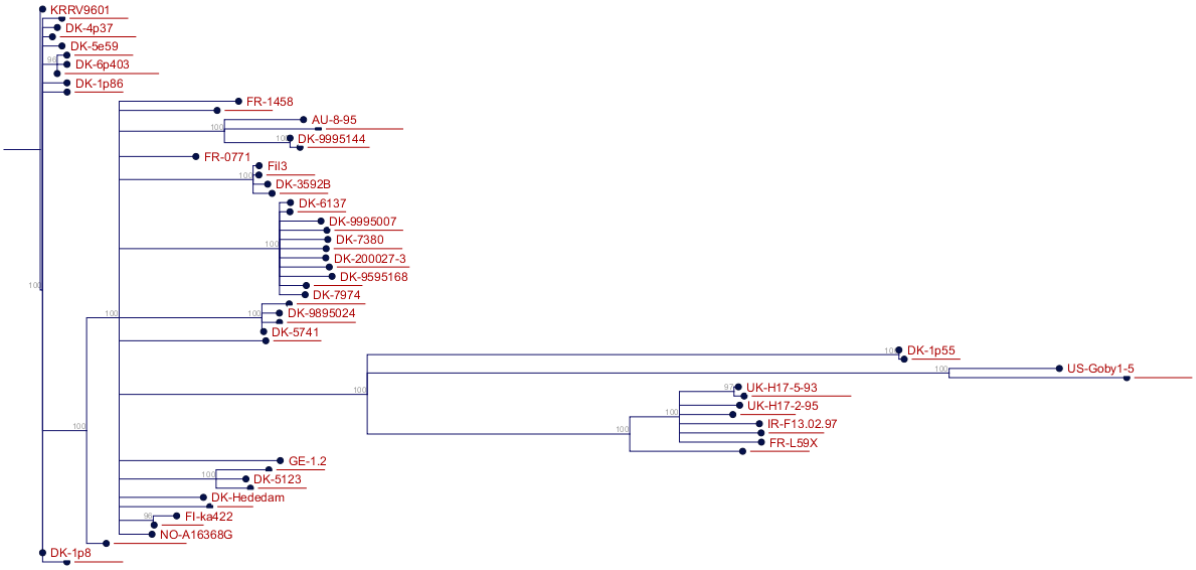

Figure 8: *The tree where edges corresponding to nodes with bootstrap values <95% are collapsed.*

- **Size** - The length of the sequence which corresponds each leaf node. This only applies to leaf nodes.
- **Start of sequence** - The first 50bp of the sequence corresponding to each leaf node.

To view a table of metadata associated with a tree click the **Show Table** button at the bottom of the View Area (figure 9). To open the metadata table in a split view, such that both the tree and the table are visible at the same time and the table information is linked to the tree view, click the **Show Table** button while holding down the Ctrl key (or ⌘ for Mac).

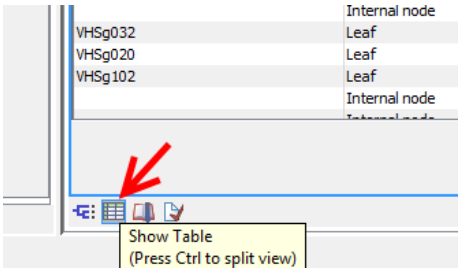

Figure 9: *The metadata table button.*

This opens a table where each row represents a node, and the columns contain different categories of metadata. When the table and tree views are open and linked in this way, clicking on a row in the table highlights the relevant location in the tree (figure 10).

Importing Metadata

Additional metadata can be imported by clicking the **Import Metadata** button below at the bottom of the table view (figure 10). The file "Phylogeny_module_example_meta_data.xls" you downloaded earlier is an Excel format file containing seven categories of metadata. These categories are listed near the beginning of this tutorial. To import the metadata contained in this file:

1. Click the **Import Medata** button.

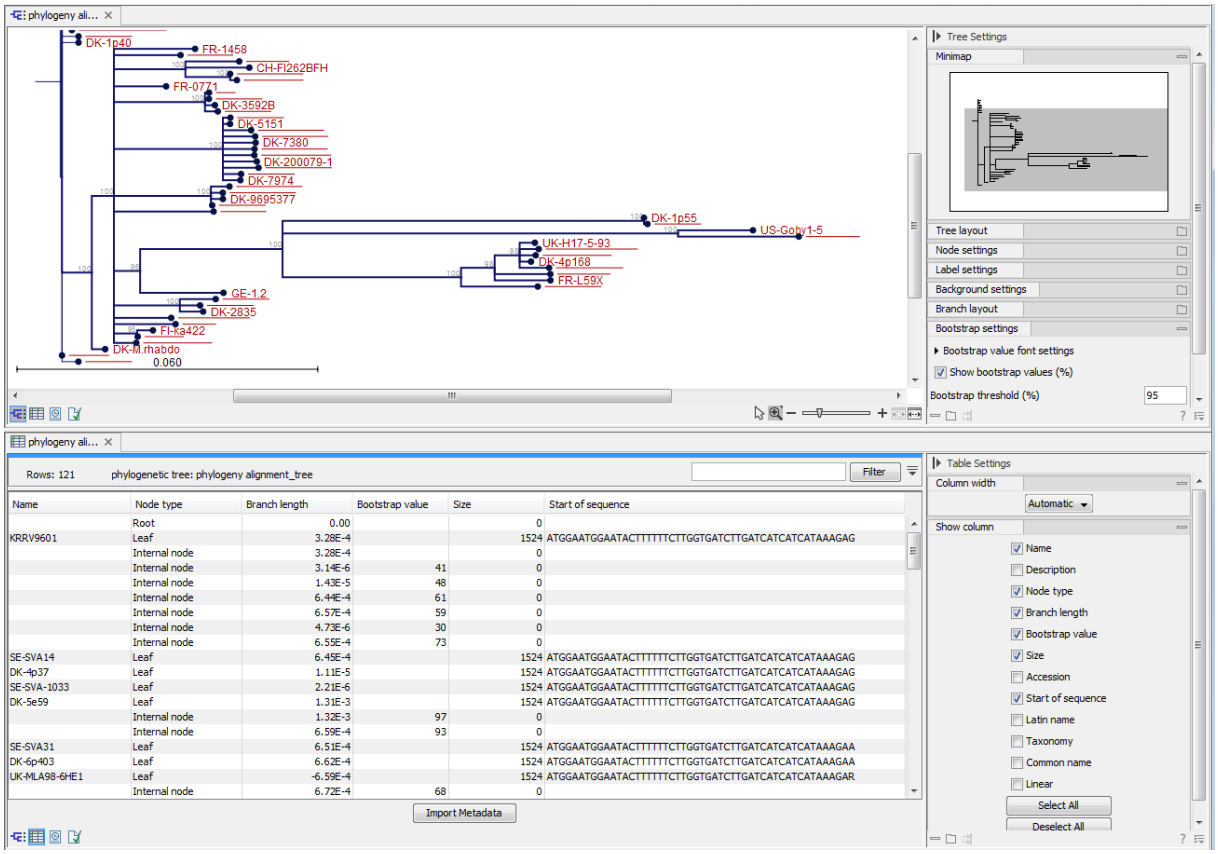

Figure 10: *Metadata table and tree viewer in split view mode. The red circle indicate the button for importing additional metadata.*

2. Click on the folder icon () next to the field in the Import section and select the file "Phylogeny_module_example_meta_data.xls" that you downloaded earlier.

3. Use the default settings, where the **Named columns** option should be checked. Information just below that option describes the mapping of categories to particular columns in the file. One column must be assigned type Name to allow the importer to associate metadata with nodes in the tree. In this case, a column in the original file called Strain will be mapped to the Name category. This means that the identifiers in the Strain column of the Excel sheet are used to link the information in a particular row of the file with the relevant nodes in the tree (figure 11).

4. Click **Finish**

After the import is complete, a number of new metadata columns have appeared in the metadata table: you can see a full list of the columns available in the right hand Table settings panel, and you can choose which ones to display in the table.

A quick way to see the metadata for a specific node in the tree is to hold the mouse over the node for a few seconds until a tooltip appears (figure 12).

Visualization of Metadata Visualization of metadata is carried out using the viewing option under the **Metadata** section in the right side panel. The tree viewer has several options for visualizing both textual and numeric metadata. Here we demonstrate three options.

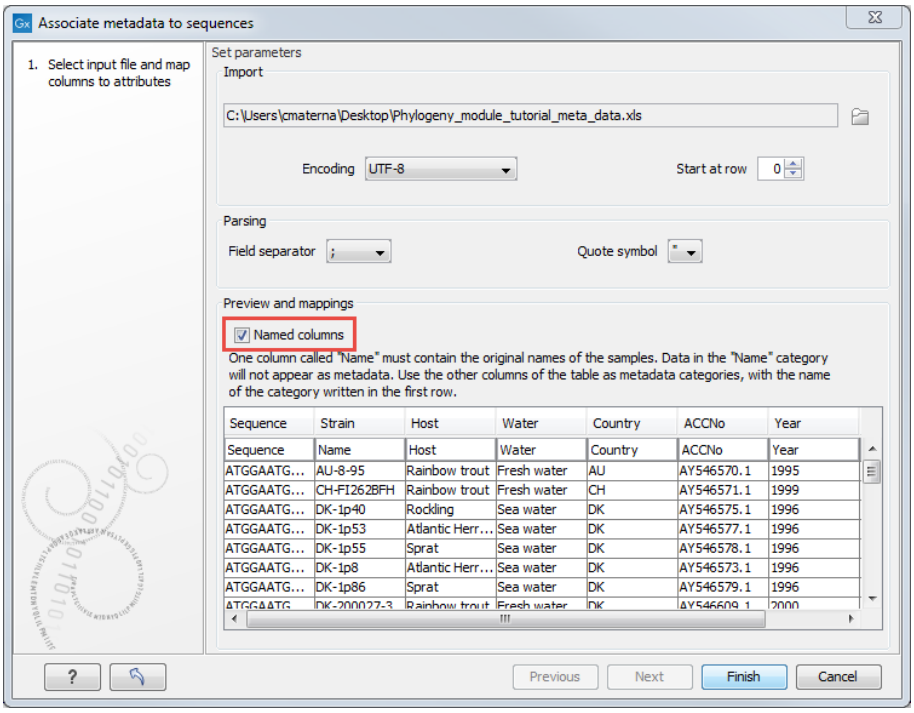

Figure 11: *The metadata import wizard. In this example the "Strain" column is mapped to the "Name" category and hereby used to map metadata to the tree nodes.*

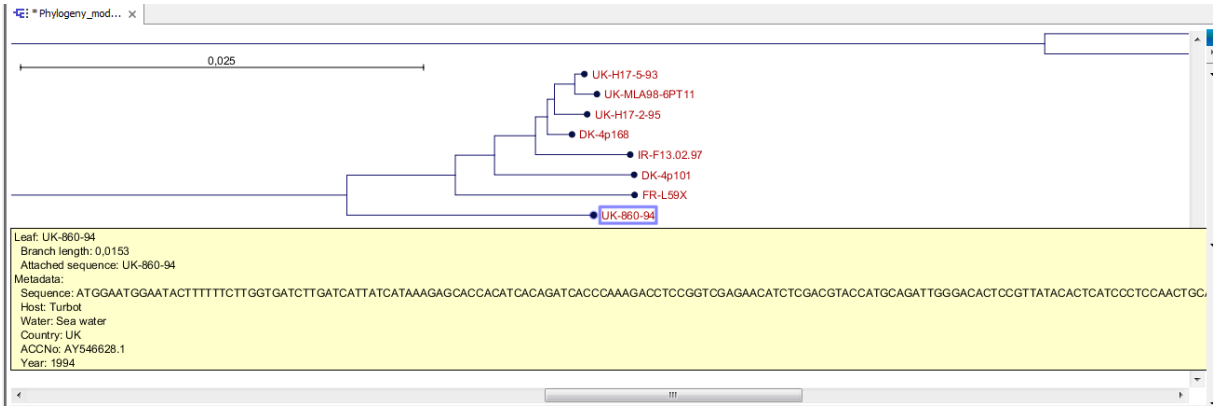

Figure 12: *When holding the mouse over a node in the tree a tooltip shows all metadata for that node.*

Node color

1. Under the **Metadata** section in the right side panel click on the text **Node color**.
2. In the dropdown menu select **Host**(see figure 13).

Nodes are now colored according to the host organism from which the virus sample was extracted. Random colors are assigned to each entry in the metadata category. To change a color go to **Node color** in the right side panel, left click on the color you want to change and choose a new color in the color chooser that pops up (figure 14). If an entry does not contain any data, the corresponding node has the label *Unknown* and is assigned a default color (usually the first color in the legend). This default color can be adjusted just like any other color. The mention of

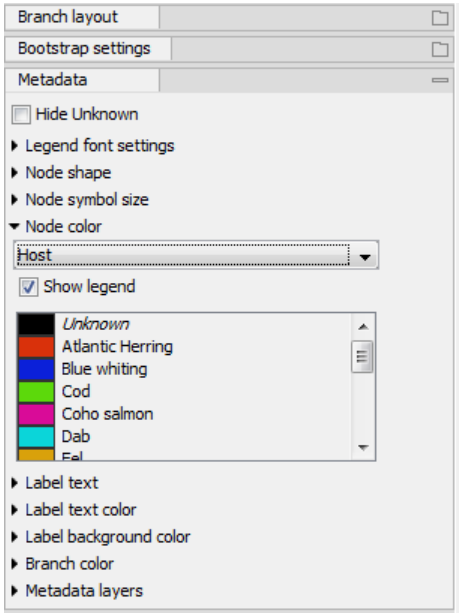

Figure 13: *The metadata visualization options in the right side panel of the tree viewer.*

Unknown can be edited by filling in all missing values of the metadata table with a value you may find more relevant.

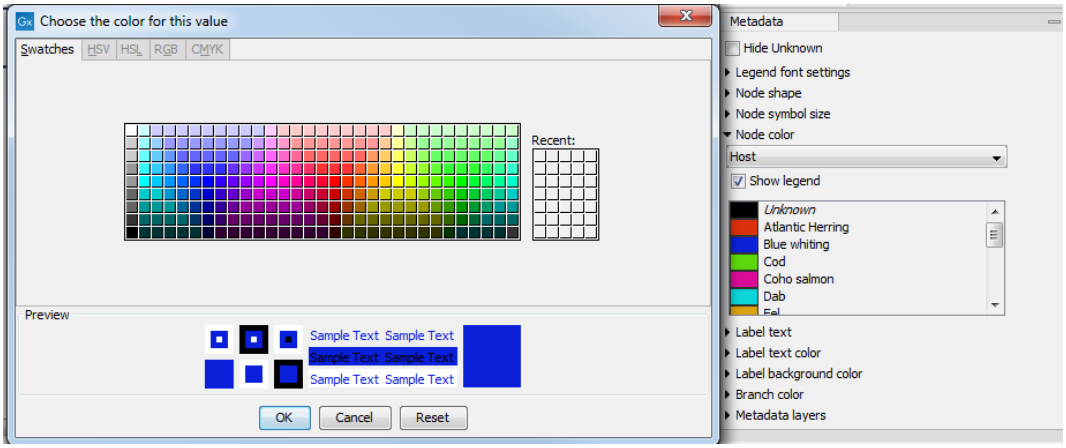

Figure 14: *The color chooser can be used to select new colors for each metadata category.*

Node symbol size or shape

1. Under **Metadata** in the right side panel click **Node symbol size** and/or **Node symbol shape**.
2. In the dropdown menu select **Water**.

All nodes are now assigned size and/or shape which symbolize the water type in which the host fish lives. Each type of water is automatically mapped to a node size. These sizes can be changed by using the sliders or drop-down menu in the right side panel.

Metadata layers

1. Under **Metadata** in the right side panel, select **Metadata layers** and then select **Metadata layer #1**.

2. In the dropdown menu select **Country**.

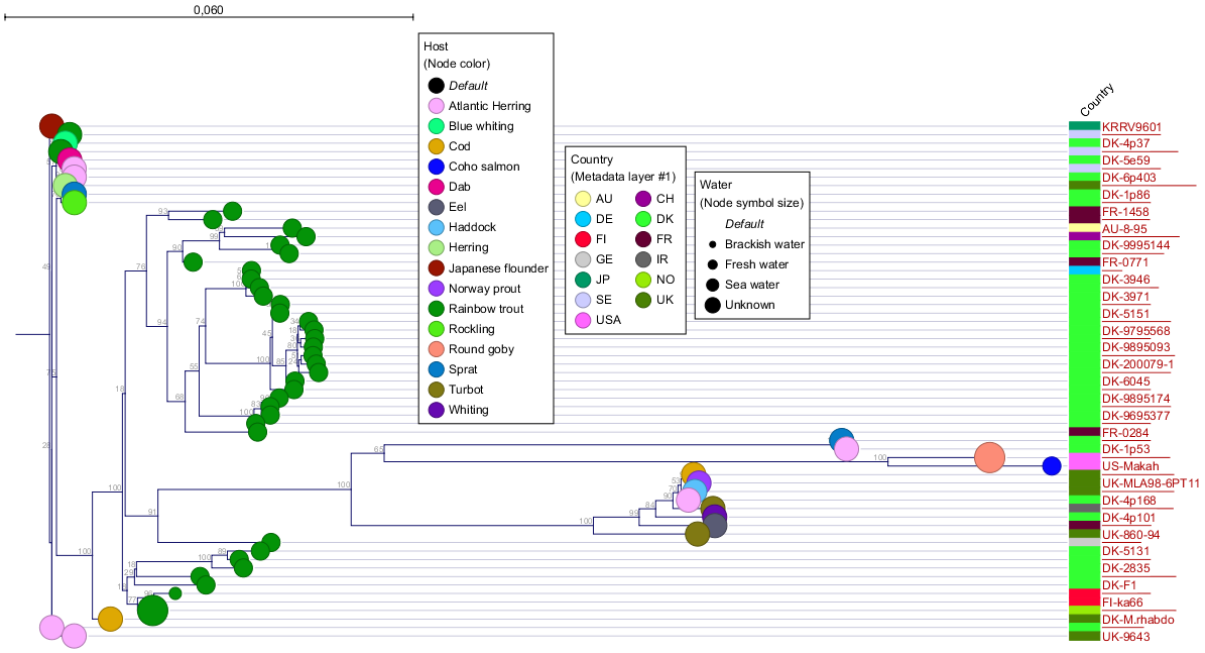

Figure 15: *Tree where water type, host and country are visualized.*

The countries where each sample was obtained are now displayed using colors in the color-code layer as shown at the bottom of the tree view (figure 15). Like the node color visualization, the color for each entry in the "Country" category can be changed under **Metadata layer #1**. It is also possible to adjust the thickness of the color-code layer by using the slider labeled **Layer thickness** under **Metadata layer #1**. If the **Show layer names** option is enabled, there is a limit to how thin a layer can be. This is to ensure that the name of different color-code layers do not overlap. By disabling the **Show layer names** option, layers can be made as thin as a single pixel.

After adding the color-code layer, a new option appears under the **Metadata layers** section in the right side panel called **Metadata layer #2**. By selecting a metadata category for this option, a new color-code layer will be added on top of the first color-code layer. By visualising metadata using multiple color-code layers, users can get a quick overview of data in different metadata categories. This makes it possible to visualize complex correlations. Figure 16 shows an example where three metadata categories are visualized using color-code layers.

Create and Modify Metadata

Metadata do not have to be imported from an external file. It is possible to both create and modify metadata through the metadata table and the tree viewer. A new metadata category can be created using the tree view by following these steps:

- Select one or more nodes in the tree view
- Right click one of the selected nodes and select **Assing metadata** (figure 17).
- Type in a name for metadata category, and use the **Value** field to assign a value to all selected nodes.
- Click on **Add**

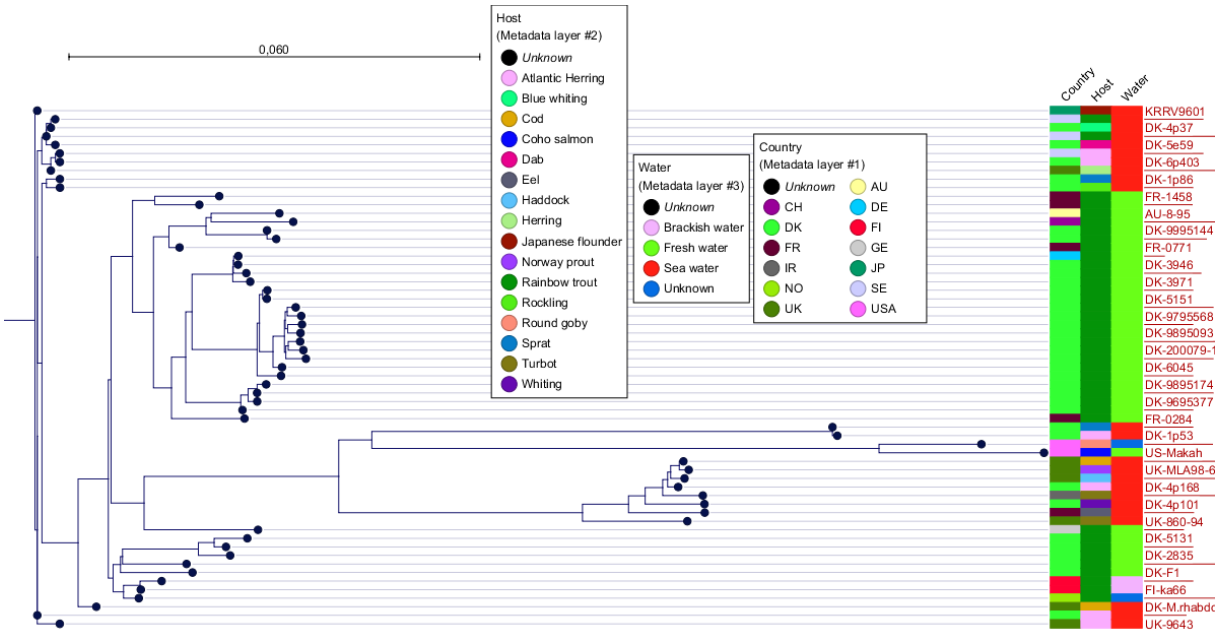

Figure 16: *Tree where water type, host and country are all visualized with color-code layers.*

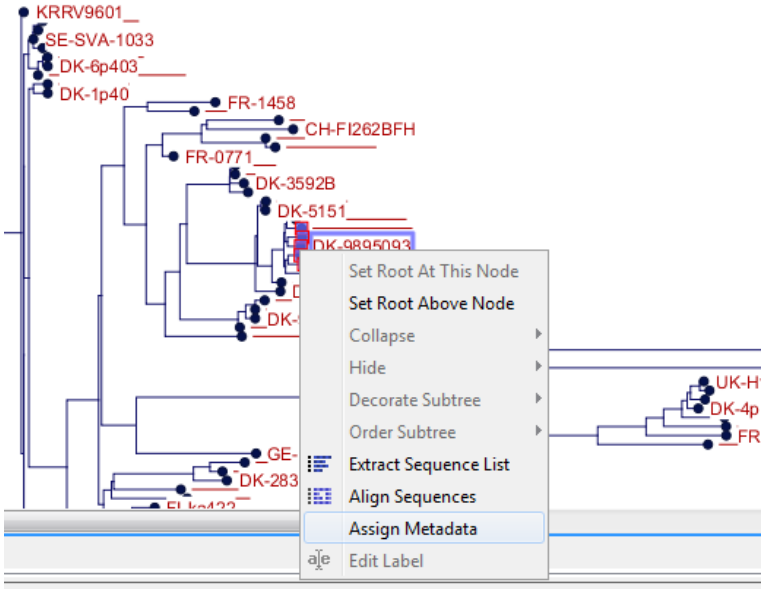

Figure 17: *Creating a new metadata category using the right click context menu.*

Now a new metadata category has been created with values assigned to all selected nodes. To assign values in the same metadata category to other nodes, follow the steps above but instead of writing a new name for a metadata category, use the drop down menu in the **Name** field to select the newly created category (figure 18). The value entered in the **Value** field will now be assigned to the selected category for all selected nodes.

Manipulation of metadata via the metadata table is very similar to manipulating metadata via the tree view. The only difference is that you need to select one or more rows in the table instead of selecting nodes in the tree.

Manually created metadata categories can be visualized in the same way as any other metadata category. An example where a new metadata category called "Continents" has been created and

Tutorial

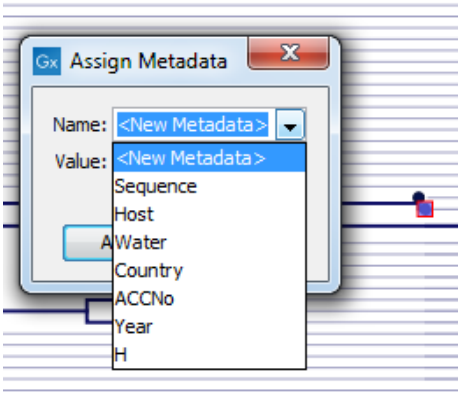

Figure 18: *Modifying an existing metadata category.*

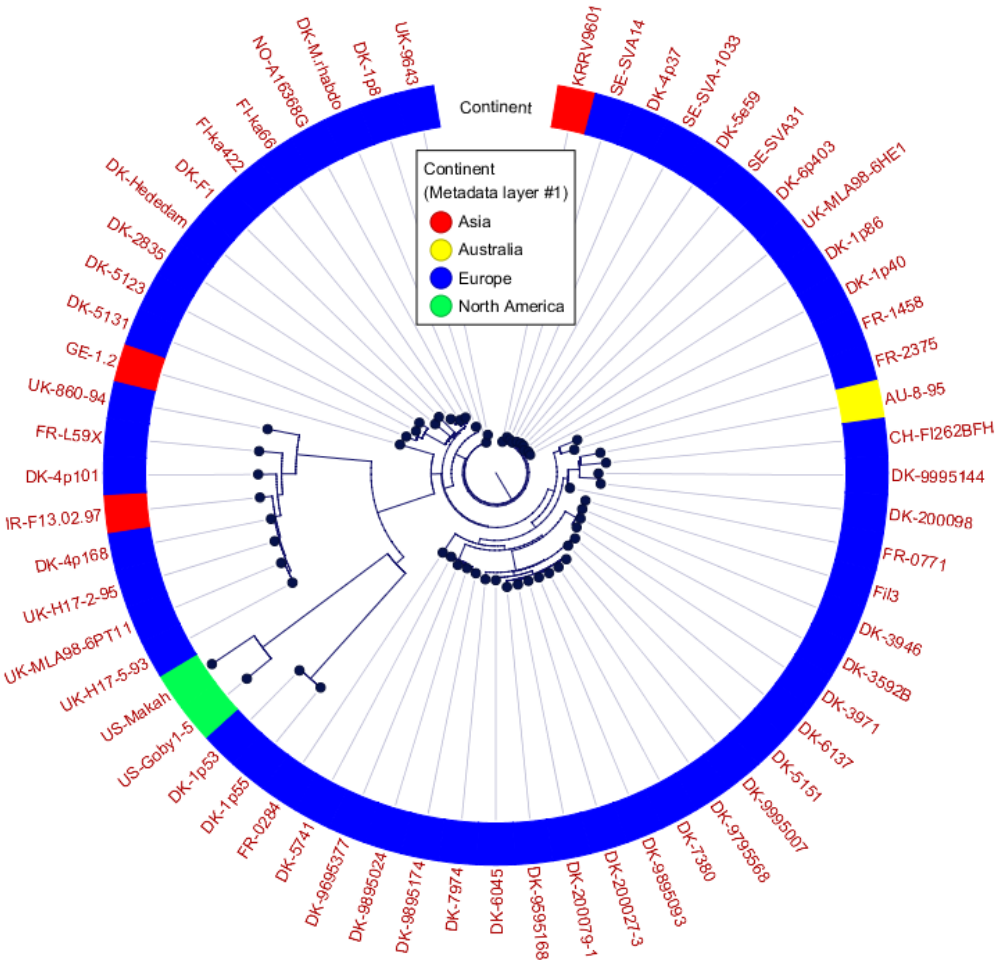

Figure 19: *Visualization of the continents where the host fish was found. The circular phylogram layout is used here.*

where all leaf nodes have been assigned values in that category according to the continent where the sample was obtained, is shown in figure 19.

Subtree Labels

It is often convenient to put labels on subtrees to illustrate what the nodes in a subtree represents. This can be done in the following way:

1. Right click an inner node in the tree.
2. Select **Decorate Subtree** and then select **Set Subtree Label** (figure 20).
3. Enter a label text and select the line color.

The result is a labeled line at the bottom of a subtree where the inner node which was clicked in step 1 is the root. Figure 21 shows an example where different subtrees have been labeled.

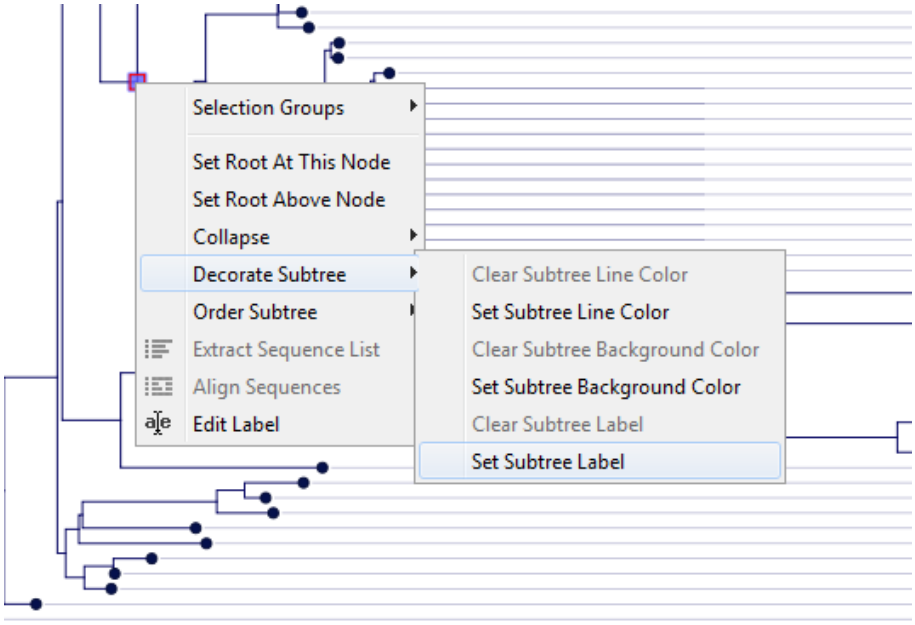

Figure 20: *Creating a label for a subtree.*

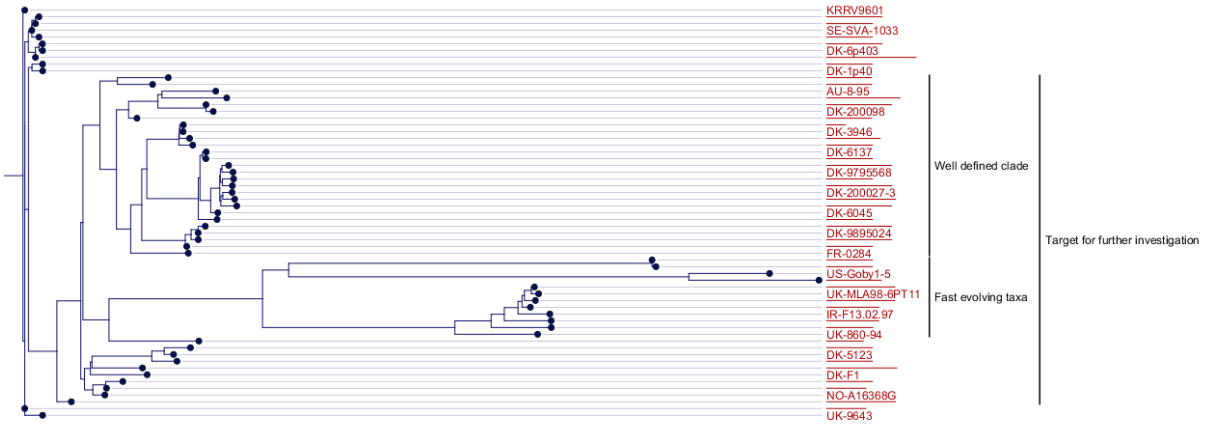

Figure 21: *A tree with three labeled subtrees. One label is for a subtree in which both the other two labeled subtrees are contained.*