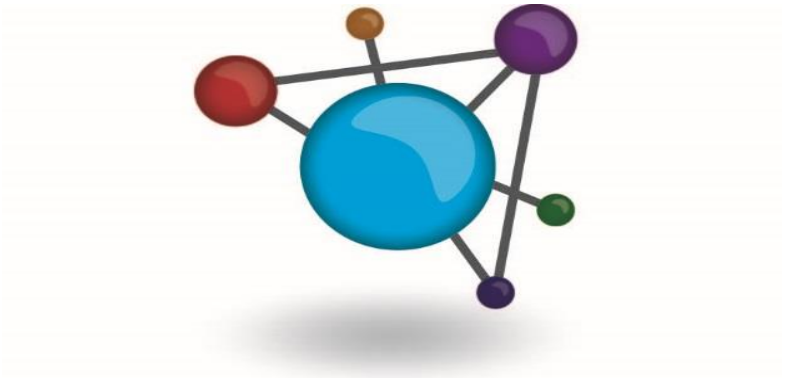


MicrobeTrace

User Manual for v 0.6.2

May 2021

Software Disclaimer

The material embodied in this software is provided to you "as-is" and without warranty of any kind, express, implied or otherwise, including without limitation, any warranty of fitness for a particular purpose. In no event shall the Centers for Disease Control and Prevention (CDC) or the United States (U.S.) government be liable to you or anyone else for any direct, special, incidental, indirect or consequential damages of any kind, or any damages whatsoever, including without limitation, loss of profit; loss of use, savings or revenue; or the claims of third parties, whether or not CDC or the U.S. government has been advised of the possibility of such loss, however caused and on any theory of liability, arising out of or in connection with the possession, use, or performance of this software.

Access MicrobeTrace at https://microbetrace.cdc.gov/MicrobeTrace/

If you use MicrobeTrace in your publications, please cite https://github.com/CDCgov/MicrobeTrace/wiki

Contents

Introduction

*A **glossary of terms** is provided after this section for details on terms commonly used in network building and analysis. As you move through the manual, you will find that many terms or references are in blue text; clicking on these hyperlinks will take you to the relevant word in the glossary section or to a website for additional information.*

MicrobeTrace is a software tool that enables rapid visualization of networks and associated data. MicrobeTrace allows users to map characteristics of their data to visual on-screen characteristics (e.g., color, size, shape) of elements of the network. In addition to network visualization, MicrobeTrace also provides other analytic tools (e.g., tables, filters, geographic maps, histograms, 3D networks, phylogenetic tree building, Gantt charts, a time player, alluvial diagrams, and scatterplots) to explore and contextualize nucleotide sequence and other data. These methods have been widely adopted in epidemiology, especially when responding to tuberculosis, HIV, HCV and COVID outbreaks, but have broad applications from molecular biology to sociology.

For nucleotide sequences, a genetic network is constructed after computing genetic distances using the TN93 (Tamura-Nei, 1993) nucleotide substitution model which computes distances between two sequences based on differences in nucleotides between the sequences per site. Potential links between the individual sequences are identified using an empirically determined genetic distance cutoff. For HIV sequences, TN93 is the nucleotide substitution model used and a genetic distance of 1.5% nucleotide substitutions/site is a good initial cutoff for examining the genetic relationships in your dataset, although smaller distances such as 0.5% may improve the specificity for recent

transmission. For other pathogens, MicrobeTrace allows importation of distance matrices determined using other nucleotide substitution models or hamming distances for pathogens with single nucleotide polymorphism (SNP) data. You can also import Newick phylogenetic tree files and MicrobeTrace will generate a network using the genetic distances calculated from the tree branch lengths using a patristic distance algorithm.

MicrobeTrace can also generate social network diagrams using contact tracing or partner services data. All networks can be customized according to available supplemental data sources (demographic, clinical, epidemiological, etc.) and mathematical inferences like the most probable transmission pathways can be determined by using the included minimum-spanning methods.

MicrobeTrace is a highly responsive, visual sequence analytics tool, which can reduce the gap between data collection and analytics and help you to discover, understand, and communicate relationships (represented as lines or edges) between individuals (represented as nodes in the network). Although it uses the capabilities of a web browser, MicrobeTrace works from a location on your laptop, not on a web-based server, and can be deployed at locations without internet access, thereby reducing both the startup cost and analysis time and effort, all while maintaining data security. Data security is of utmost importance when using sensitive data and should be given the highest consideration when using MicrobeTrace. **Please follow your institution's data security policies when using MicrobeTrace.**

This user manual is a step-by step guide on how to use the software beginning with loading data to generation of many visualizations using textual descriptions of each step along with corresponding screenshots. We begin with a glossary of terms commonly used in network analysis as well as descriptions of various types of files used with MicrobeTrace. This manual serves as a stand-alone guide. However, we are happy to schedule trainings or help users navigate MicrobeTrace as needed and as our schedule permits. We provide technical support details in the Support and Questions section.

MicrobeTrace Users

MicrobeTrace is a versatile, powerful, open source data visualization tool that could be valuable to researchers from a wide range of disciplines. State and local public health workers investigating active microbial transmission clusters and researchers (academic and government) conducting

transmission network analysis will find MicrobeTrace especially useful. Although the software was originally designed for HIV and contact tracing transmission analysis, MicrobeTrace is pathogen agnostic and can be used with any pathogen or disease data.

MicrobeTrace Help

MicrobeTrace help is available by selecting the "Help" link on the right of the menu options as shown in Fig. 1. Help consists of very useful information and step-by-step procedures to assist you in using the tool.

- **Website:** https://microbetrace.cdc.gov/MicrobeTrace/ - Follow this link to launch MicrobeTrace
- **GitHub: http://github.com/cdcgov/microbetrace -** GitHub is the software repository where MicrobeTrace is hosted. You can find brief descriptions of the software and its capabilities as well as code for various components. This is also a platform to report any issues or bugs you encounter.
- **Support and questions: microbetrace@cdc.gov** Email us with any questions about MicrobeTrace. Someone from our team will be happy to respond.

Glossary of Terms

Network Terminology

Edge – A link or line in a network that connects two nodes is referred to as an edge. An edge consists of the unique ID for both connected nodes that are typically labeled as "Source" and "Target". An edge can be the close genetic relatedness between two sequences in your data, but cannot infer directionality of transmission between these two sequences.

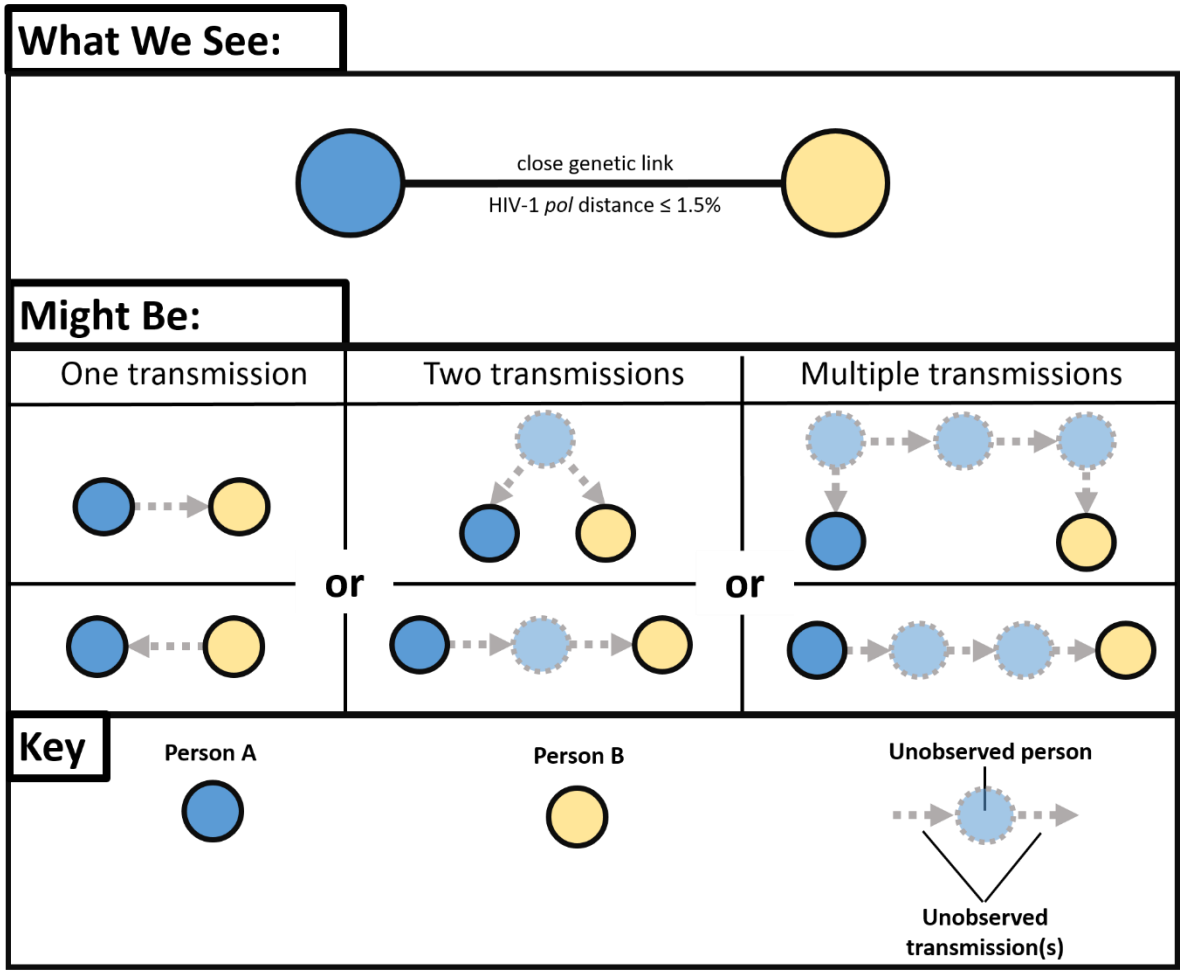


Fig. 1. Possible transmission relationships between nodes (persons)

A close genetic link between HIV-1 polymerase (*pol*) sequences (distance ≤1.5%) can represent many actual transmission scenarios that could involve at least one unobserved person. Six potential transmission scenarios are shown above.

Edge Attribute – A data field associated with an edge (i.e., a characteristic of the edge) that can be a categorical or numerical value. For example, one could calculate the absolute difference in ages

between individuals connected by an edge. Figure 1 above outlines various possible transmission scenarios between Node (Person) A and Node (Person) B that make determination of directionality difficult without inclusion of additional epidemiologic information.

Edge List – A list in which all edges and associated information (e.g., genetic distance and/or contact type data) occur exactly once. Edge Lists are also referred to as Link Lists. For MicrobeTrace, these data are included in a CSV (comma separated values) file or a Microsoft Excel file. CSV files can be prepared by storing the metadata in an excel file that is then saved as a CSV file. Note that reciprocal edges (**Person A → Person B** and **Person B → Person A**) are considered unique. Below is an example of an edge list.

Source	Target	Genetic Distance	Type of Contact
Person A	Person B	0.004	Sexual
Person B	Person A	0.004	Social

Metadata – Data that provide information about other data. Metadata can exist for both edges and/or nodes: for example, the record entry date of a new case or the type of high-risk contact associated with a link. For MicrobeTrace, this data is included in a CSV (comma separated values) file. CSV files can be prepared by storing the metadata in an excel file that is then saved as a CSV file.

Node – A discrete object in a network that typically represents a person (as in a contact tracing network) or a viral nucleotide sequence (as in a genetic distance network) from an infected person.

Node Attribute – A data field associated with a node (i.e., a characteristic of the node like a person's age) that can be a categorical or numerical value.

Node List – A list in which each node and its associated information (e.g., demographic and behavioral details) occurs exactly once. For MicrobeTrace this data is included in a CSV file. CSV files can be prepared by storing the metadata in an excel file that is then saved as a CSV file. Below is an example of a node list.

ID	Gender	Age (years)	Race/Ethnicity
Person A	Male	26	White

ID	Gender	Age (years)	Race/Ethnicity
Person B	Female	23	Black

Networks – For the purposes of this software and manual, there are social or contact networks and genetic distance networks. Social and contact networks are determined from behavioral data collected by partner services during the investigation. Genetic networks are inferred from the microbial nucleotide sequences of the pathogen being studied. Both network types should be included in the analysis for optimal epidemiological understanding of the transmission network, which is a network that combines data from both the social/contact tracing and genetic networks.

Source – The node where an edge begins. For example, if **Person A** names **Person B** (**Person A → Person B**), then the source is **Person A**. Please note that in this context, source does not imply the source of transmission.

Target – The node where an edge ends. For example, if **Person A** names **Person B**, then the target is **Person B**. Please note that in this context, target does not imply the target of transmission.

Genetic Analysis

Cluster – A cluster is defined as a group of nodes in which each node can be reached either directly or indirectly from any other node. A cluster can be identified using molecular sequences and/or contact tracing data. If no path or edge exists between two nodes, then they are considered to be in different clusters or they are singletons. ***It is important to note that the identification of a cluster will change depending on your chosen genetic distance threshold or the information provided in the contact tracing data.***

Dyad - A cluster containing only two nodes.

FASTA File – A text-based file format for representing a nucleotide sequence that consists of the standard IUPAC single letter characters for a nucleotide or amino acid. FASTA files can have the file name extensions .FASTA, .FA, .FAS or even saved as a text file (.TXT). The FASTA file extensions do not need to be uppercase. The first line in a FASTA file starts with a “>” (greater than sign without the apostrophes) and includes the code or text used for the name of the sequence, specimen or person. The next line in the FASTA file contains the actual nucleotide sequence using the one-letter IUPAC code. **If you are uploading a sequence file as well as a corresponding**

node list with demographic data (CSV file), IDs used for sequences in the FASTA file must match exactly those in the CSV (or Excel) file and must also be unique. A multiple sequence FASTA file would contain multiple iterations of unique sequence names and their corresponding sequences. Blank lines do not have to separate the first and subsequent sequences in the multiple sequence FASTA file. Here's an example of the contents of a multiple sequence FASTA file containing three different short sequences.

```
>Sequence ID 1
ATCGATCGATCGATCGATCG
>Sequence ID 2
ATCGATCGATCGGGGGGGG
>Sequence ID 3
ATCGATCGATCGCCCGGGTT
```

Genetic Distance Threshold – When analyzing nucleotide sequences, the genetic distances between all possible pairs of sequences are determined using a nucleotide substitution model (MicrobeTrace uses the TN93 model). Therefore, a genetic distance threshold or cutoff for the analysis must be selected to determine potential transmission linkage. For HIV transmission, a 1.5% genetic distance threshold corresponding to 0.015 nucleotide substitutions/per site is used as a starting point to link closely related viruses. For comparison, a distance of 1.0% between two HIV *pol* sequences represents about 10 years of viral evolution within an individual mono-infected with HIV-1 subtype B. However, users will need to select an applicable threshold depending on the situation (e.g., recent vs. distant evolutionary past) and specific pathogen under investigation. The potential cutoff for your analysis can be determined by identifying a threshold that best differentiates a bimodal distribution of the genetic distances that are typically present in your sequences.

Figure 2 below is an example of a frequency distribution of genetic distances with the blue boxes representing genetic distances from known pathogen transmission cases, and the red boxes representing genetic distances from cases without evidence of pathogen transmission. For this data set, a genetic distance of 0.02 nucleotide substitutions/site would likely best differentiate the genetic distances from viruses/pathogens associated with and without transmission (see dashed line in figure). A lower threshold will result in fewer identified linkages, but with increased specificity for recent transmission.

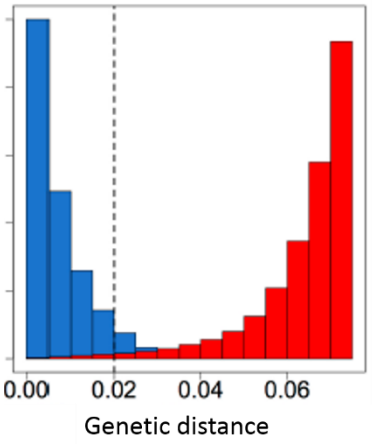


Fig. 2. Genetic distance histogram

Patristic distance algorithm — Sum of the lengths of the branches that link two nodes in a tree.

Singleton – An isolated node. For example, a node that does not link to any other nodes in the network.

SNP – Single nucleotide polymorphism. This is a single nucleotide difference between two sequences that occurs at a specific position in the genome, oftentimes referred to as a genetic mutation. SNPs are more frequently used with bacterial pathogens.

Network Visualization Parameters Network visualization is optimized using physical properties, including charge, friction and gravity, to determine how densely packed the nodes are in the network.

Charge – Nodes repel each other in the network visualization to maintain separation so all nodes are visible. To change the degree to which the nodes repel one another, this setting can be modified as desired.

Friction – The rate at which a node can move across the network view on your computer screen. High friction means nodes won't move much.

Gravity – Nodes are drawn to the center of the network view in your computer screen in proportion to a gravitational constant. Low gravity means nodes will float toward the edges and high gravity will ensure that they are tightly clustered on-screen.

Tool tip – A software visualization tool that displays information when the mouse pointer hovers over an object.

System Requirements

MicrobeTrace can run on any computer using Google Chrome, Firefox, or Microsoft Edge**.**

***Please note* MicrobeTrace is not compatible with any version of Microsoft Internet Explorer**.

Downloading Sample Data

Click here to download example files, including example node lists, edge lists, FASTA files and a SecureHIVTrace file that can be imported into MicrobeTrace. SecureHIVTrace is a private bioinformatics tool for registered public health jurisdictions in the US to perform HIV cluster detection using *pol* sequences. The names of the example files are listed below.

Node list: Demo_outbreak_NodeList.csv

Edge list: Demo_outbreak_EdgeList.csv

Sequence file (FASTA): Demo_outbreak_Sequences.fas

Distance matrix: Demo_outbreak_DistanceMatrix.csv

Newick (tree) file: Demo_outbreak_PhylogeneticTree

See descriptions below for details on each type of file.

When you click on Download Sample Data on the MicrobeTrace website, you are prompted to save or open the file. Select **Save As** and save the zipped file to the location of your choice (Fig.3).

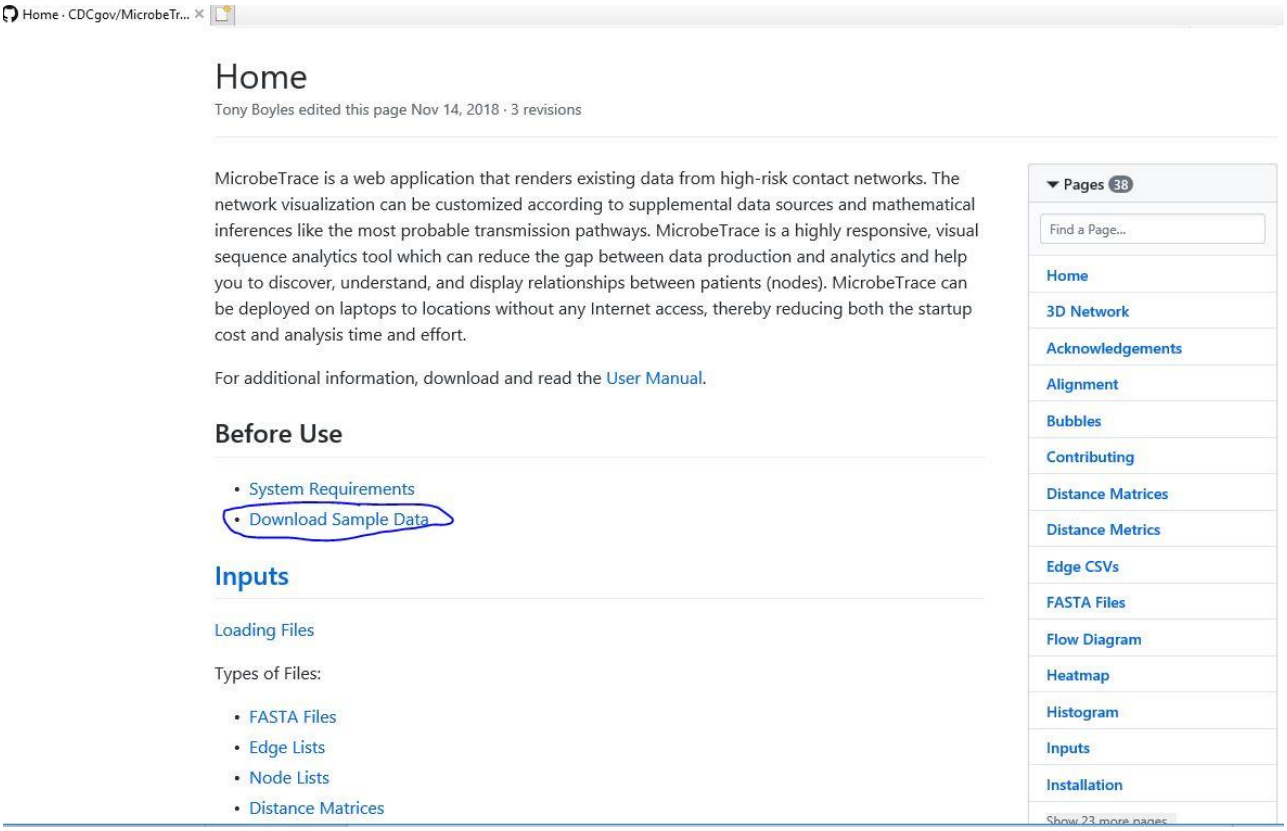


Fig. 3.a. Downloading example files from the MicrobeTrace website

Open the file location on your computer and you will see three zipped files. When you click on **Extract all Files**, a dialog box opens up; click on **Extract**.

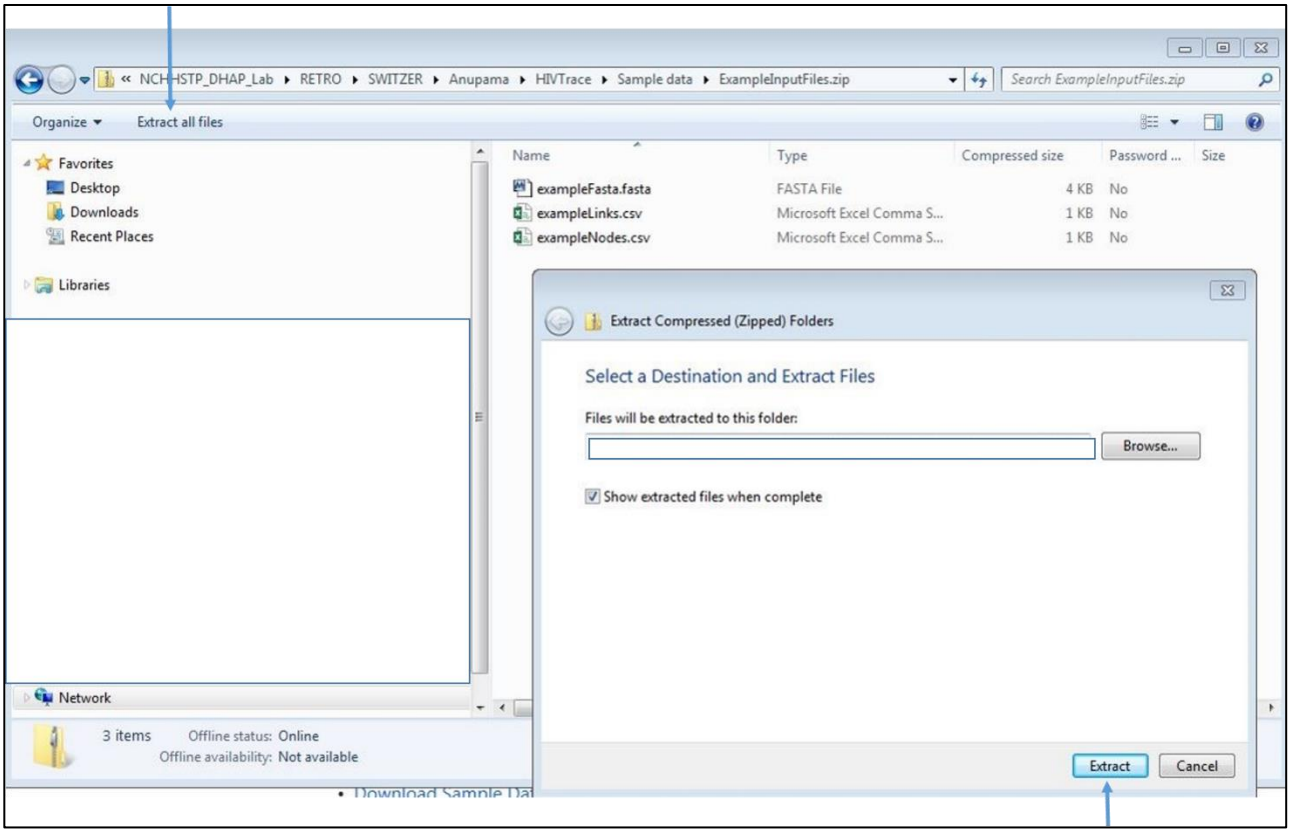


Fig. 3.b. Extracting MicrobeTrace example files from zipped folders

You can use these unzipped example files to explore MicrobeTrace.

Creating and importing files

MicrobeTrace accepts the following file formats:

- Nucleotide sequences in the FASTA file format
- Standard Microsoft Excel files, or comma- or space-separated files (.CSV files). These files can be edge lists or node lists (files with sequence/patient IDs with corresponding data such as age, sex, risk-type, method of transmission, diagnosis date, sequence subtypes). A node can represent many things, but in the context of partner services (e.g., contact tracing) they typically represent either an infected person or their high-risk partners. In a genetic distance network, nodes represent the pathogen sequences that appear in your FASTA file. IDs in the FASTA file appear as the text after the ">" and before any space in each sequence.

While node attributes are not required to visualize networks, they are a vital component of characterizing and exploring the network. To associate the node attributes to a network, each node ID in the node list CSV or Excel file must match exactly to its corresponding ID in the provided edge list, or to sequence IDS if sequences are loaded separately as a FASTA file. All available metadata that might help with the network analysis should be appended as additional columns that follow the node ID column in the node list CSV or Excel file.

NOTE: If your node list file contains a column with sequences in it, MicrobeTrace will include those sequences in the analysis, and you will have all the functionality associated with a separate sequence file. Alternatively, sequence names may be stored in the CSV or Excel file in any column with the column header named "ID". If more than one "ID" column exists, the leftmost one in the file will be used

PLEASE NOTE*:** ***In a CSV or Excel file, rows with identical node names cause repeating rows to be dropped. You MUST ensure that ID names are unique.

Edge lists: As an alternative input to a FASTA file, a list of edges can be provided which indicate connections between nodes defined in the node CSV or Excel file. This is called an "edge CSV" file and is typical of person-to-person linkages determined during contact tracing. In MicrobeTrace, this is also called a Link file or Link List or Link CSV file. Networks from edge CSV or link files are called contact tracing or social networks. Here is an example of data in an edge file.

Source ID	Target ID	Edge Attribute
John	Jacob	High-risk contact
John	Mary	High-risk contact

Additional edge properties (or data) can be visualized by adding data columns to the edge list. Edge properties can be any characteristic that further define relationships between two nodes. It is important to note that edge properties should reference both nodes that are connected by an edge. Some examples of edge properties include genetic distance between two pathogens, the type of high-risk contact that occurred between two people, or the age difference between two people. In a contact-tracing network, an edge represents an epidemiologic link between two people. In a genetic distance network, an edge represents the genetic relationship between two pathogens. Edges can be directed or undirected. Directed edges are represented by arrows between nodes. ***PLEASE NOTE*: Arrowheads are turned off by default. We STRONGLY advise that this default setting be used unless directionality has been supported with strong confidence using additional epidemiologic information (see Directionality for more information).**

- **MicrobeTrace session and style files**: If you are working on a dataset and need to save your data and settings for later use, this is a useful feature. The file will be saved with the filename extension .microbetrace. You can then load your saved .microbetrace file directly into MicrobeTrace. When you are saving your session, you also have an option to download data by clusters (Fig. 4).

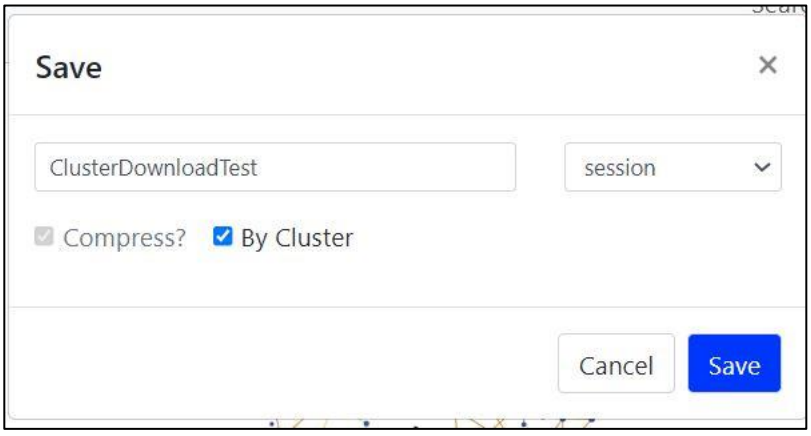


Fig. 4. Saving a file: downloading node and edge lists by cluster

If you select this option, MicrobeTrace will create a zipped folder in a location of your choice. This folder will in turn contain sub-folders for each cluster which will have node lists and edge lists for that cluster (Fig 5).

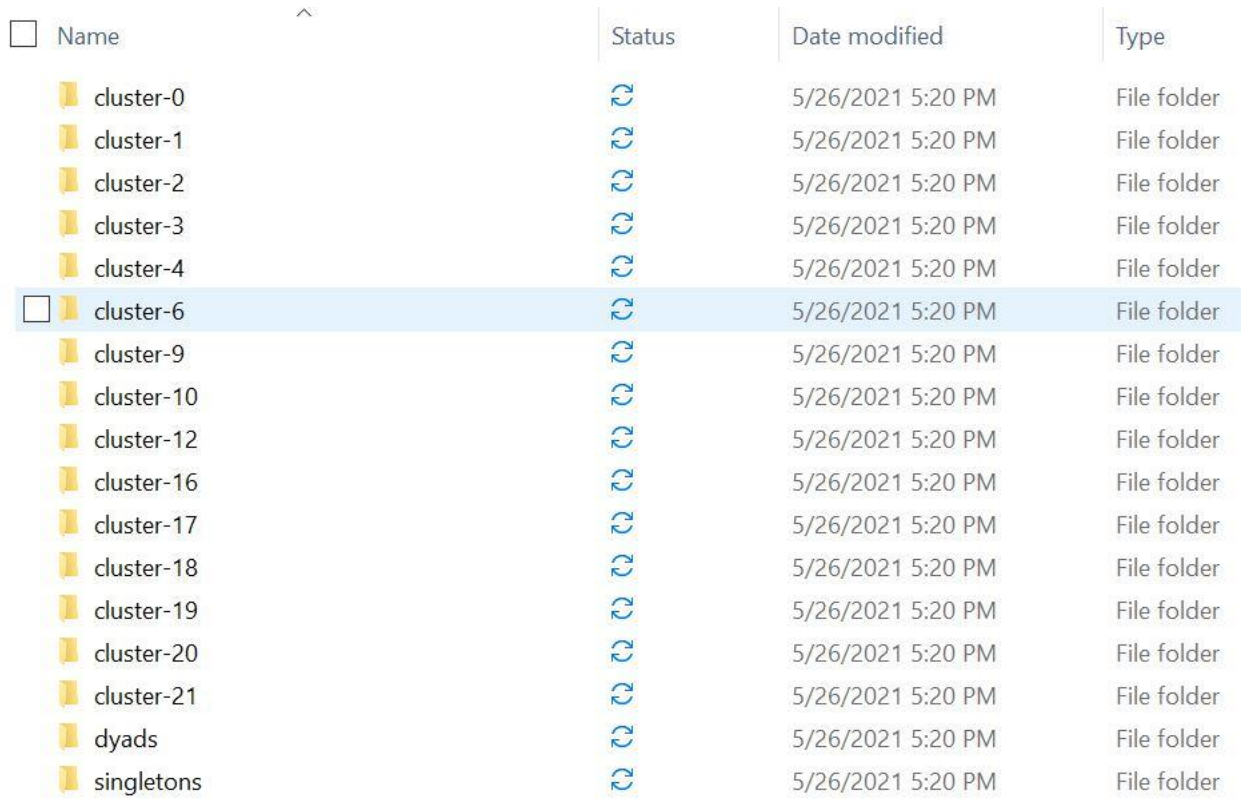


Fig. 5. Cluster level data downloads. Each folder contains node lists and edge lists with data on nodes within that cluster.

Similarly, if you have changed node and link settings (coloring or sizing by various attributes), you can save these as a style file (Fig. 6 below) You can then use this style file with another dataset via the **Global Settings** menu under the **Styling tab. Importantly, you must remember to use the same file types (i.e. node vs link) when you apply the style file in your new session or the saved styles will not apply correctly.**

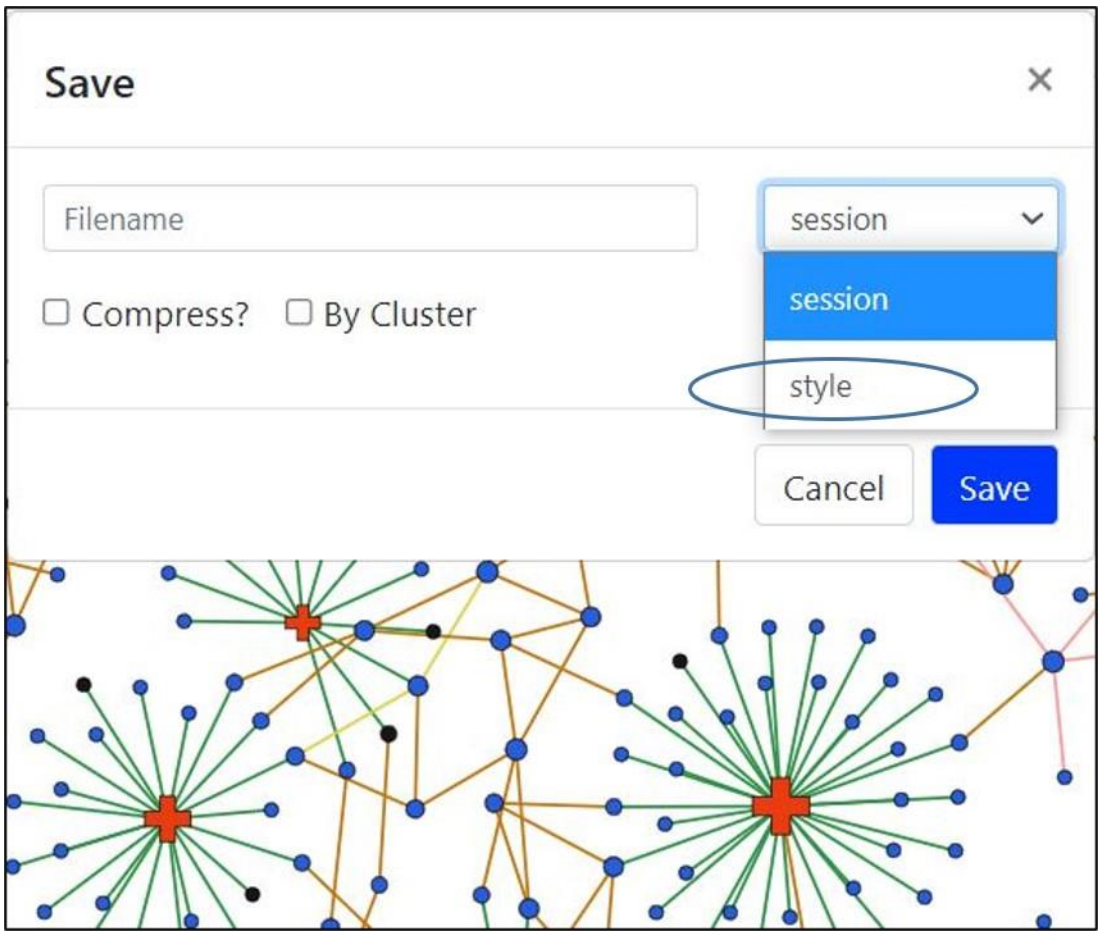


Fig. 6. Saving a style file

- **Distance matrix files (.csv):** This file type is especially useful if you plan to analyze a genetic distance network from a large dataset of sequences. Processing time in MicrobeTrace is significantly reduced if the genetic distances are pre-computed, and the resulting distance matrix file imported into MicrobeTrace.
- **Newick (.nwk) tree files:** This file type is useful if you have a tree generated from an external phylogenetic tool and would like to visualize it as a network in MicrobeTrace in conjunction with other associated data for the taxa in that file. MicrobeTrace uses a patristic distance method to generate a network diagram from your Newick file. This feature is particularly useful in cases where you may not have access to raw sequence data, but have the output from a phylogenetics tool like Nextstrain (https://nextstrain.org/) or if you want to convert a phylogenetic tree to a network and then add the associated metadata for further exploration.

***IMPORTANT*: IDs used for sequences in the FASTA file must match exactly those in the CSV file and must also be unique.** For best practices, duplicate IDs should not be used in a FASTA file. If duplicate IDs are detected, the sequences with identical IDs are automatically modified with an underscore and a consecutive number (e.g., PersonA_1, PersonA_2) to make them unique. New unique IDs will propagate to all data visualization layers.

Possible File Input Combinations

As shown in Fig. 7, a combination of data files can be input into MicrobeTrace depending on the specific analysis or network visualization desired. The examples shown in Fig. 7 are just some commonly used examples. You can also load two edge lists or two sequence files to overlay networks from either multiple pathogens, or to compare networks generated from contact tracing data and sequence data. These features are described later in the manual (Overlaying Networks).

The speed of network generation by MicrobeTrace will depend on the number of data files and amount and type of data and your specific computer configuration. The table below gives you an estimate of average time to process genetic data. These results include calculation of distance matrices and network computation.

Number of sequences (~1000-bp length)	**Time duration estimates**
≤ 50	<1 second
100	<1 second
150	<1 second
200	<1 second
250	1 second
350	1.5 seconds
500	2.1 seconds
750	3.6 seconds
1000	6.0 seconds

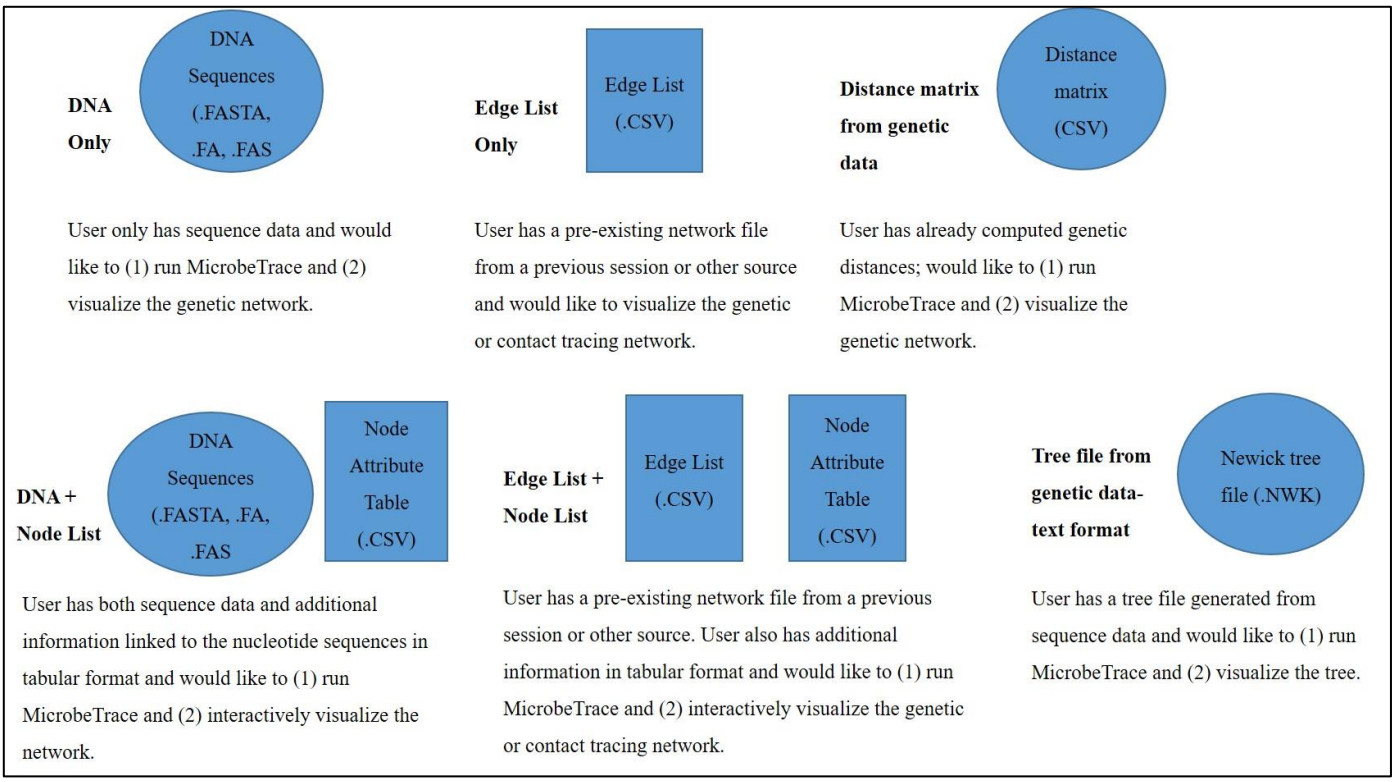


Fig. 7. Possible file combinations to create new networks or to visualize previously created networks

Accessing MicrobeTrace and Loading Files

Open a Chrome browser window and navigate to the address http://microbetrace.cdc.gov. Once loaded, the following home screen for MicrobeTrace (Fig. 8) is displayed, and you are ready to select and load files.

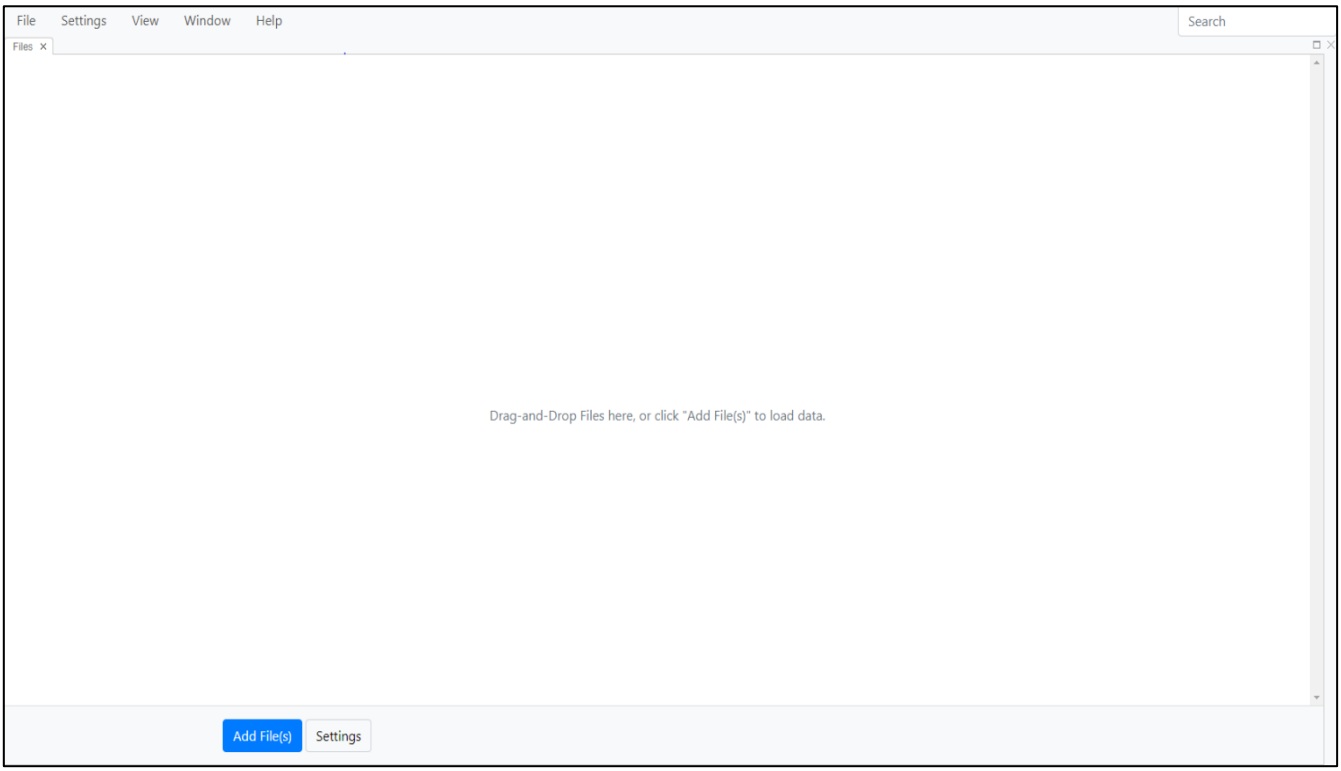


Fig. 8. MicrobeTrace home screen for selecting and loading data files

Main menu

From the main menu, you can access various options (Fig. 9).

File: Lets you open a new file, save current session as a .microbetrace file, or to add data files to the current analysis.

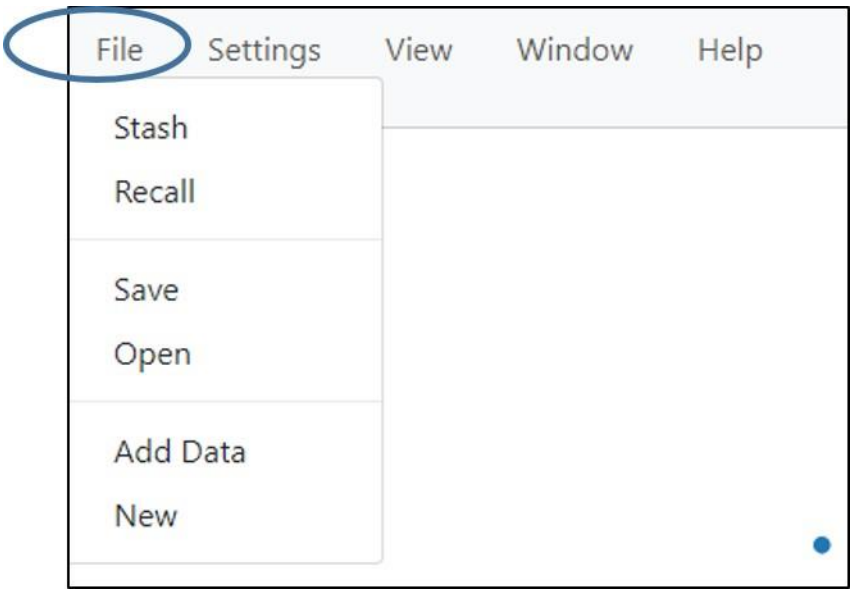


Fig. 9. File menu

Selecting **Stash** (Fig. 10) will allow MicrobeTrace to save your session so that if MicrobeTrace shuts down unexpectedly, you can then re-launch the program and use the **Recall** function under the file menu to recover the data you were working on. This is especially useful to save settings that you apply to your data.

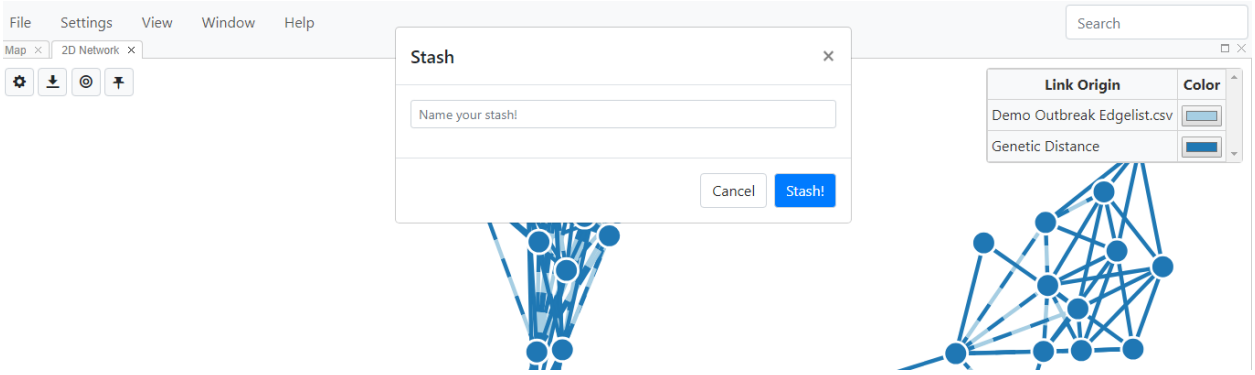


Fig. 10. Stashing your MicrobeTrace session

When you select Recall from the dropdown menu, a dialog box opens up with a list of stashed sessions (Fig. 11). Select the one you want and click Recall.

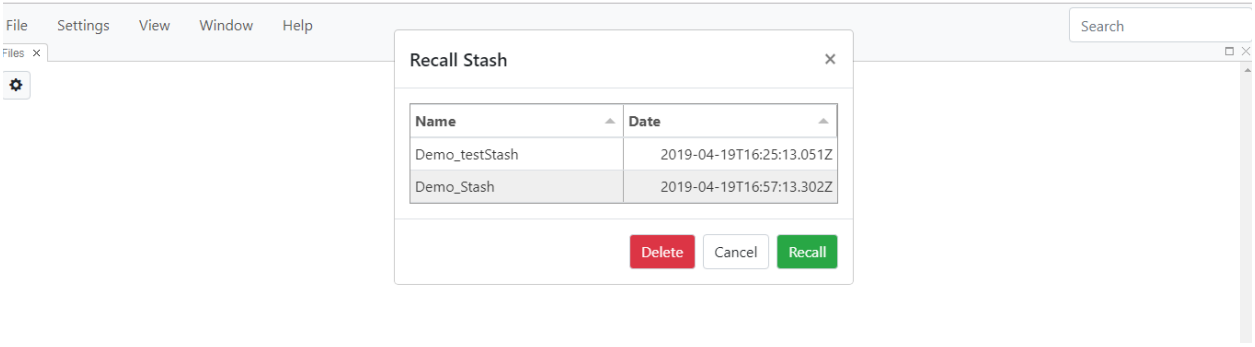


Fig. 11. Recalling a stashed session.

Settings: Opens the Global Settings dialog box. Allows link and node style settings and genetic distance cut-off values to be customized. Descriptions of each of these settings are in the Network configuration section.

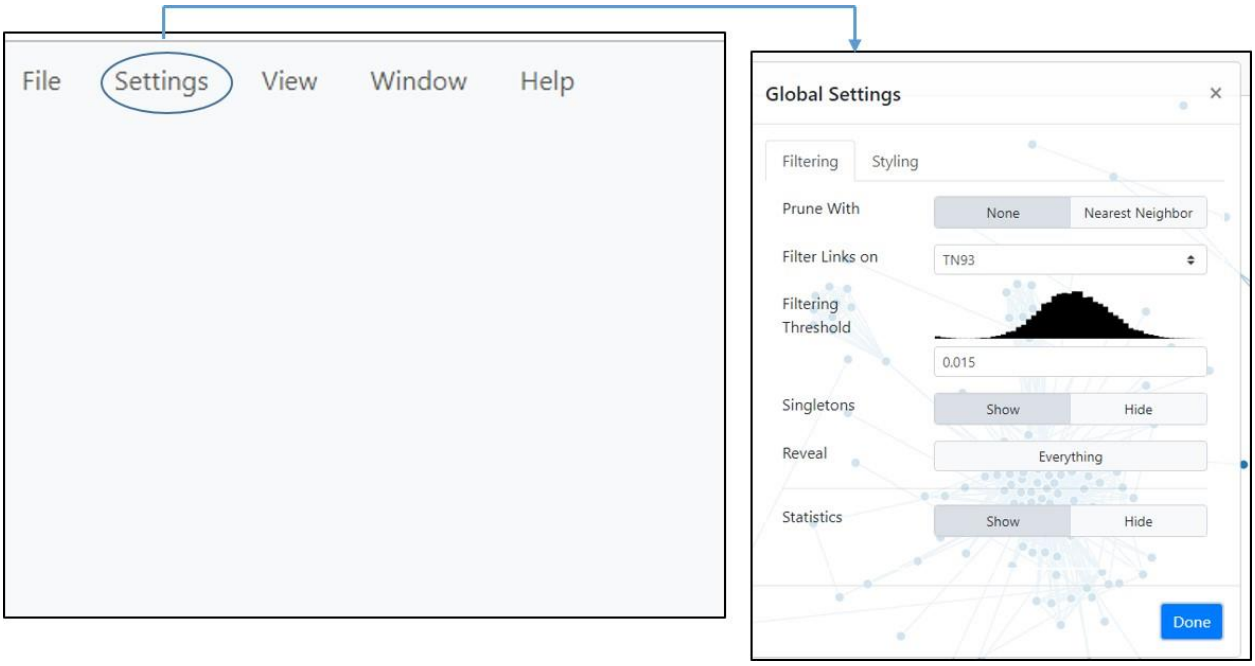


Fig. 12. Global settings menu

View: Gives you a list of data visualization views to choose from, each of which will be described in detail in the sections below. If you would like to read about a specific view first, you can click on the relevant heading in the table of contents to be taken directly to the description of that view.

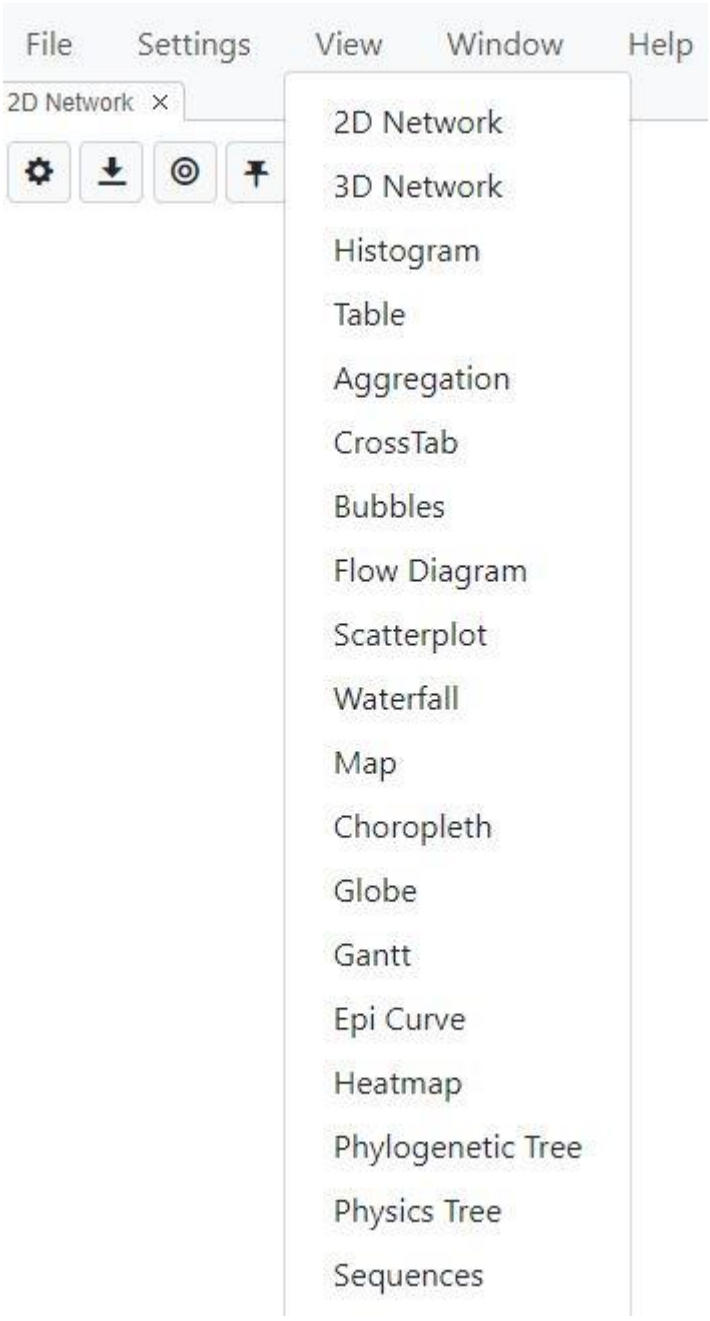


Fig. 13. View menu

Window: This menu gives you the option of reloading your data and lets you toggle to and from a full screen view.

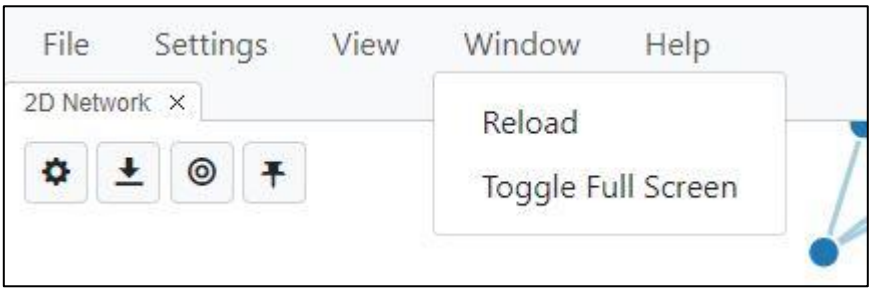


Fig. 14. Window menu

Selecting reload exits the session and goes back to the Add File screen. Toggle Full Screen toggles to and from a full-screen view.

Help: Provides access to the information about MicrobeTrace at the MicrobeTrace GitHub site (https://github.com/CDCgov/MicrobeTrace/wiki), and allows you to report any issues encountered (https://github.com/CDCgov/MicrobeTrace/issues).

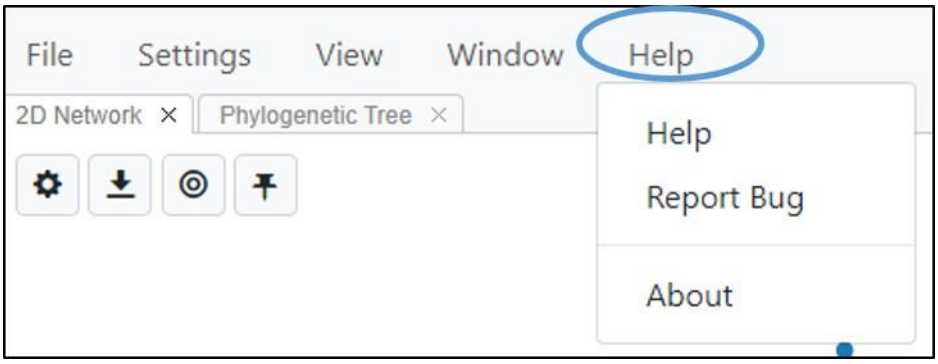


Fig.15. Selecting the help option in MicrobeTrace

Loading Files

Depending on the type of data you plan to analyze, you can load multiple data files simultaneously. You can also add files during a MicrobeTrace session. Below are a few common scenarios with the corresponding file types that you would use. The detailed instructions for each file type are described below.

1. You have only sequence data and want to construct a genetic distance network: Load **FASTA file**

2. You have sequence data, and also a node list with demographic or other information about the cases associated with the sequences: Load a **FASTA file** as well as the **node list CSV file** containing the node information in columns.
3. You have a node list which contains demographic or other information, including a column with sequences, and also a contact tracing file (edge list): Load the **node CSV** and the **edge CSV**. This will give you two overlaid networks - one computed from the edges (contact tracing network), and one computed from sequences in the node CSV (genetic distance network).

Loading a FASTA file

Step 1. Select the **Add Files** button (Fig. 16). The system displays a standard windows explorer page for navigating to the desired file. You can also drag and drop files from your computer into the MicrobeTrace window.

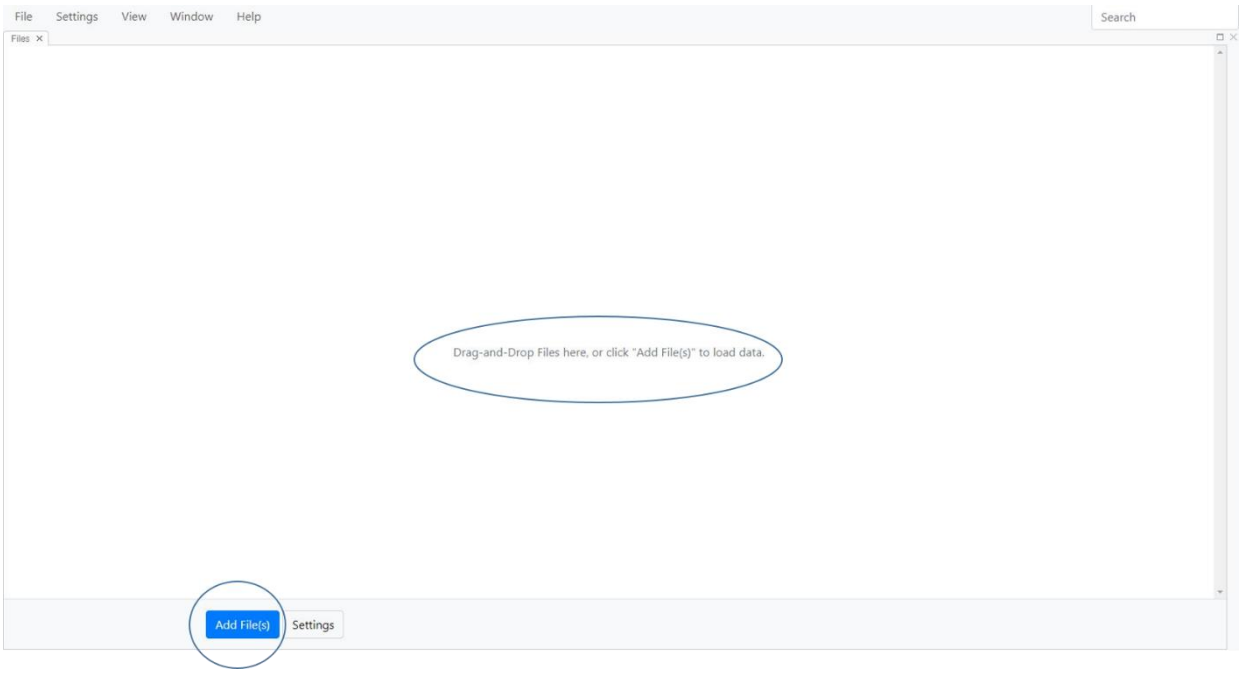


Fig. 16. Loading a file into MicrobeTrace

Step 2. Navigate to the example FASTA file and double-click on the file to add it to the analysis, or select the file and choose **Open**. This loads the FASTA file into MicrobeTrace.

Step 3. Selecting metrics (Fig.17) - Click on the settings button  on the top left corner of the screen. This will display two tabs: **Files** and **Experimental**.

The **Files** tab allows you to customize settings for the following:

- Distance metric (TN93 or SNPs)
- Handling of ambiguities (options below)

  -average (count any resolutions that match as a perfect match)

  -resolve (average all possible resolutions)

  -skip (skip all positions with ambiguities)

  -GapMM (count character-gap positions as 4-way mismatches, otherwise same as average)

  -HIV-Trace-g (Any sequence with no more than the selected proportion [0 - 1] will have its ambiguities resolved [if possible], and ambiguities in sequences with higher fractions of them will be averaged. This mitigates spurious linkages due to highly ambiguous sequences.)

- Link threshold settings (cut-off to determine if two nodes are linked)
- Opening view to launch (2D network, 3D network, Bubbles, Table, phylogenetic tree)

If you would like to see the network diagram, choose 2D or 3D network. If you would prefer to view the nodes without the links (in case of sensitive data), choose Bubbles. If you want to see data in a tabular format, rather than a network diagram, choose Table. If you are uploading sequences, or a Newick file and would like to start with a tree view, choose Phylogenetic Tree. Note that this sets only the opening view. You can then visualize your data in all these formats and more.

Default settings are:

Distance metric: TN93, **Ambiguities**: Average, **Link Threshold**: 0.015

View to launch: 2D network

Once you change these preferences, they will be saved as your defaults and applied whenever you open MicrobeTrace. You can also change these settings anytime during your MicrobeTrace session.

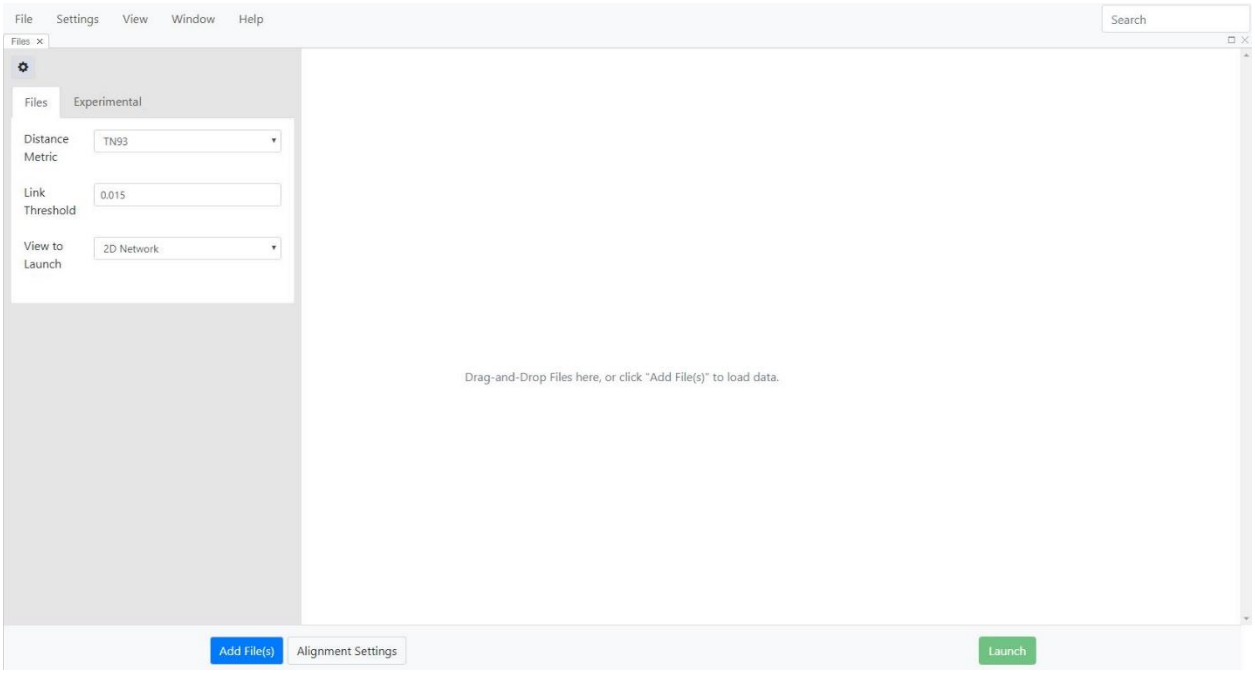


Fig. 17. Selecting preferred settings for distance metric, threshold, and opening view.

Experimental tab:

These options are purely "experimental" and allow you to try some of these less commonly used features. You can generate a set of random sequences to explore MicrobeTrace and select how many sequences you would like in this dataset.

You can choose whether or not you would like to turn on directionality or autostashing.
IMPORTANT NOTE: Predicting directionality with sequences is not reliable unless it has been supported with strong confidence using additional epidemiologic information (see Directionality for more details)*

Turning **Autostashing** on allows MicrobeTrace to autosave your session for re-loading later. When turned on, the session is saved every minute to local storage.

These two parameters are turned OFF by default.

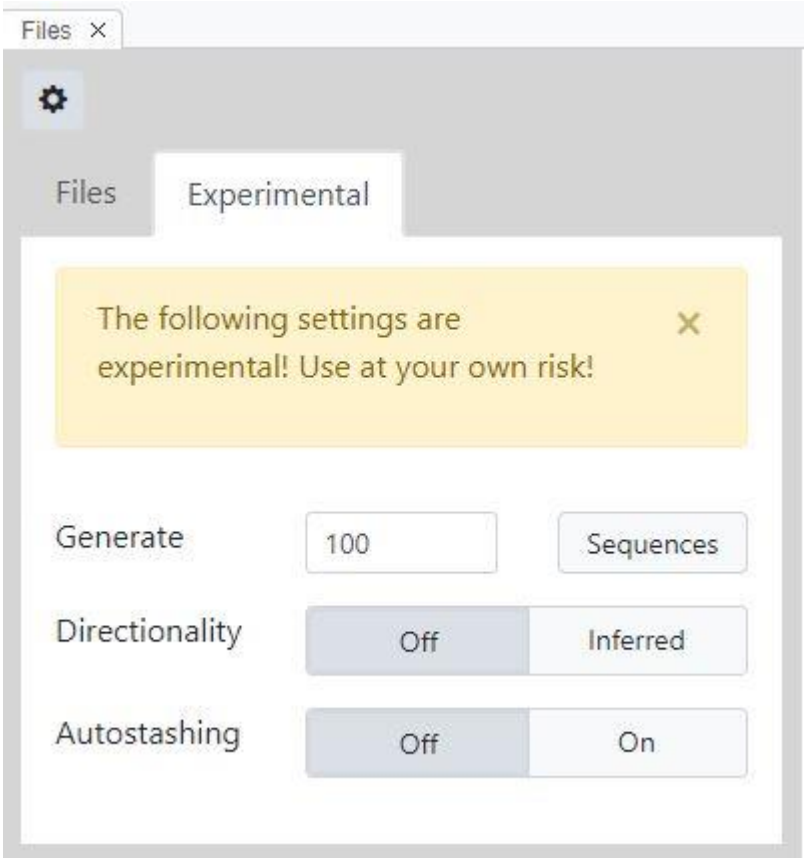


Fig. 18. Experimental tab settings

Step 4. The **Settings** button gives you the choice of aligning your sequences against a reference sequence. ****IMPORTANT NOTE*: The default option is that your sequences will NOT be aligned. Please make sure that your sequences are aligned before network analysis; either in MicrobeTrace (steps below) or using an alignment software of your choice.***

****IMPORTANT NOTE*: If you are using pre-aligned sequences, please proceed directly to the Submit step, this submits sequences with the default settings. MicrobeTrace may give you faulty networks if you ask it to align sequences that have already been aligned.***

If you want MicrobeTrace to align your sequences, select **Settings.** The sequence file is loaded with the default **Align** button set to none.

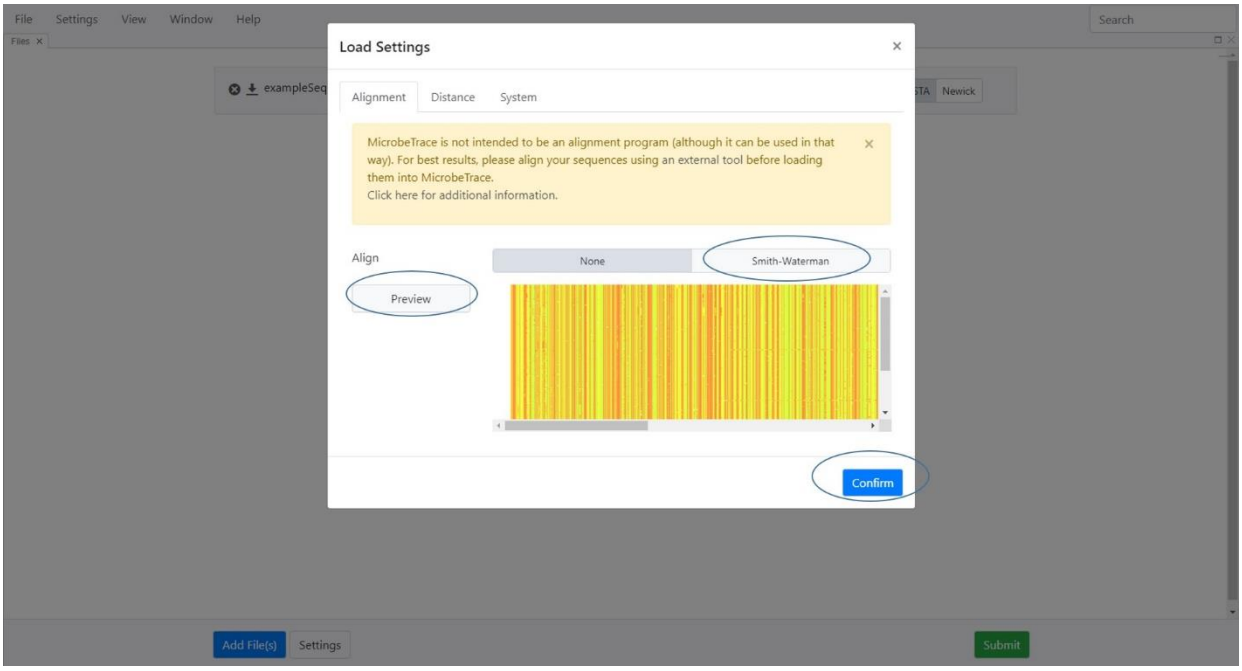


Fig. 19. Load Settings view for analyzing sequences with sequence preview selected

If you need to align your sequences, select **Smith-Waterman** (Fig. 16). The default setting is to align your sequences to the HIV-1 HXB2 *pol* region reference (HIV-1_HXB2 GenBank accession number K03455). If you wish to align your sequences to a reference of your choice, use the **Browse** button to navigate and load your reference.

You will see HIV (HXB2.pol) in the window. Select the **Confirm** button to accept this option. If you prefer to load a specific reference sequence file, select the **Browse** button and load your file from the appropriate location on your computer, then click **Confirm**.

Once you click Confirm, you will see the screen below (Fig. 20)

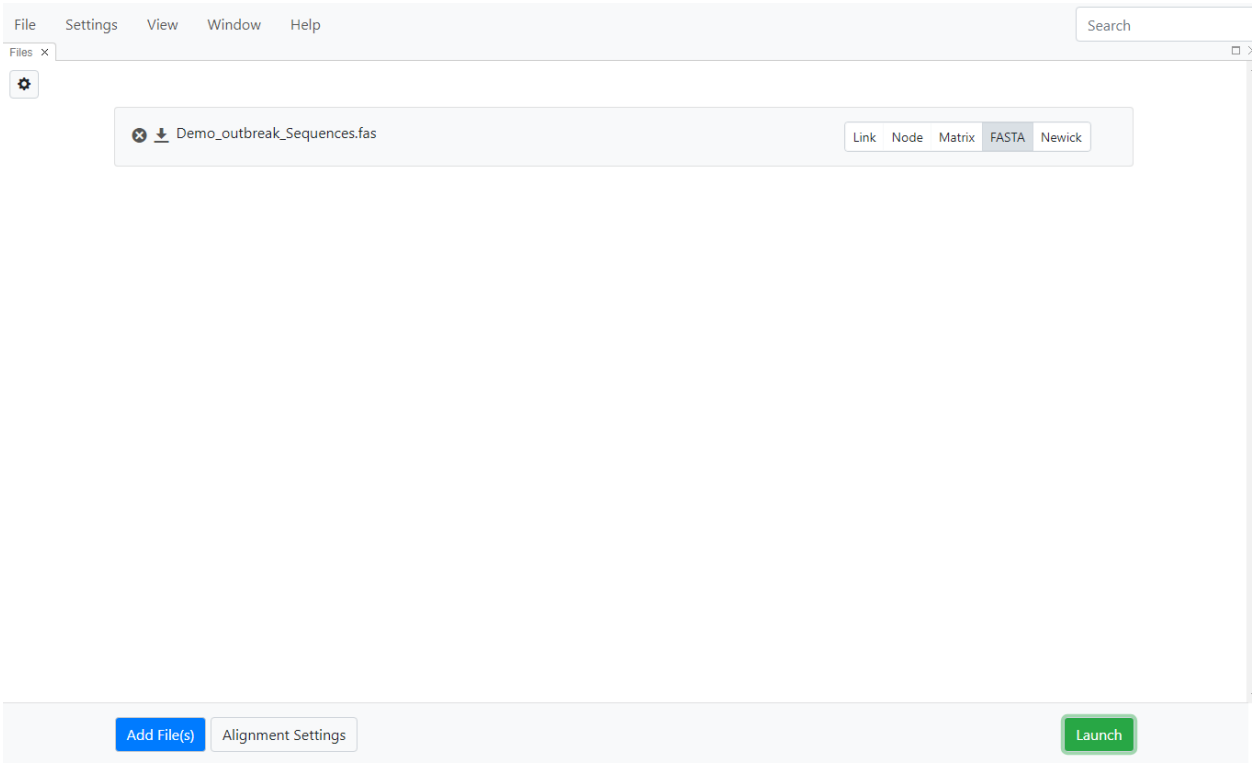


Fig. 20. Display after choosing alignment options for uploading/analyzing your sequences

Step 5. If you do not need to load a node CSV file and would like to visualize only the genetic distance network, select **Launch** now. If you would like to load demographic/epi data in a node file, proceed to the next section.

****IMPORTANT NOTE*: We STRONGLY recommend that you use pre-aligned sequences for non-HIV pathogens because MicrobeTrace is primarily not an alignment program for other pathogens.*** MicrobeTrace is currently configured for determining genetic distances between only HIV-1 *pol* sequences in the FASTA file with a reference HIV *pol* sequence (HIV_HXB2 GenBank accession number K03455) in the embedded alignment algorithm. For other pathogens, we recommend loading a **pre-aligned** nucleotide sequence file as the FASTA file input needed for MicrobeTrace and skipping the alignment options step and use the default settings (no alignment).

Loading a Node List and/or Edge List

If you have additional data associated with the nodes in the network, that data can also be imported into MicrobeTrace (Fig. 21). This data must be prepared in the CSV or Excel file formats and contain an ID column with values that match the source or target columns of the Edge CSV file and/or the sequence IDs in the FASTA file. If more than one ID column exists, the left-most (first) column in that file will be used. To load a node or edge file, select the Add File(s) button, or drag and drop into the main MicrobeTrace window as before. Fig. 21. shows the screen with three types of files loaded. MicrobeTrace tries to automatically determine the type of file that has been loaded (e.g., Node, Link, Distance Matrix, or FASTA) and selects it on the right side of the file upload menu and it will then be greyed out to show the selection. Please always check to ensure that the selection is correct.

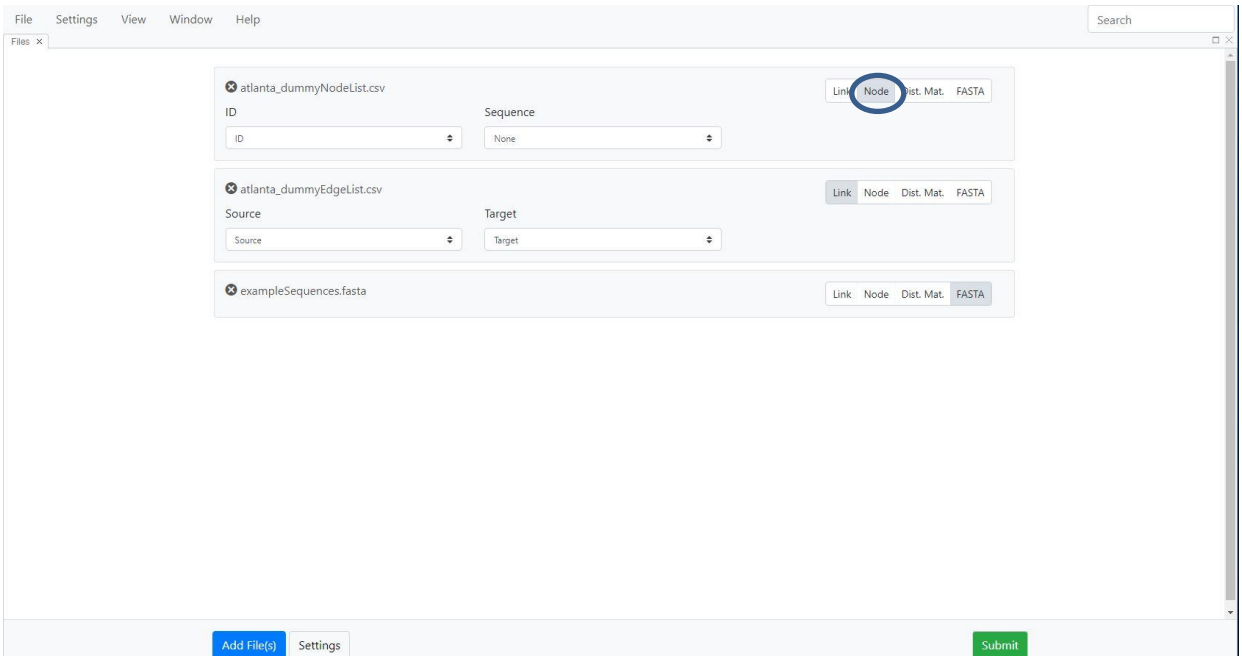


Fig. 21. Screen view with three types of files loaded (node, link, FASTA) and ready for analysis

***NOTE* If your node list file contains a column with sequences in it, MicrobeTrace will automatically recognize and analyze these sequences, and you will have all the functionality associated with a separate sequence file. In such situations, there is no need to load a separate FASTA file.**

In the **ID Column** drop-down menu for the node file, the default selection for node names is ID; however, if your node file has a different column heading for IDs that you wish to use, you may select that one using the drop-down menu.

If there is a **column with sequences in the node list**, make sure to select it from the dropdown menu once you load your node list so it is included in the analysis (Fig. 22).

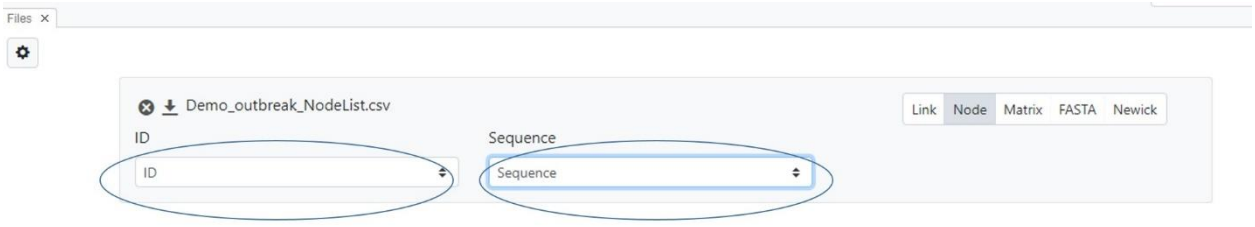


Fig. 22. Loading a node list with a sequence column.

***IMPORTANT NOTE*: Rows in a node list with identical node IDs cause previous rows with the same ID to be overwritten. Please ensure that node IDs are unique.**

Select **Submit** to start the analysis. A status bar will appear to show the progress of the analysis as the files are loaded and the genetic and/or social network is inferred (Fig.23).

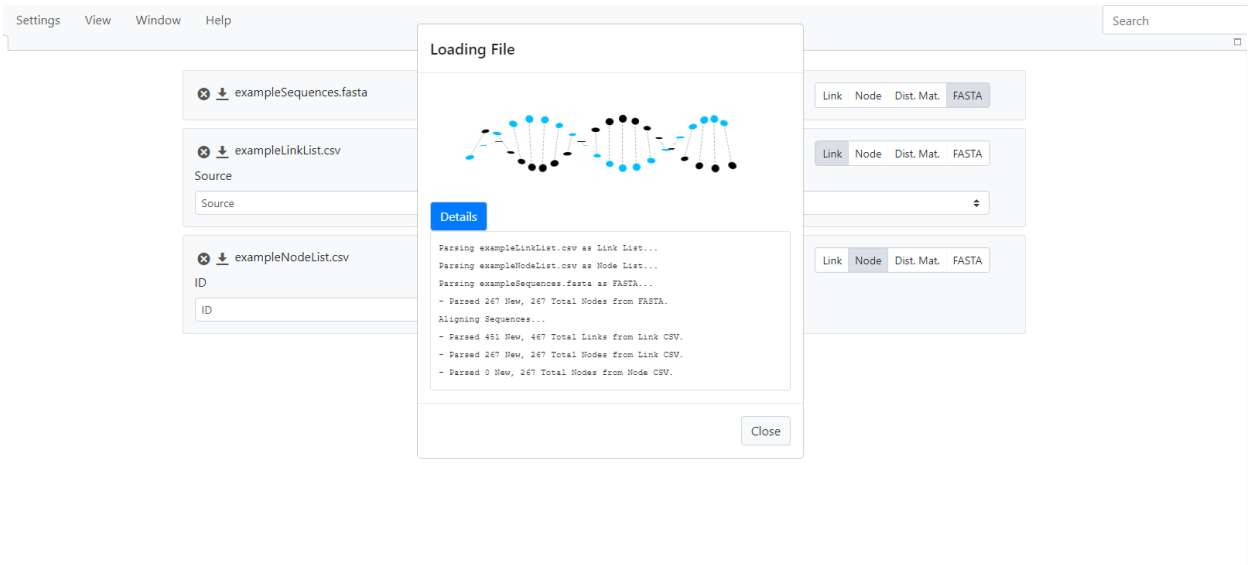


Fig. 23. Status window showing progress of file loading and data analysis

Loading an edge list instead of a FASTA file

If you have already prepared an edge list (in CSV or Excel format), then you can load that file directly into MicrobeTrace (Fig. 24). The link file must contain a source (e.g. person with the infection) and target (e.g. recipient of the infection or contact of the infected person) column. Any additional edge properties can be included as additional columns in the edge list file. Select the Add File(s) button, navigate to your CSV file and double-click on the file, or select the file and then select Open. Once the file is loaded, you can either click Submit to run the analysis, or you may add a node file if you have demographic and epi data linked to persons in the study. To do this, please refer to the preceding section on loading optional node data.

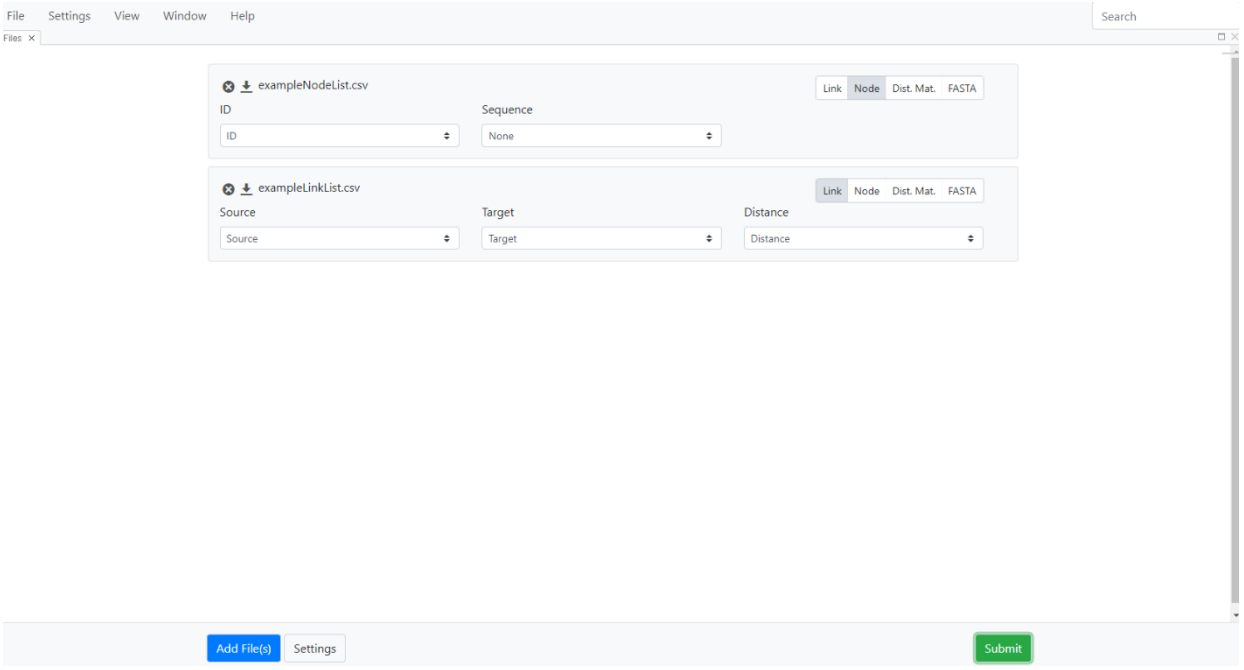


Fig. 24. Screen view after loading link and node files

Once the file loading and data analysis is complete, the software displays a network diagram and a summary statistics table at the bottom-right corner of the page (Fig. 25). This table displays the number of nodes, the number of links (edges), clusters and singletons.

Data Visualization

The default data visualization and exploration method is the two dimensional (2D) Network View (Fig. 25). You can select a number of different data visualizations from the View menu to display the data as a table, a flow (also called alluvial) diagram, a histogram, a heat map of the pairwise genetic distance matrix, an alignment of the nucleotide sequences, a phylogenetic tree of the genetic relationships of the nucleotide sequences, or a geographic map showing the location of the nodes. Some of these viewing options are specific to file type. For example, histogram, sequence view, heat maps or phylogenetic trees require nucleotide sequence data. The additional data visualization views are described in the following sections. Each time you select a view, it opens in a new tab, so you can move easily between views. You can also click on the tab and drag it to create side by side windows, to enable you to compare views. It is often helpful to visualize data in two or more different formats together. You will see examples of this feature in later sections describing individual views.

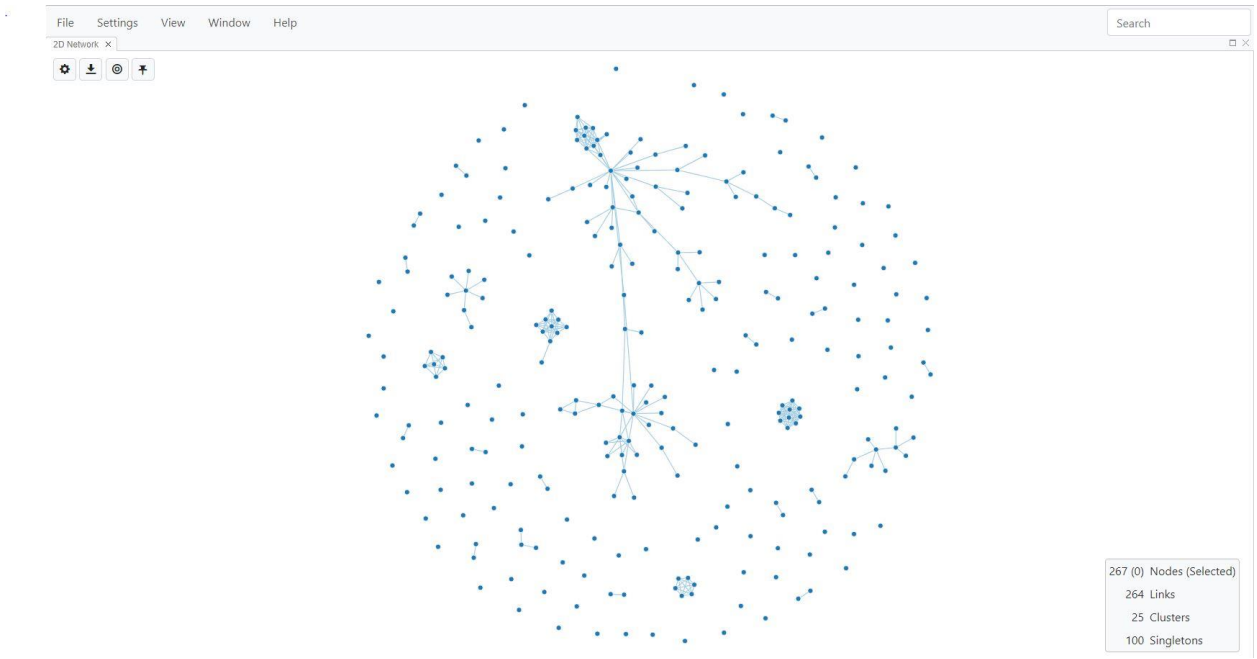


Fig. 25. Default view-2D network

Tiling Different Views

Each time you select a view from the main dropdown menu, a new tab opens. You can switch between views by moving between tabs, much like you would in a browser. In addition,

MicrobeTrace also allows you to visualize multiple views side by side, or in a tiled arrangement in the same window. You can do this by clicking in a view tab, and simultaneously dragging it to a spot in the window either next to the currently open view, or below it. In the example below (Fig. 26), a 2D network window is open, then the Table and Map Views are opened in new tabs. Go to the 2D network window, click on the Table View tab, and drag it so it's positioned to the left of the network view. Now click on the Map View tab and drag it below the two open windows. You can tile as many views as you wish.

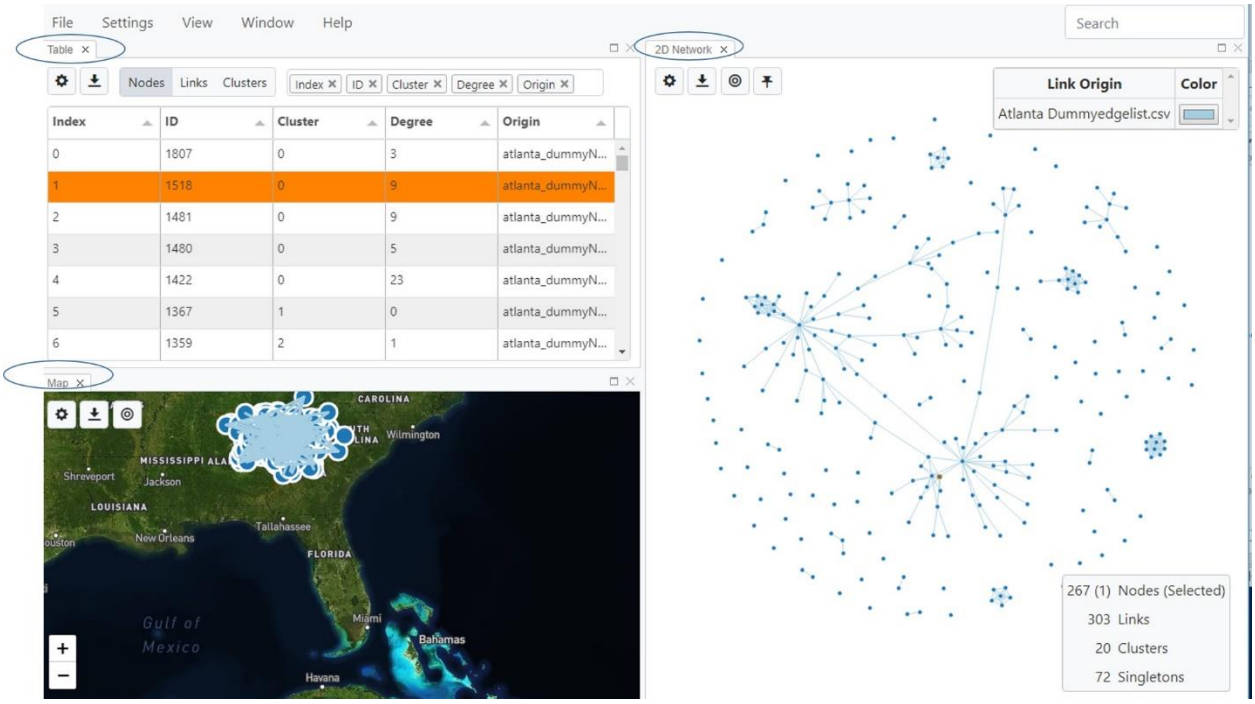


Fig. 26. Tiling views by selecting multiple views, and arranging them in the same window so they are all visible.

All views contain two or more of these four icons  on the top left corner of the screen (Fig. 27). Hovering over each of these with your mouse will give you descriptions of each icon.

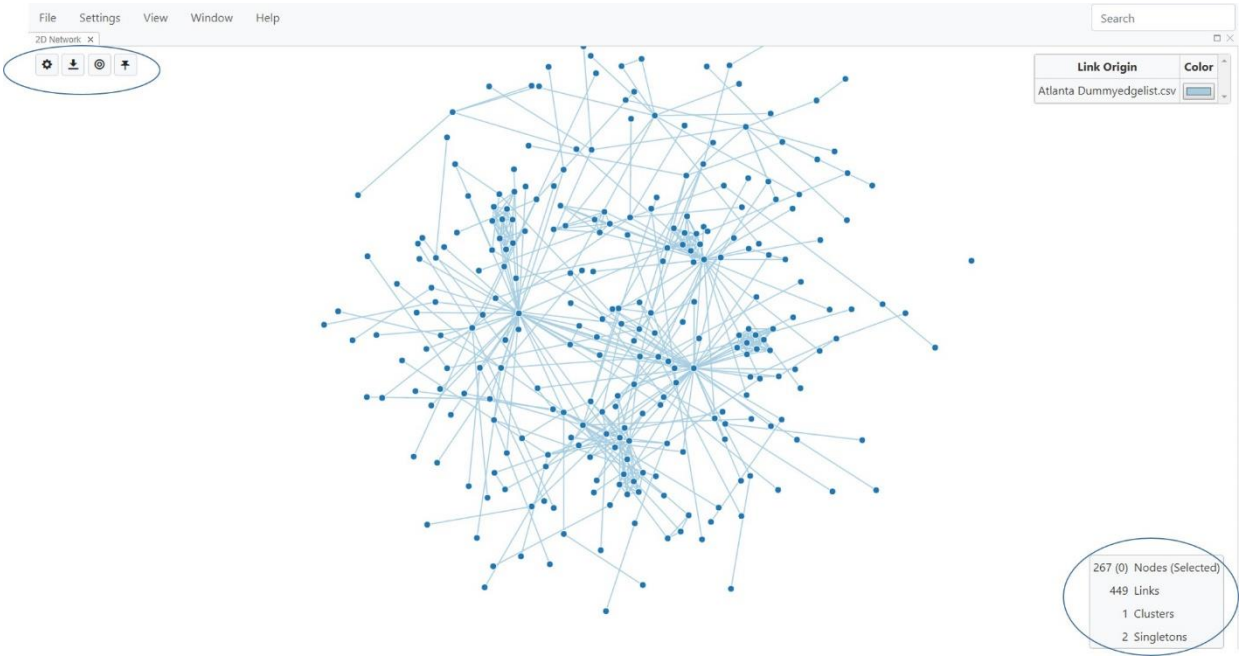


Fig. 27. 2D Network View with settings buttons circled in the upper left corner of the window

1. **Toggle Settings:** Lets you adjust various parameters like shape, size, color, and physics (charge, friction, and gravity) for nodes, links and the network itself. Details are provided in the Network Configuration section below. In other views, this button lets you customize settings for that particular view. Selecting the Toggle Settings button once displays the menu and selecting it again hides the menu, hence you toggle between displaying and hiding the menu.
2. **Export**: Lets you export the current view as an image, or as other file types as applicable. You can select file type using a dropdown menu. If you are exporting as an image, advanced settings allow you to adjust size depending on your application. There is also a .svg option which allows you to increase image size without loss of resolution. This is particularly useful for posters and publications. The downloaded image will include the MicrobeTrace logo as a watermark. You can adjust the opacity of the MicrobeTrace watermark using a slider bar in the export settings.
3. **Center and Scale**: You can zoom in and out on the network using the mouse roller. This button re-scales and centers it back to the original size on the screen.

4. **Pin all Nodes**:  The network is a dynamic structure, and you will see it floating slightly as it renders. This button will "freeze" the network. You can do this after you drag nodes around to give you the best visual for your data (see section on nodes below).

2D Network View

This is the default visualization window you will see after MicrobeTrace processes your data (Fig. 26). In the Network View, you can:

- pan around by clicking anywhere in the window, drag nodes around or zoom in or out by using the roller on your mouse
- select or de-select individual nodes (Use Ctrl+Click to select individual nodes; use Shift+Click to select multiple nodes). You can select multiple nodes and see them in a different view, which can be especially useful in Table View described in more detail below). This node selection feature allows you to view all the characteristics of only the selected nodes. You may want to look at the epidemiologic data for just the nodes in a cluster or any that you find interesting in the Network View.
- right-click on a node to see various options (Fig. 28)
    - **Pin node:** Pinning a node allows you to drag it out for better visualization of that node in a cluster and to explore the cluster.
    - **Copy ID:** Copy the node ID
    - **View attributes:** allows you to view node properties (metadata associated with that node)

Note that the rendering of the network is updated each time you drag a node, so you will see the nodes moving around before settling into a static position. You will see this rendering each time you change any of the network settings (see below for options in settings). You can use the **Pin All Nodes** icon to freeze the network.

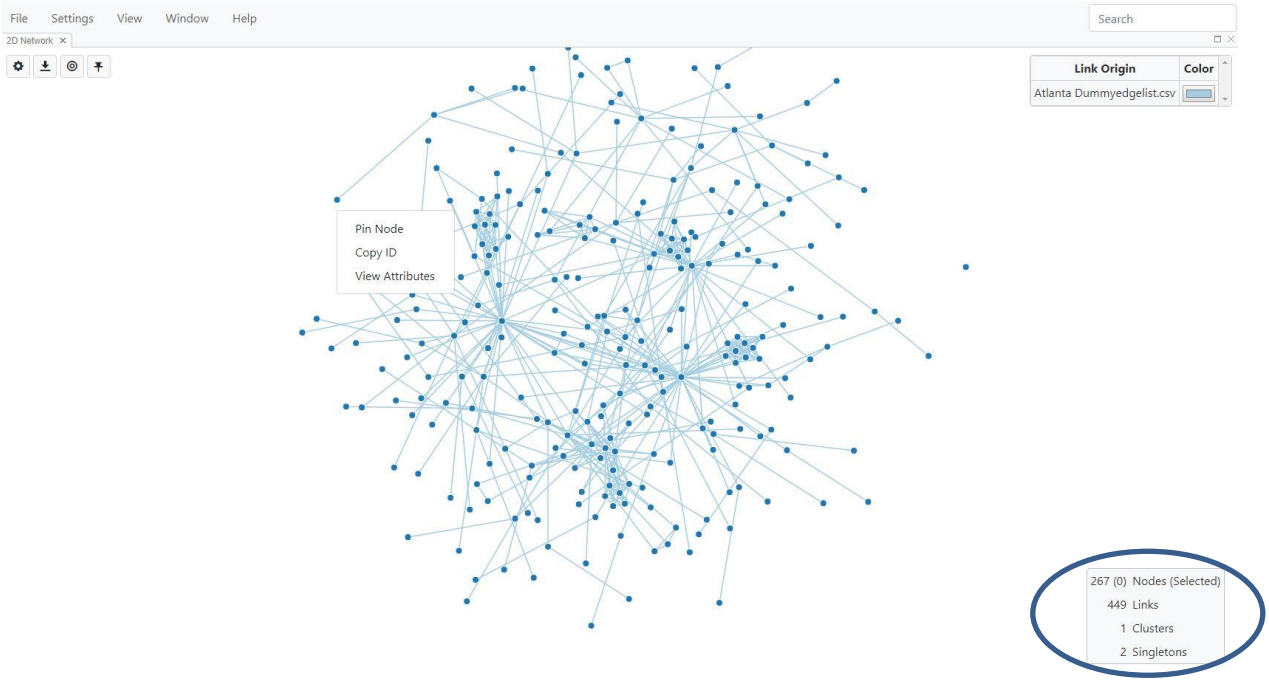


Fig. 28. Viewing and exploring node options by right-clicking on a node

Network Configuration

You can modify multiple visual characteristics of the displayed network. First, we will describe the Global Settings menu, and then describe settings specific to nodes, links and networks.

The Global Settings menu is accessed by selecting Settings on the main menu bar in the upper left window. This menu option has three tabs: the **Filtering** tab, used to modify cut-off values and other features of the network, the **Styling** tab used to color nodes, links and background (the Styling tab is described in the following section on nodes), and the **Timeline** tab, which gives you a time player to track cluster growth over time.

***Filtering*:**

This tab lets you set distance thresholds, prune links using the Nearest Neighbor algorithm, and set the threshold for the minimum number of nodes you would use to define a cluster. This feature allows you to filter by cluster size, which could be useful in some outbreak scenarios, especially if you want to remove all singletons from the Network View. This tab also allows you to show or hide the network statistics that are automatically generated during the analysis and shown in the box in the lower right of the window (see circle in Fig. 29).

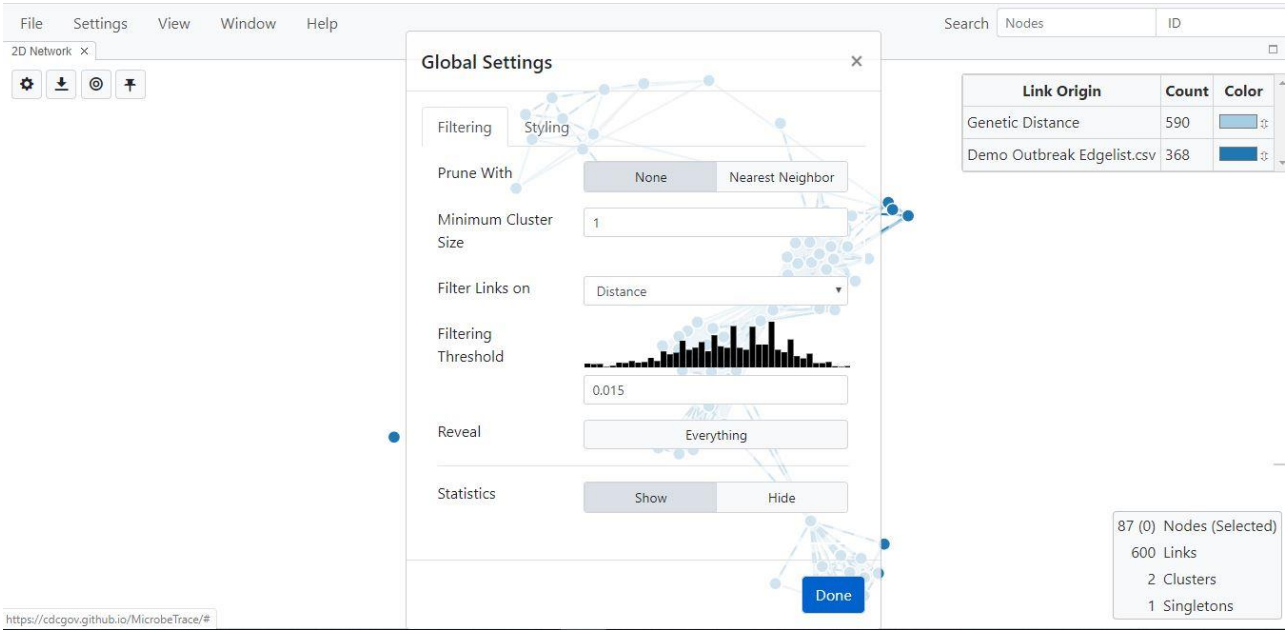


Fig. 29. Global settings, Filtering tab.

Minimum Cluster Size: Allows you to set the number of nodes that you use to visualize a cluster. The default is set to 1, which means singletons are also displayed. To hide singletons, or nodes that are not connected to any others, you should set this number to 2 (clusters of size two or greater will be displayed).

Timeline tab

The Timeline tab allows you to monitor the growth of a network and its clusters over time. It provides a snapshot of transmission at any given point within the date range of your data. You can select from any date field in your dataset using the dropdown menu (Fig. 30). Once you select an attribute, a time player is displayed at the bottom of the screen. Please note that the time player will also be visible in other views. This can be a valuable feature in the map view to look at spatiotemporal growth in networks.

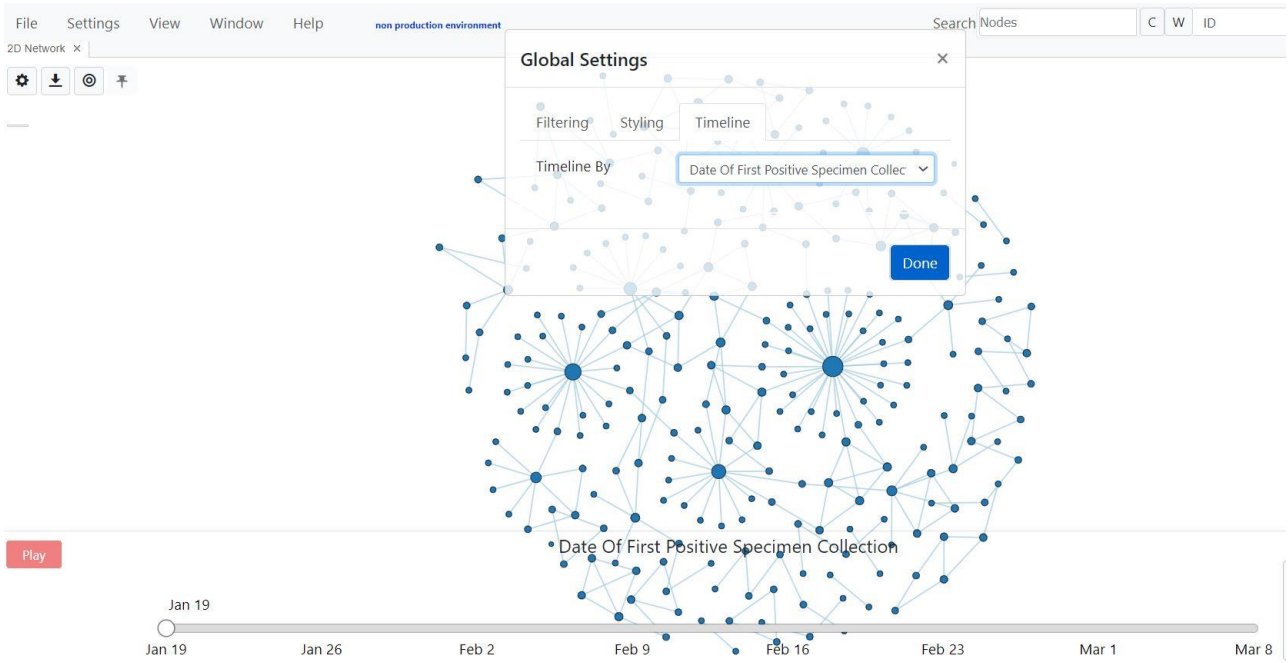


Fig. 30. Timeline tab on the Global settings menu. You can select a date field from your dataset.

Clicking on the **Play** button will start the time player and you can see the network changing over the date ranges in your dataset (Fig. 31). You can pause, rewind or advance.

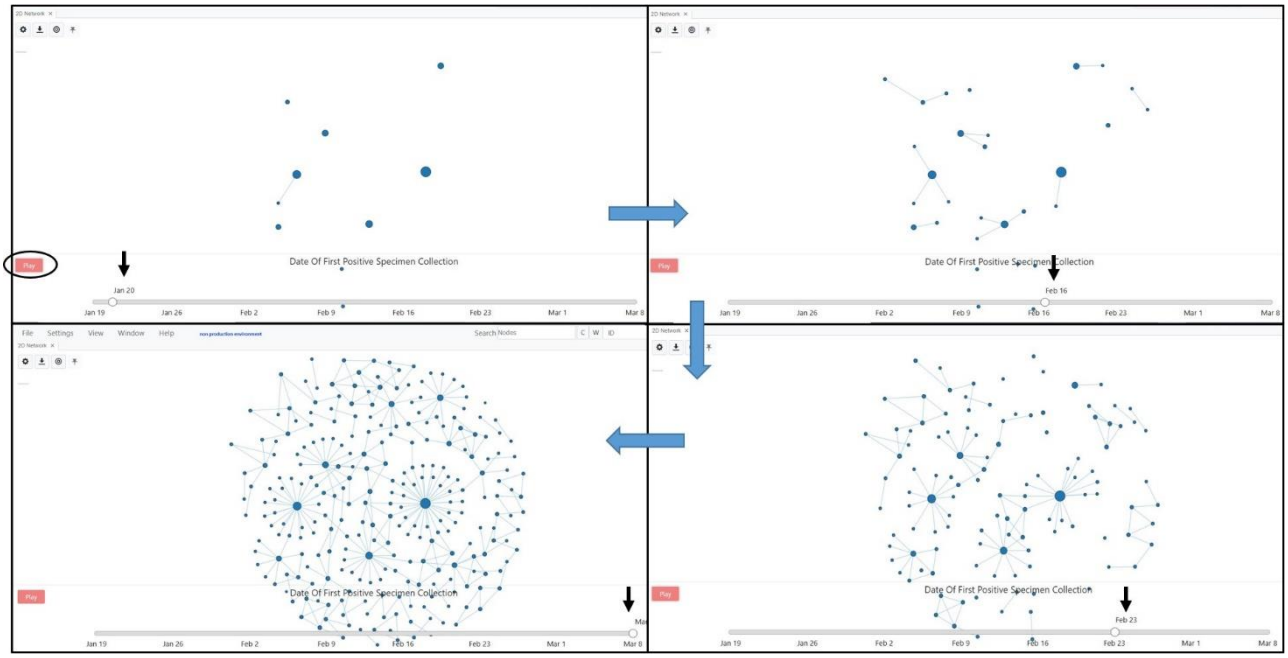


Fig. 31. Screenshots from the time player of the network at various timepoints in the date range. You can see how links form over a period of time as the network changes.

Node and Link Settings

Select the Toggle Network Settings icon mentioned earlier to display a context menu (Fig. 32). You can choose from three tabs: "**Nodes**", "**Links**", or "**Network**" to adjust the settings of the various network components. Using your mouse, hover over each property name to see what it represents.

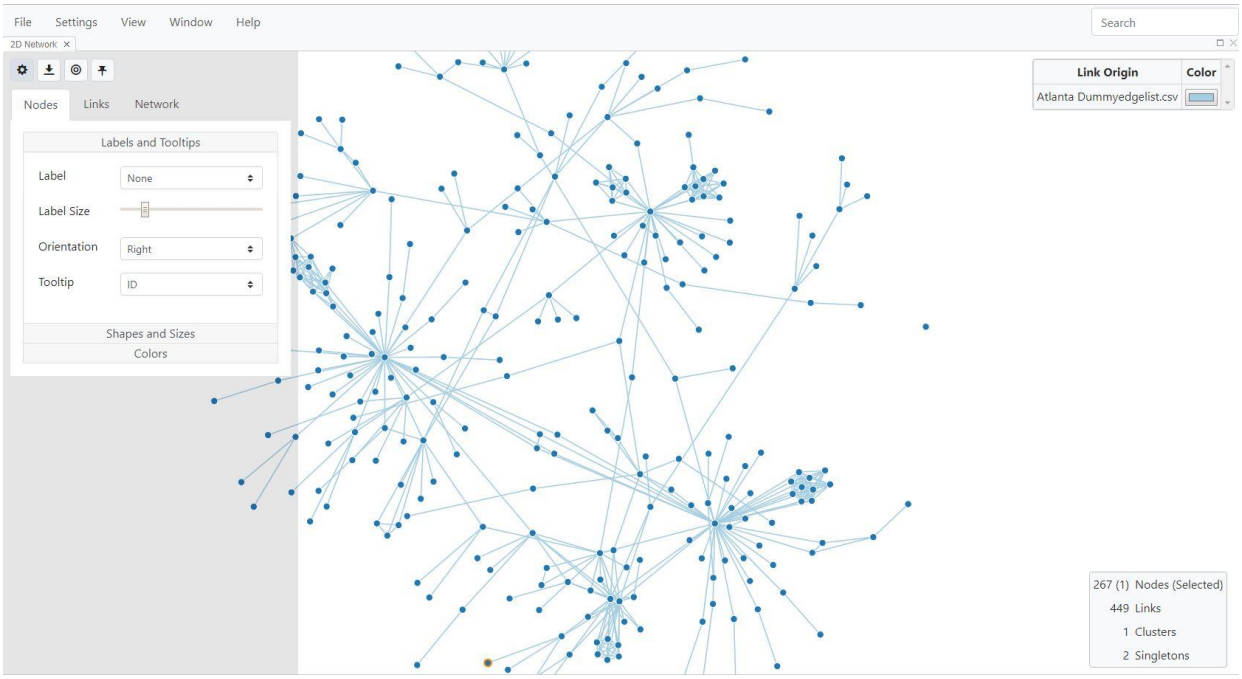


Fig. 32. Available settings for changing the network configuration in the Nodes tab

Node Properties

Labels and tooltips: You can change labels and tooltips of the node by selecting from dropdown menus for each parameter. You can choose more than one variable for tooltips. You can use the slider bar to change label size. You can also select the orientation of the label relative to the node.

***Shape and size*:** You can select **Shapes and Sizes** to map shapes to the nodes and map sizes of the nodes to demographic characteristics picked from the dropdown menu. The default shape of the node is a circle, but can be changed using the dropdown menu. You can set the minimum and maximum sizes of nodes so that all nodes are visible even when sized by variables. You can also change the thickness of the node border.

In Fig. 33, shapes have been mapped to gender, and a key is displayed in the top right corner of the window. Nodes are sized by degree (number of links to a node).

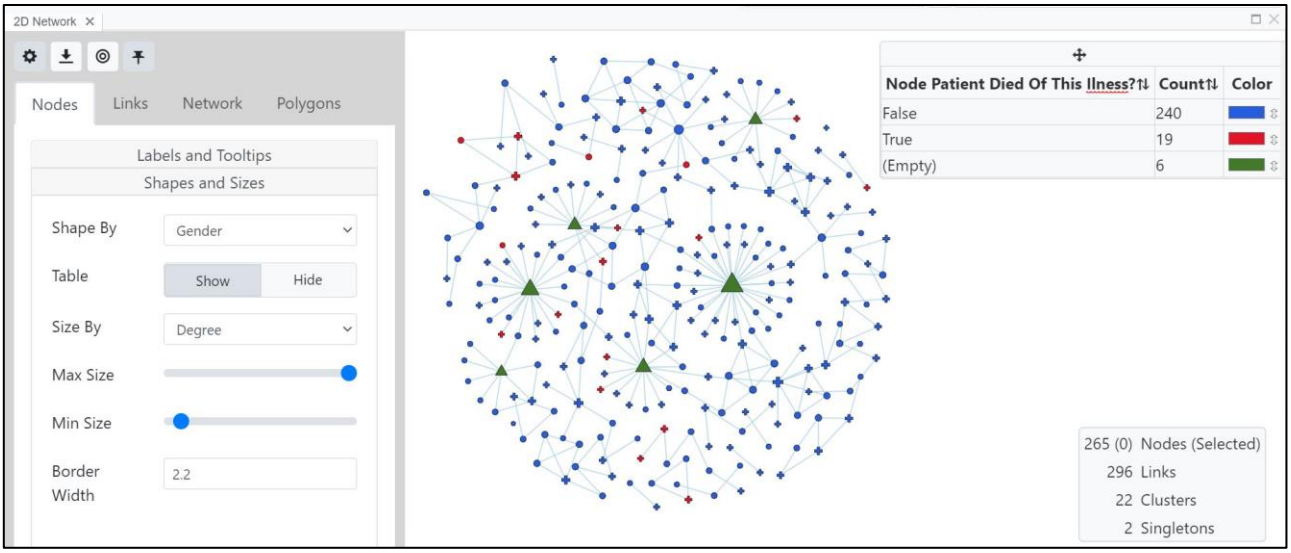


Fig. 33. Shapes and Sizes menu for changing node characteristics with node shapes mapped to gender, sized by degree, and border width increased.

***Colors*:** Selecting the **Colors** button takes you to the styling tab of the **Global Settings** menu where you can use dropdown menus to map the color of nodes or links to demographic data. You can also change node border color as well as background color (Fig. 34). There is also an option called **Apply style** where you can browse and choose a previously saved MicrobeTrace style file. This will enable you to recreate that style with the current dataset. **Importantly, you must remember to use the same file types (i.e. node vs link) when you apply the style file in your new session or the saved styles will not apply correctly.** The process for saving a style file is described here.

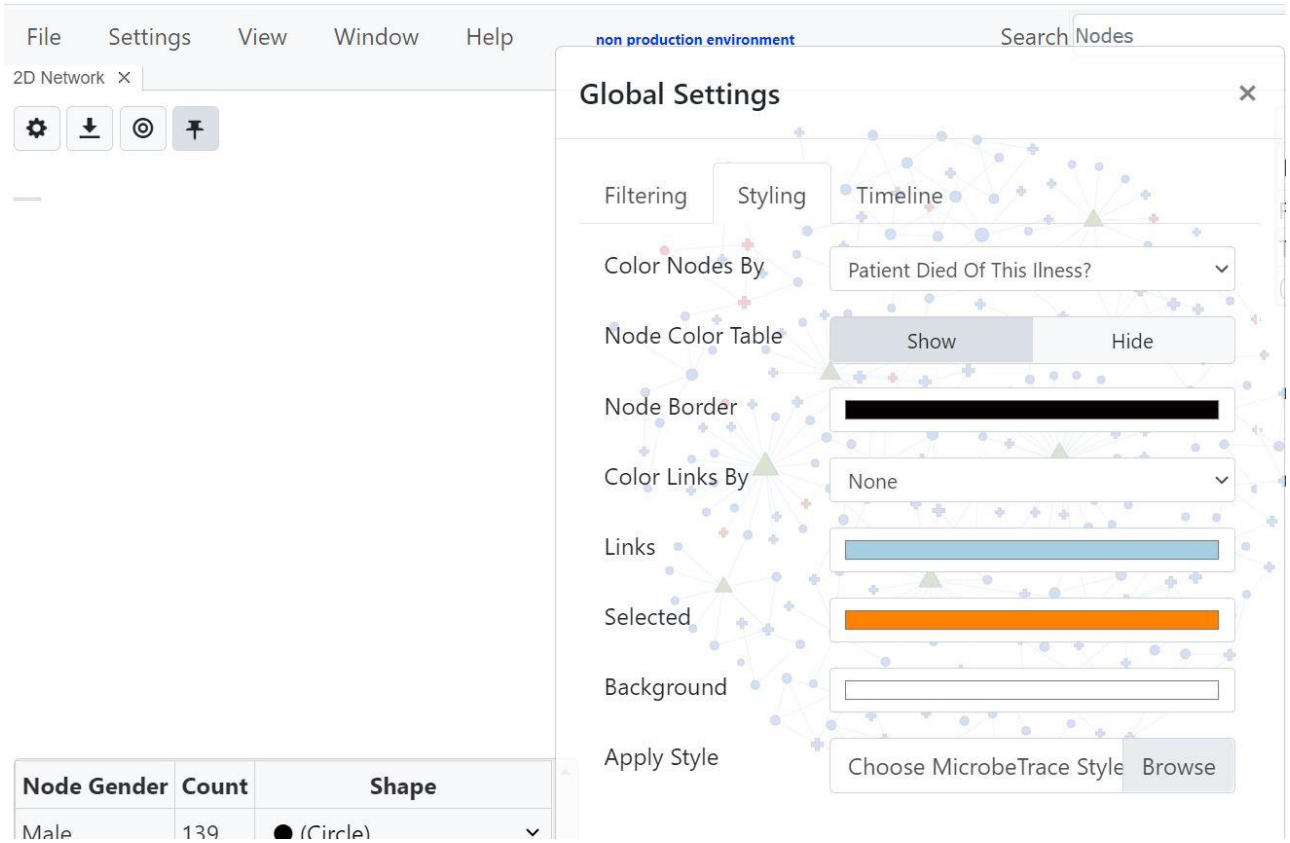


Fig. 34. Colors and styling of nodes, borders, links and background

In the example below, node labels have been set to ID from the **Label** drop-down menu and have been colored by risk factor, by selecting **Risk Factor** from the **Color Nodes By** drop-down menu, and then selecting **Done.** You will then see the ID labels for the nodes on the network and a color-mapping key will be available in a text box in the top-right corner of the window (Fig. 35).

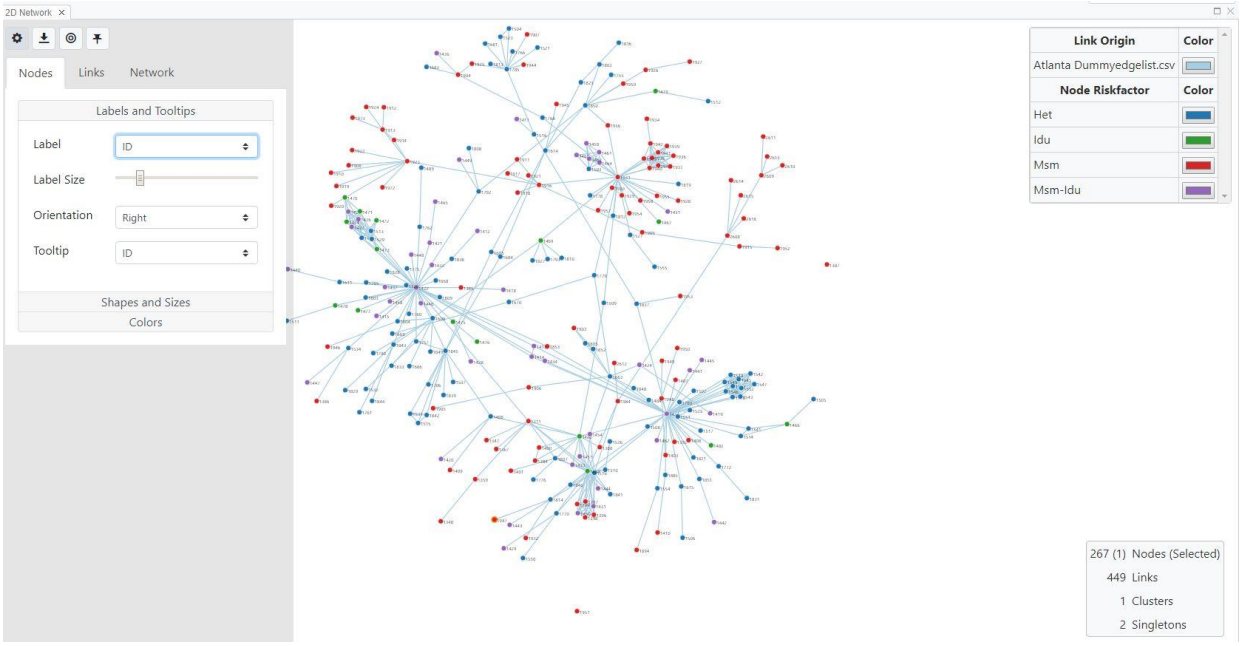


Fig. 35. Example of node settings with nodes labeled with the ID and colored by risk factor

The key on the top right corner can be edited (Fig. 36). You can change colors of each variable by clicking on the color bar, which then pulls up a color chart to choose from. Clicking on the two-sided arrow next to the color will give you a transparency slider to change the transparency of that color. You can also edit the text column by clicking on it.

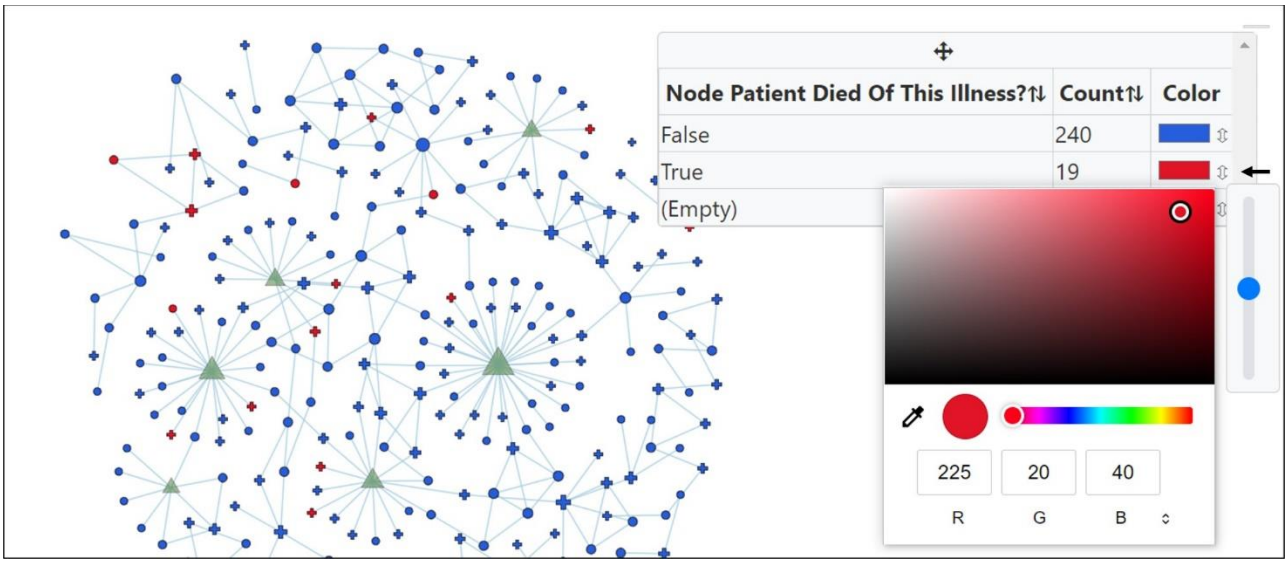


Fig. 36. Editing color key

This key will be displayed at the top right corner of each fresh view that you open. You can move, hide and unhide it, and expand it to show all variables by right clicking on the box and selecting Drag, Pin, Hide, Expand, Toggle Counts and Toggle Frequencies. Once hidden, you can go to the Global Settings Menu and select **Reveal Everything**. Toggling allows you to select whether or not to display node counts and frequencies for all the categories in the key (Fig. 37) shows two options). All key boxes can also be moved using the arrow keys on your keyboard, allowing for finer control and positioning of boxes.

Fig. 37. Moving and expanding the network key box/table. **A.** The options are Drag, Hide, Expand, Toggle Counts and Toggle Frequencies. **B.** Once the box is dragged to a location of your choice,

right clicking on the box shows the available options Pin, Hide, Expand, Hide Counts and Hide Frequencies.

Link Properties

Genetic links or edges are typically generated using a defined nucleotide distance cut-off. The **Links** tab on the Toggle Network Settings Menu lets you customize link colors, width, etc. and to map these properties to demographics just as you can with nodes (Fig. 38).

***Labels and Tooltips*:** These are customizable as they are with nodes.

***Shapes and sizes*:** You can user slider bars to change the transparency, width and length of links. For example, you can increase the length of the links (drag the slider bar on the **Length** option) if your clusters are too tight (very dense). This option allows the cluster structure to become more open (less dense) so the cluster nodes and edges are more easily viewed. You can also map width to any variable in your link list.

Colors: Selecting **Colors** will take you to the styling tab of the Global Settings Menu as described in the nodes section above.

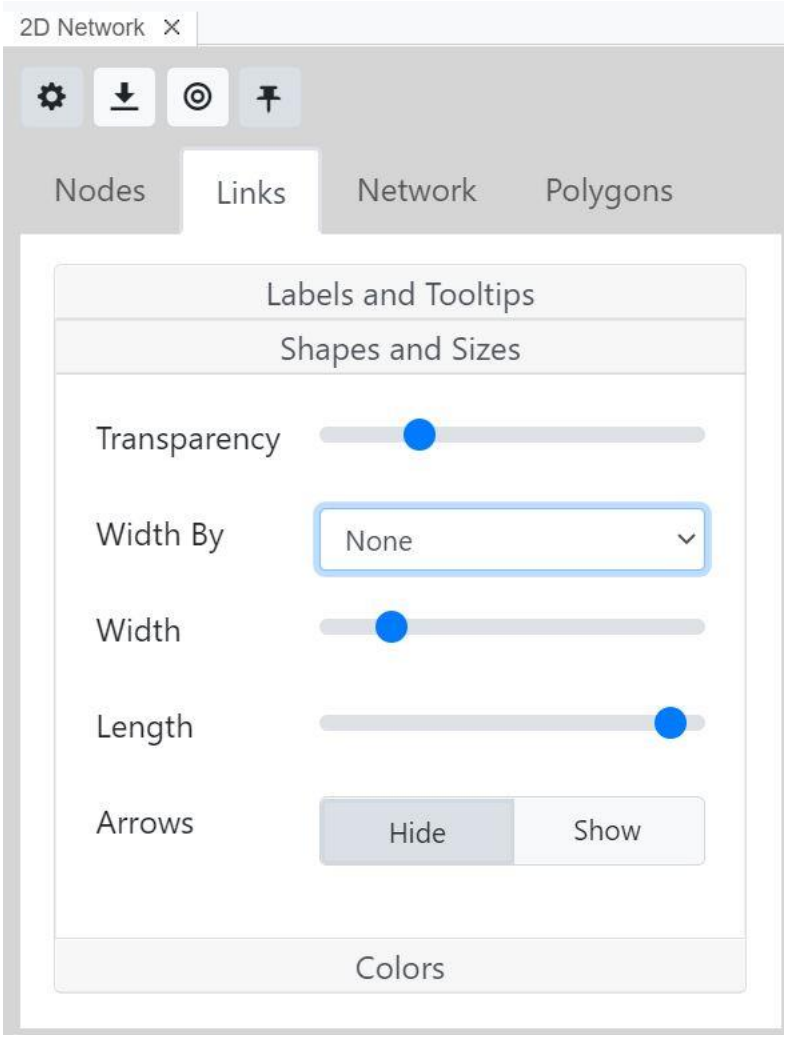


Fig. 38. Options to customize link properties and enable directionality (arrowheads)

Enabling directionality in the network *(PLEASE READ IMPORTANT CAVEATS IN THE GLOSSARY ABOUT INFERRING TRANSMISSION DIRECTIONALITY)*

Directionality is only valid for edge lists that contain contact tracing data, and not for datasets containing only sequence data.

The default Link setting for HIV-1 analyses is that links are undirected. If you upload an edge list containing contact tracing information, then you can use the **Arrowheads** button (Fig. 39) to predict directionality. Before enabling this feature, please see the notes above about the limitations of using directionality to explore HIV or other pathogen transmission.

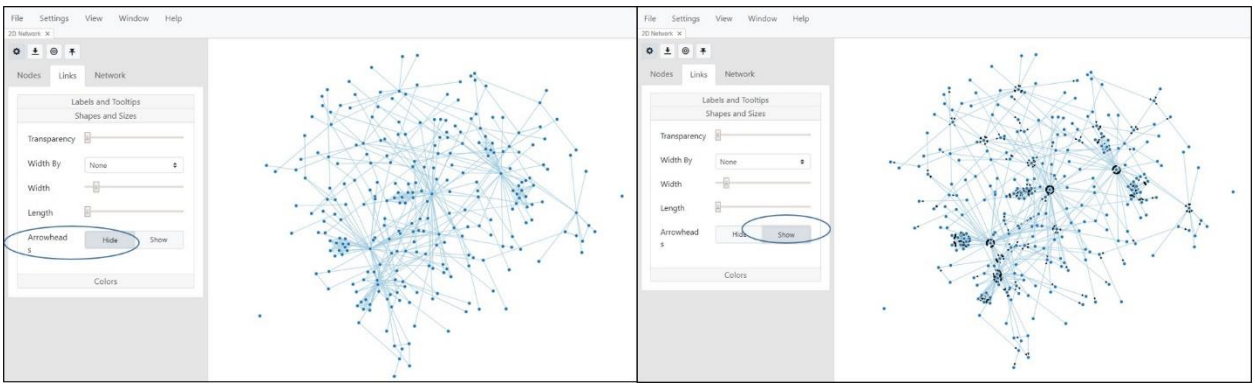


Fig. 39. Link settings without (left) and with (right) arrows between links for an edge list containing both genetic information as well as contact tracing data.

Network Properties

***Neighbors*:** Use this feature to adjust display options for neighboring nodes. Selecting **Normal** will display all nodes when you hover on a specific node. However, if you set this parameter to **Highlighted**, hovering on a specific node of interest in the network will highlight all its neighboring nodes, or nodes linked to your node of interest (Fig. 40).

Gridlines: Default is for gridlines to be hidden, but you may choose to display them by selecting **Show** (Fig. 40). The grid utility enables you to align nodes and clusters in a quantitative way, both visually and analytically. It can also help reproduce visualizations in subsequent analyses.

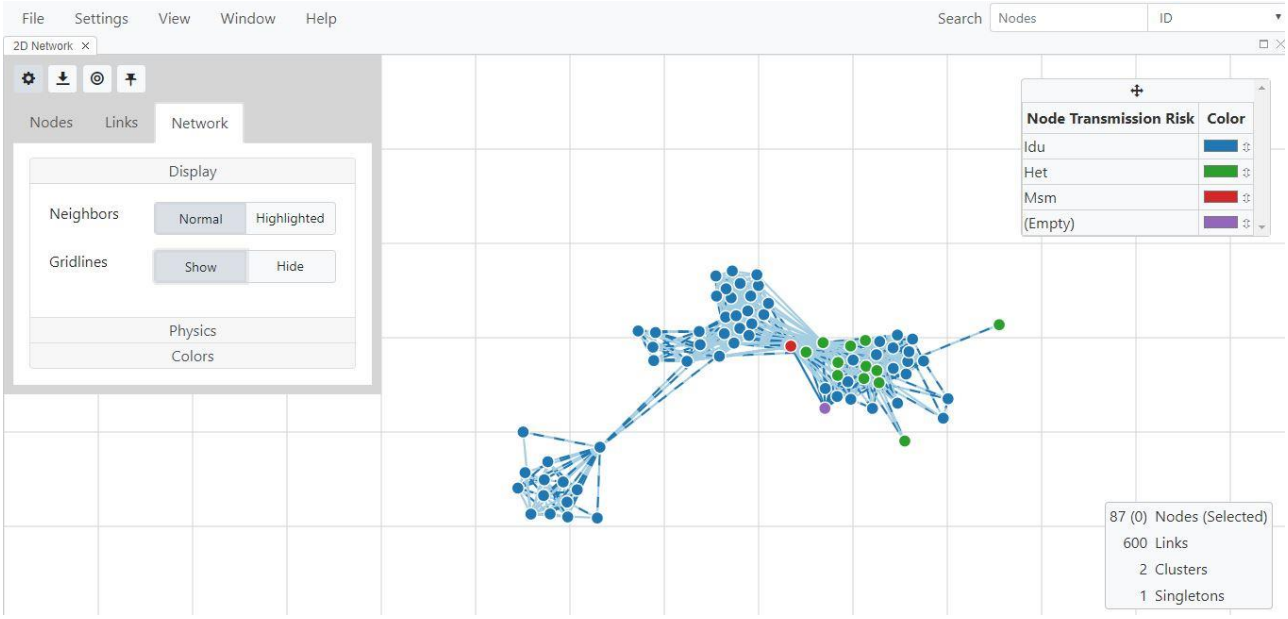


Fig. 40. Network display options to highlight neighboring nodes; gridlines are displayed.

***Physics*:** This button lets you change the friction, charge and gravity of the network, which are parameters that determine how densely packed the nodes are in the network. Learn more about these properties in the glossary by clicking the highlighted links for these terms in the previous sentence (Fig. 41).

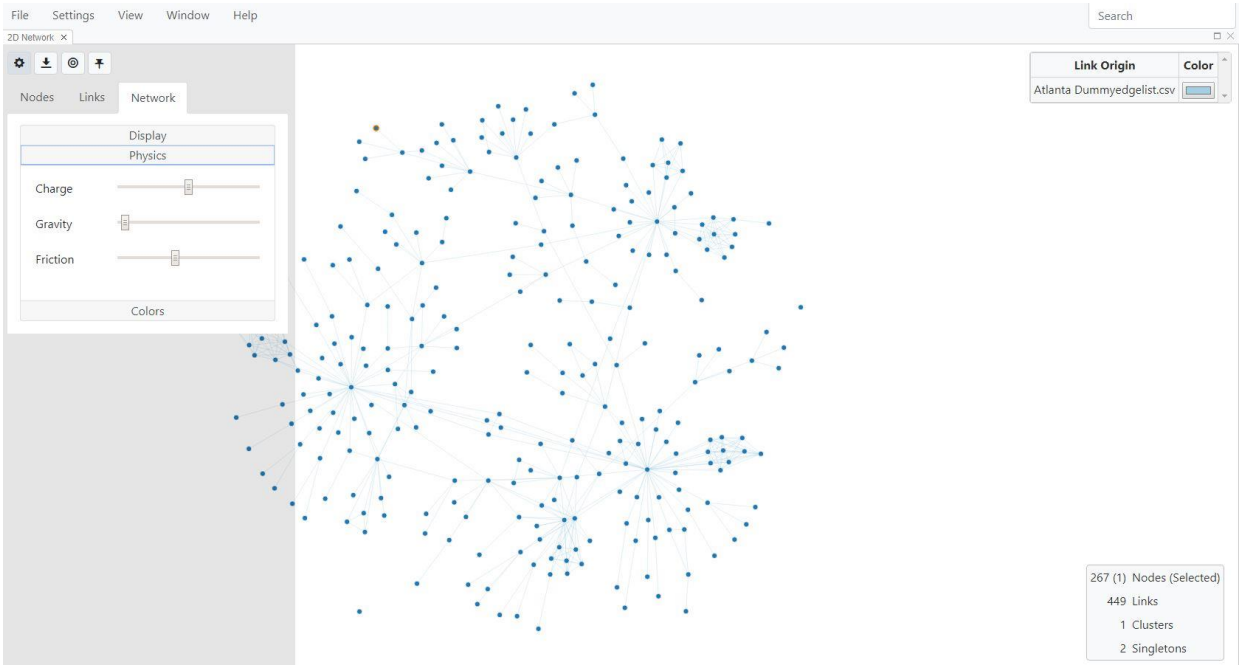


Fig. 41. Network options for changing the physical properties of the network.

Colors: Selecting **Colors** will take you to the styling tab of the Global Settings Menu as described in the nodes section above.

Polygons

The Polygons tab allows you to group nodes by any node attribute of your choice in your dataset, selectable via a pull-down menu (Fig. 41A), and MicrobeTrace draws a colored polygon around the group or groups. The gather slider bar lets you adjust the physics of the nodes inside of a polygon. Increasing the gather pulls the nodes within each polygon closer together. You can also have the polygons colored differently and labeled (Fig. 41 B and C), with the ability to change polygon colors and transparency via the color chart in the key. Label orientation and size are also adjustable.

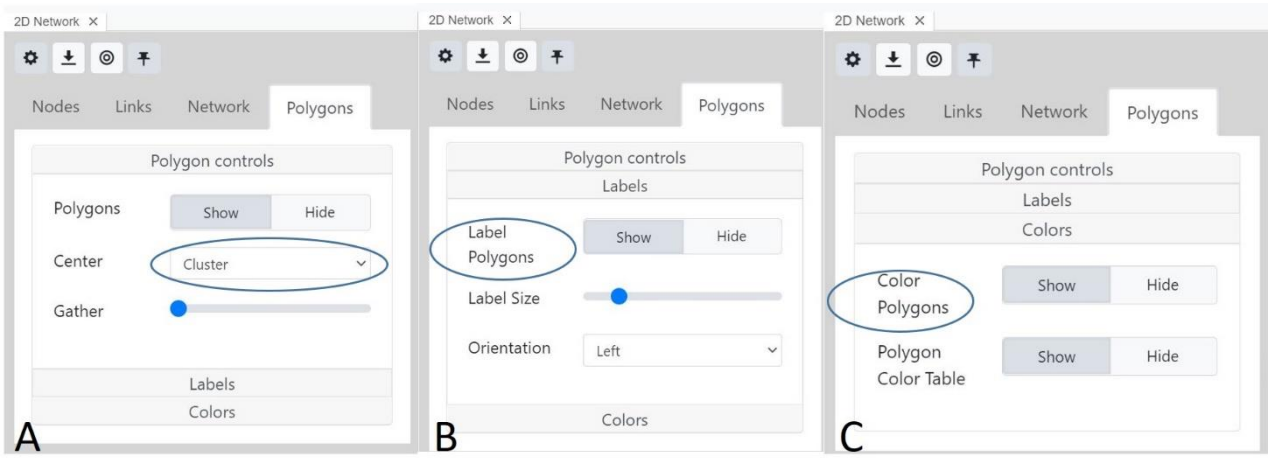


Fig. 41. Polygons settings. The polygons tab has three sub tabs **A**. Polygon controls: select the attribute by which to group nodes and adjust gather strength. **B**. Labels: select label size and orientation. **C**. Colors: Modify polygons and colors and transparency.

The figure below shows an example of the polygon feature where nodes are grouped by cluster, labeled in the center, and transparencies adjusted to show labels clearly.

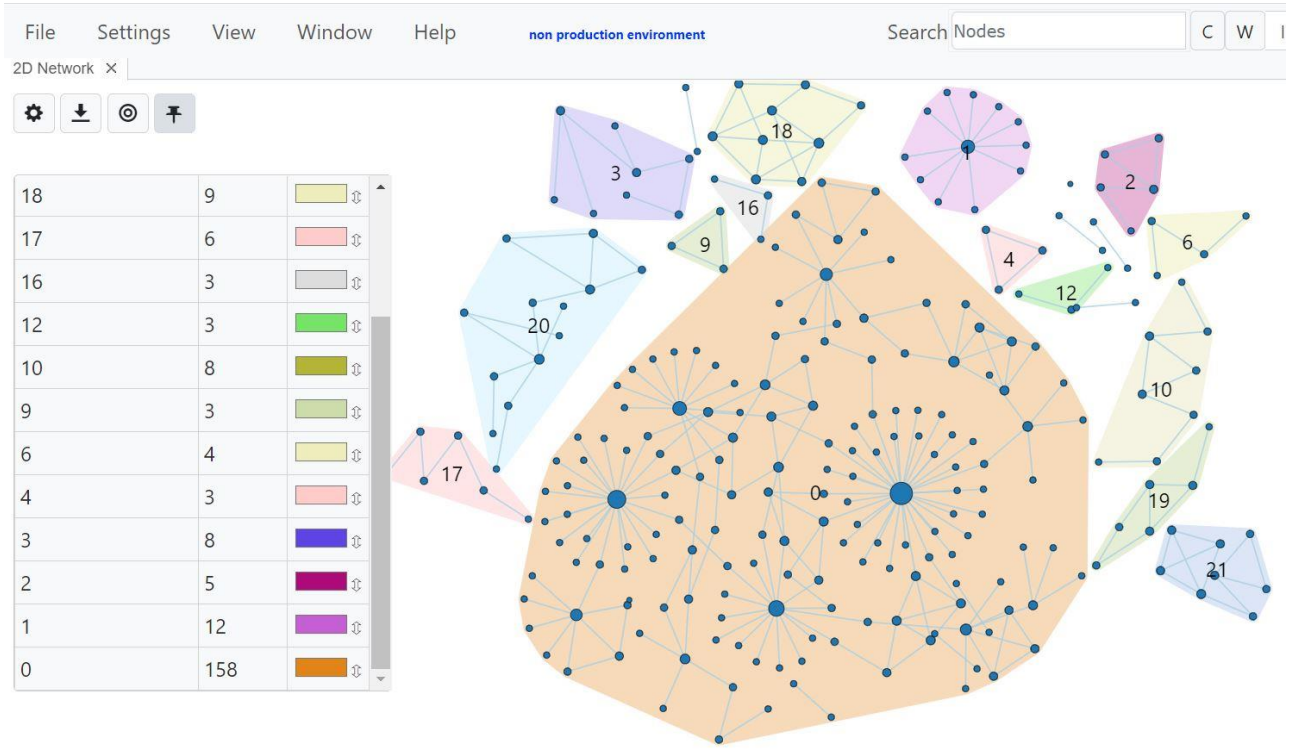


Fig. 42. Nodes are grouped by cluster, gather strength has been increased slightly to separate the polygons, and polygons are colored and labeled.

3D Network View

You can generate a 3D version of the Network View by selecting **3D Network** from the drop-down **View** menu (Fig. 43).

Fig. 43. Selecting the 3D Network View

The network will be rendered in 3D, which can sometimes be useful for further exploration (Fig. 44). In the 3D view, you can zoom in or out by using the mouse wheel. Although you cannot reposition nodes in the 3D view, you can rotate the network to find the position that gives you the best visualization of clusters and links.

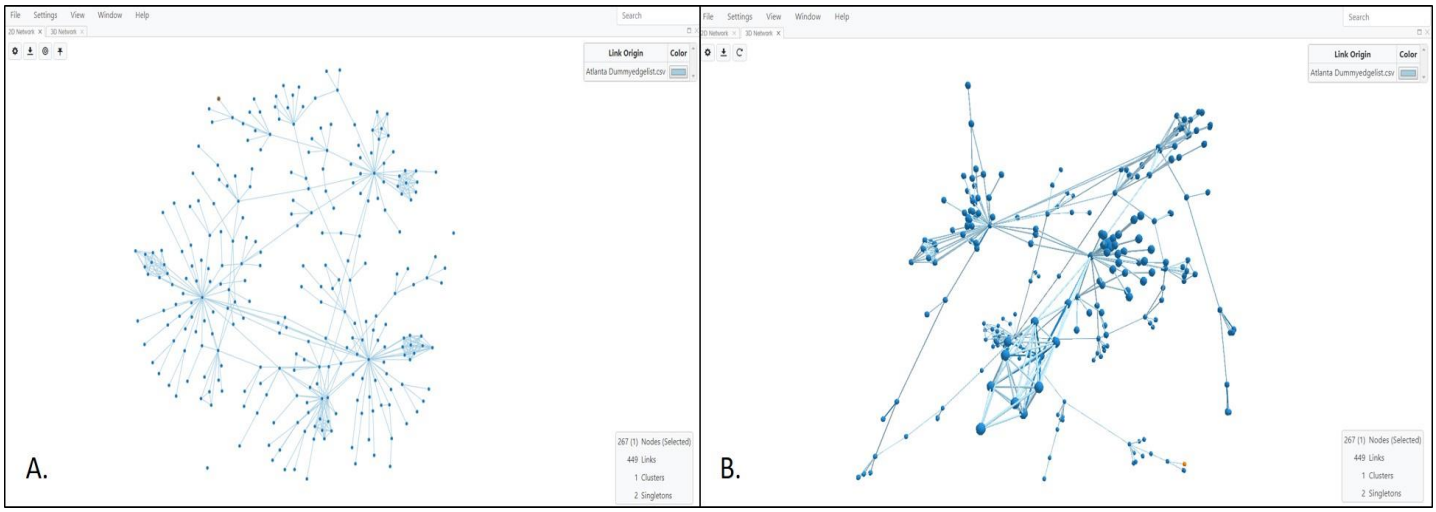


Fig. 44. 3D Network View. **A.** Standard network view. **B.** 3D rendering of the same network. View was also after zooming in slightly to show the 3D aspect of nodes as well as rotation of the network.

As with the 2D Network View, you can change the node and link settings in the 3D view using the Toggle Network Settings icon on the top left corner of the screen. In addition to this icon, and the Export Network icon , there is also a third icon , which lets you refresh the network after you make changes. Figure 45 below shows the open Toggle Network Settings menu, along with the Global Settings menu used to map node color to risk factor. NOTE: Any color options set in the 2D network will be retained in the 3D network and vice-versa.

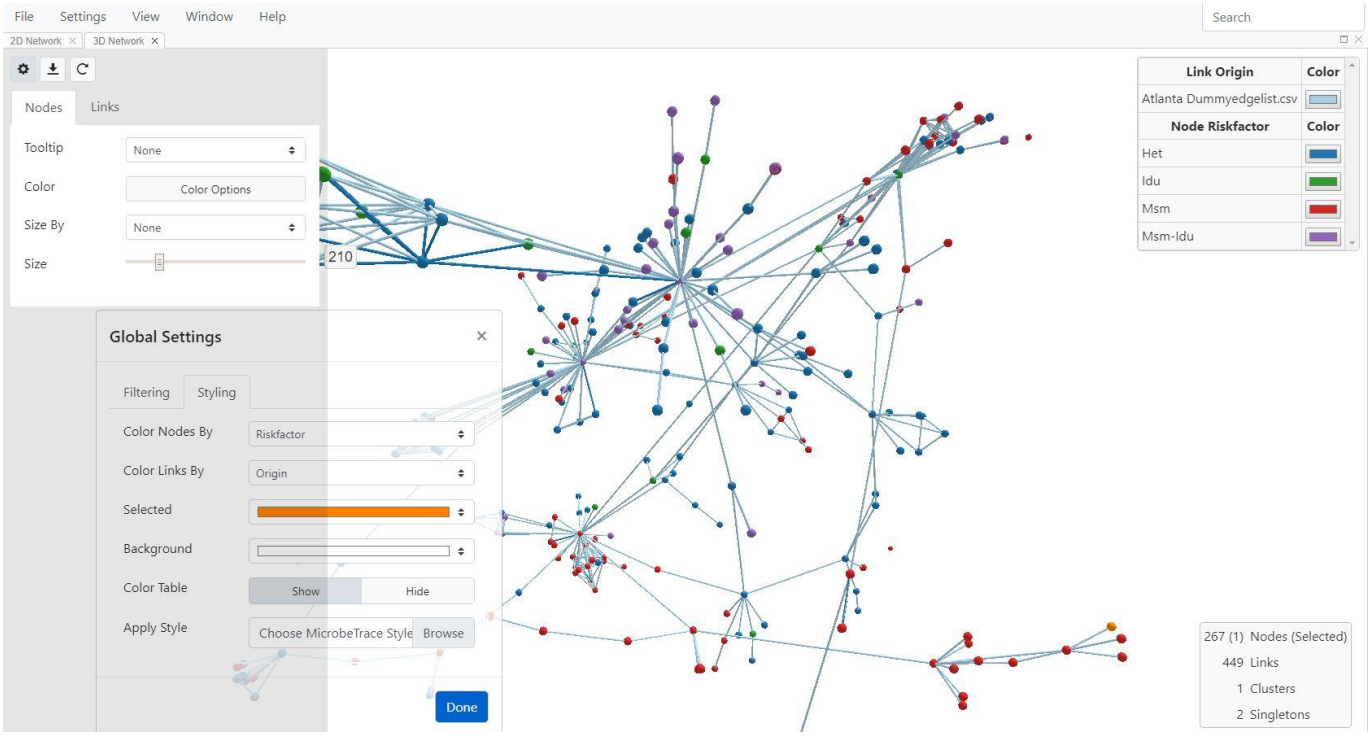


Fig. 45. 3D network settings with nodes colored by risk factor.

Histogram View

If the analysis included a FASTA file with nucleotide sequences, then the genetic distance results calculated with the TN93 nucleotide substitution model can be viewed as a histogram. Although viewing the distribution of genetic distances in the sequences is the main use of the histogram view, it can also be used with contact tracing data as an alternate visualization of the relationship between variables.

The genetic distance histogram is a bar chart that shows the frequency with which a particular genetic distance occurs in the data set. Typically, the frequency distribution in the histogram chart appears bi-modal (two peaks). One peak will contain genetic distances of very closely related sequences and the second will contain more distantly related sequences. The genetic distance which best separates these two peaks can be used to refine the genetic distance threshold selection for your specific analysis. A miniature version of this histogram is also available on the **Global Settings Menu** (accessed by selecting the Settings on your main MicrobeTrace window) for your

convenience. You can use this feature to more easily determine genetic distance cut-off values for your data.

Click **Histogram** under **View** to display the genetic distance histogram (Fig. 46).

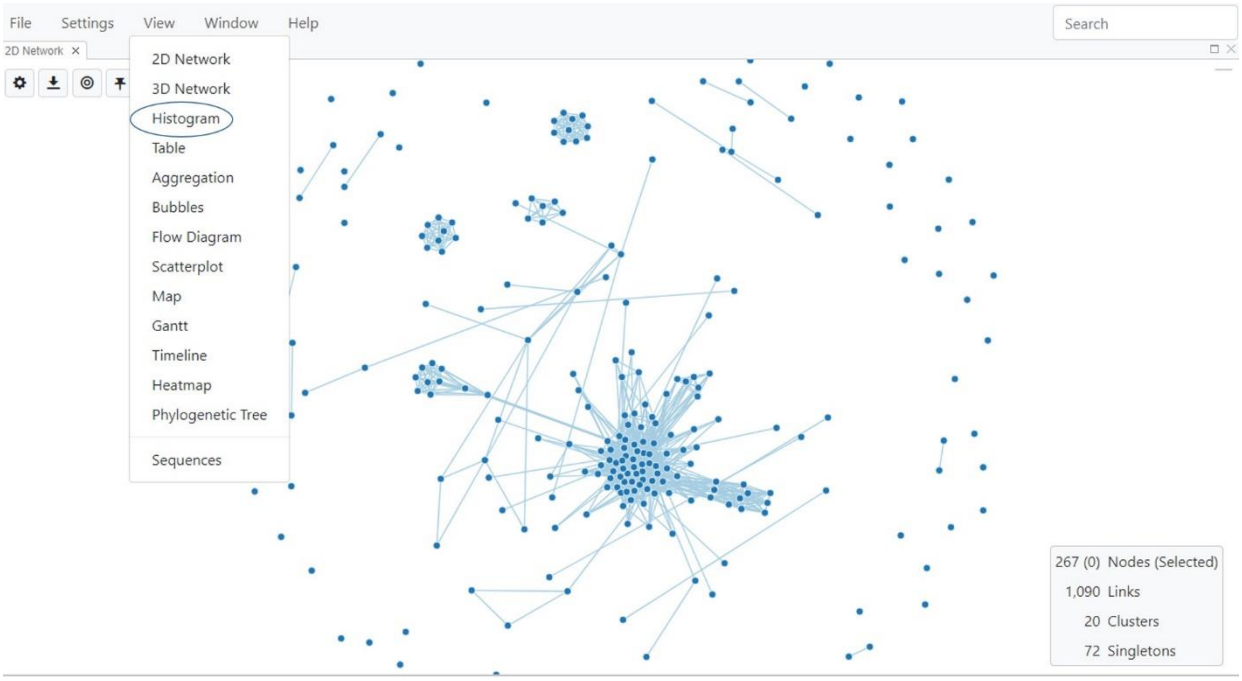


Fig. 46. Selecting Histogram View

The Histogram settings can be changed by selecting the Toggle Histogram Settings button  . Select the relevant variable from the pull-down menu to choose the type of links used for configuring the histogram (Fig. 47). The default histogram setting uses genetic distance as the variable, displays the frequency of the distances between all links and uses a linear scale for the distances. The genetic distances can also be plotted in linear or log scales by toggling between linear and log scales when distance is the chosen variable to plot. If you have chosen to color nodes or links by any variable in the 2D view, those same colors will be displayed in the Histogram View.

The distances on the bimodal curve can be examined to determine the genetic distance cut-off used for determining linkage of the nucleotide sequences (see genetic distance threshold in the glossary for details). If a different genetic distance cut-off value needs to be used for your specific analysis, then go to the Network View and input this value in the genetic distance threshold under Network Configuration.

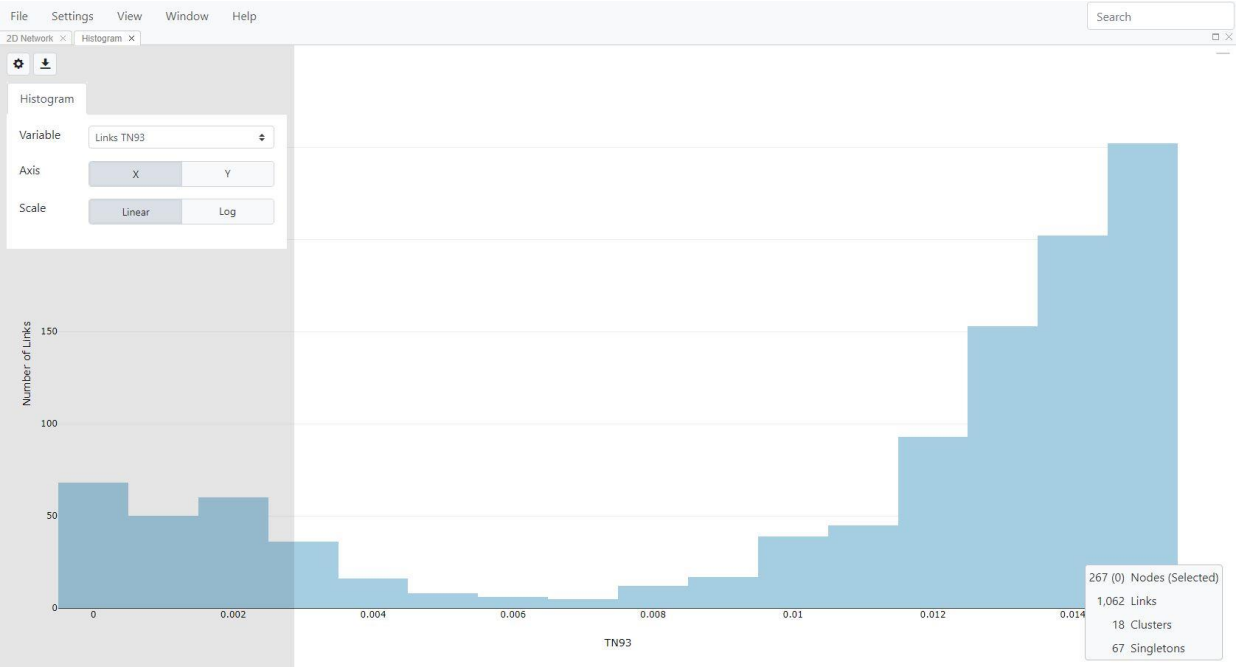


Fig. 47. Histogram view with the number of links plotted against genetic distance calculated using the TN93 nucleotide substitution model.

Table View

Table View shows the data associated with the nodes in the form of a table like an Excel worksheet. Select **Table** under **View** (Fig. 48).

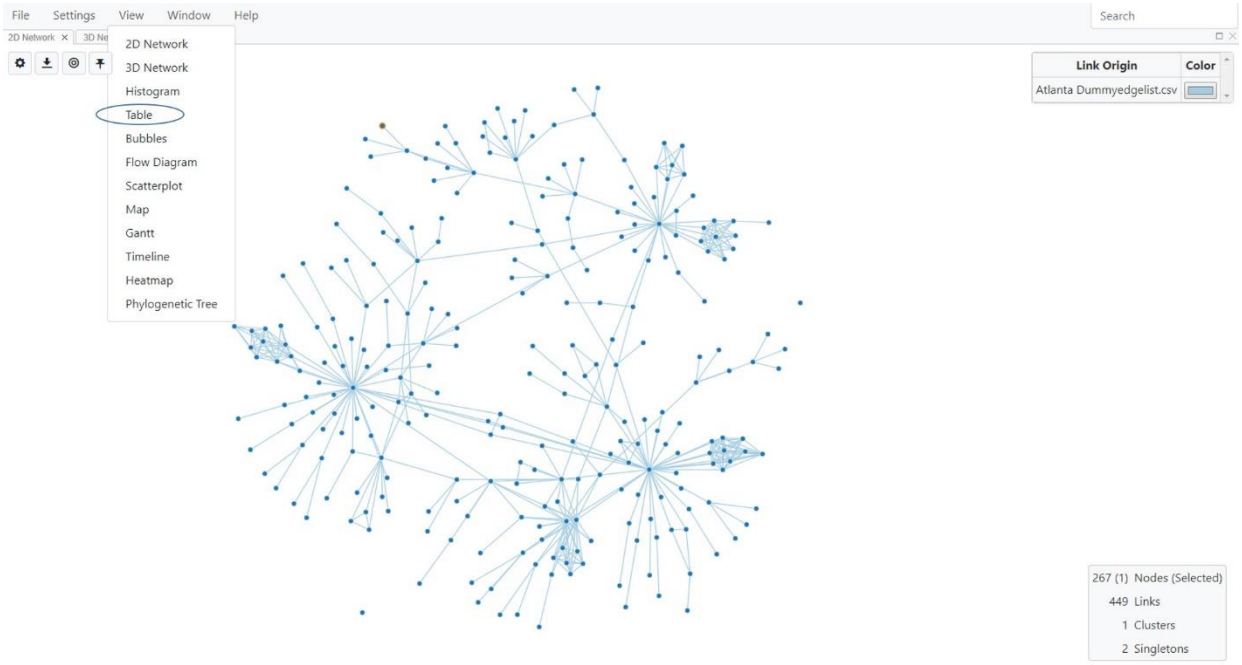
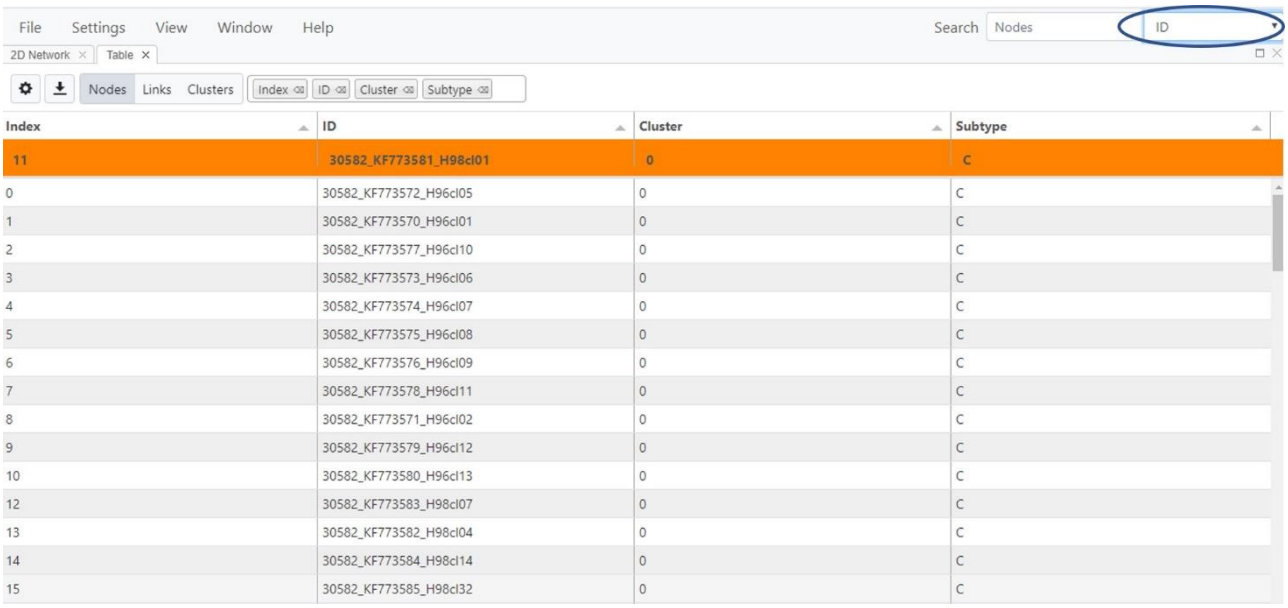


Fig. 48. Selecting the Table View

The system displays the table in a new window (Fig. 49).



Index	ID	Cluster	Subtype
11	30582_KF773581_H98cl01	0	C
0	30582_KF773572_H96cl05	0	C
1	30582_KF773570_H96cl01	0	C
2	30582_KF773577_H96cl10	0	C
3	30582_KF773573_H96cl06	0	C
4	30582_KF773574_H96cl07	0	C
5	30582_KF773575_H96cl08	0	C
6	30582_KF773576_H96cl09	0	C
7	30582_KF773578_H96cl11	0	C
8	30582_KF773571_H96cl02	0	C
9	30582_KF773579_H96cl12	0	C
10	30582_KF773580_H96cl13	0	C
12	30582_KF773583_H98cl07	0	C
13	30582_KF773582_H98cl04	0	C
14	30582_KF773584_H98cl14	0	C
15	30582_KF773585_H98cl32	0	C

Fig. 49. Table View display of data in the node list file.

A search box allows you to find information for nodes of interest in the table data. You can search nodes by any of the attributes in your csv files, available as a dropdown menu to the right of the search box (see circled in Fig. 49)

If you choose to highlight data for specific nodes in the table format, then you can select as many nodes as you want to include in the table view by holding down the **Ctrl** key in the 2D view to select the nodes. Now go back to the Table View window to see the data for the nodes that were selected. MicrobeTrace brings the selected rows to the top of the table for convenience. You can also click and drag the Table View tab alongside the Network View so that the two windows are side by side in the same browser window (Fig. 50).

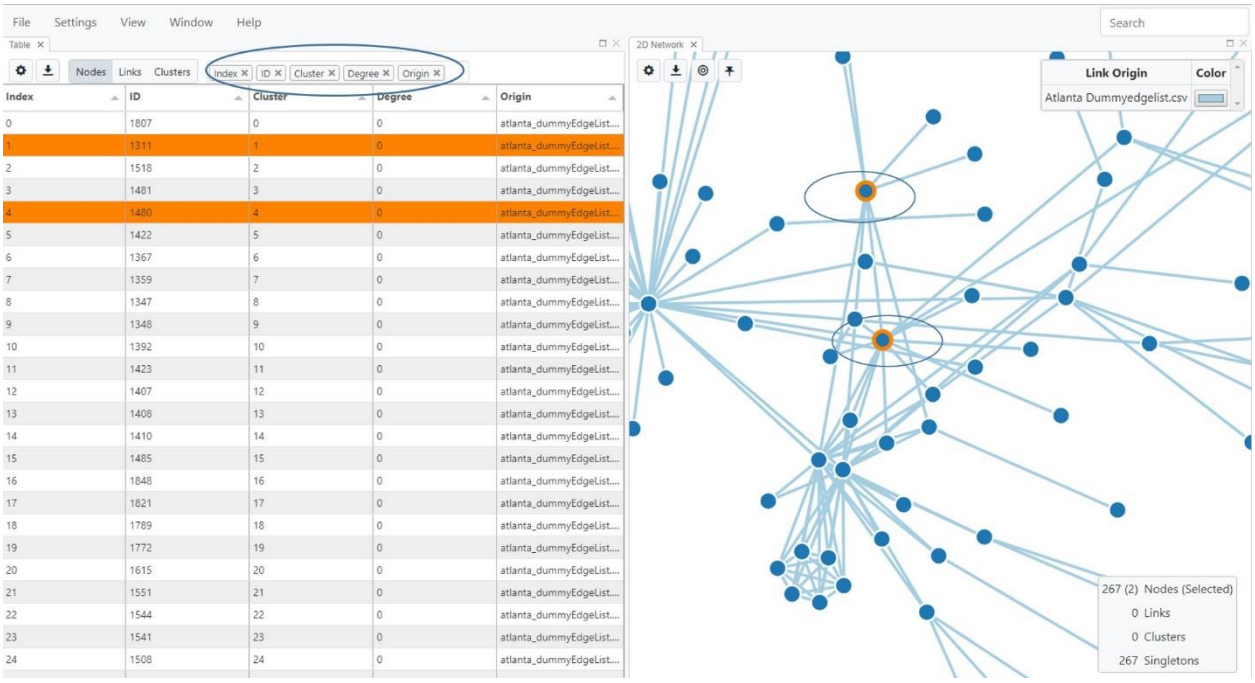


Fig. 50. Table View (right) alongside 2D Network View (left). Selected nodes are highlighted in orange in the network, as are the corresponding rows of data in the Table View. The selected nodes are brought to the top of the table.

By default, only four columns are displayed in Table View. If you would like to display more columns, click on the "white space" (circled in Fig. 50) to select more columns from the dropdown menu. Alternatively, you can start typing the column names in the "white space" and MicrobeTrace will automatically try to guess your selection and complete the word like spell checking does. You can also remove columns from the Table View by selecting the "x" in the column name box.

Clicking on the column headers in the table will sort all rows according to data in that column. Selecting rows in the table will highlight the corresponding nodes in the network and also the

linked geographic data in the Map View (but only if the geographic data is included in your node list data file). Searches can be performed on all textual entries in the table. You can search for a specific node using an ID or other identifiers, select it in the table, and then see the corresponding highlighted node in the Network View (node is highlighted in orange) to see links and cluster or network positions. Please note that if you chose not to display singletons (using the Global Settings Filtering tab), and you select a node while you are in the Table View that happens to be a singleton, the singleton will not be displayed in the Network View.

The default Table View setting displays node data in the table columns. You can use the buttons above the table, to the right of the settings buttons to switch between viewing the link or cluster data in the table. The same features are available for Link and Cluster Views, including selecting IDs, sorting, selecting which columns to display, etc.

Table Settings

You can adjust text sizes in the Table View by using the provided sliding scale options on the Toggle Table Settings button  .

Aggregation View

Aggregation View is used to total variables in your node data. Aggregation View provides a table with two columns: the values for the variable you're totaling, and the total number of entities with those values. This function is especially useful to generate tables or reports. To use this feature, select **Aggregation** from the View menu (Fig. 51).

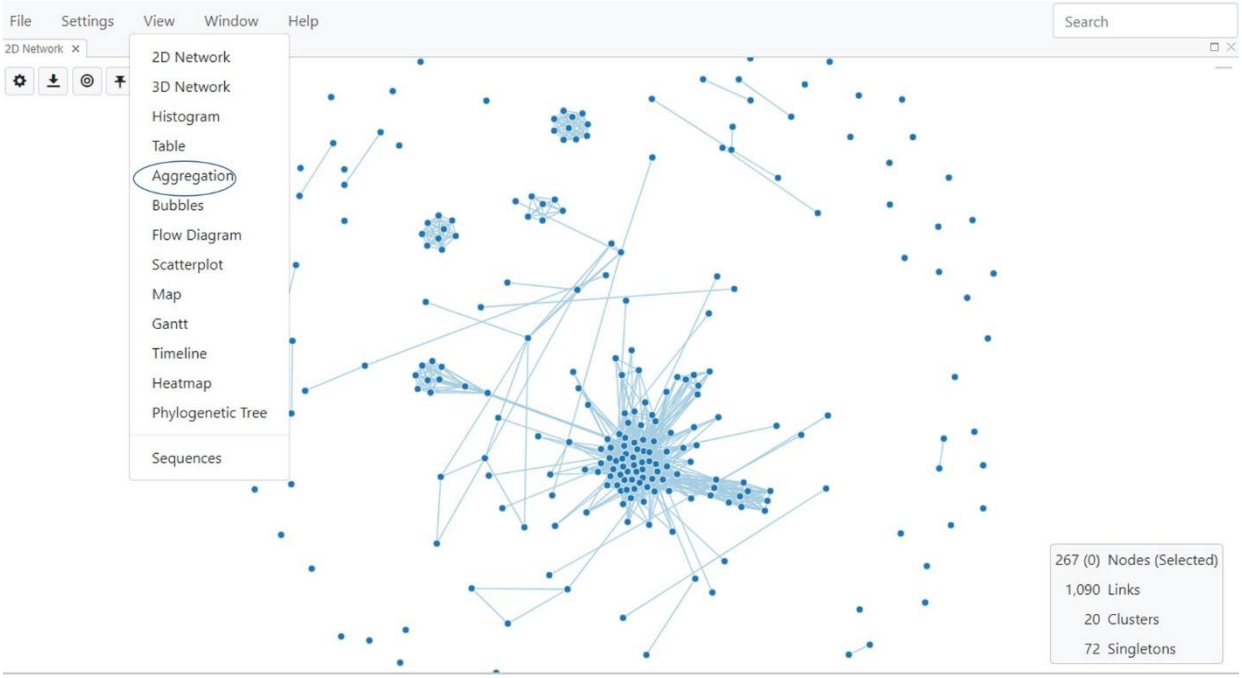


Fig. 51. Selecting Aggregation View

In the example below, when you first load the aggregation table, the default setting for variable is **Cluster**. The number of nodes in each cluster is provided in the second column, and the third column lists the percentage of nodes in each cluster. Select the Toggle Aggregation Settings  button which will bring up the settings menu to customize the variable to which you would like to compare node, link or cluster data.

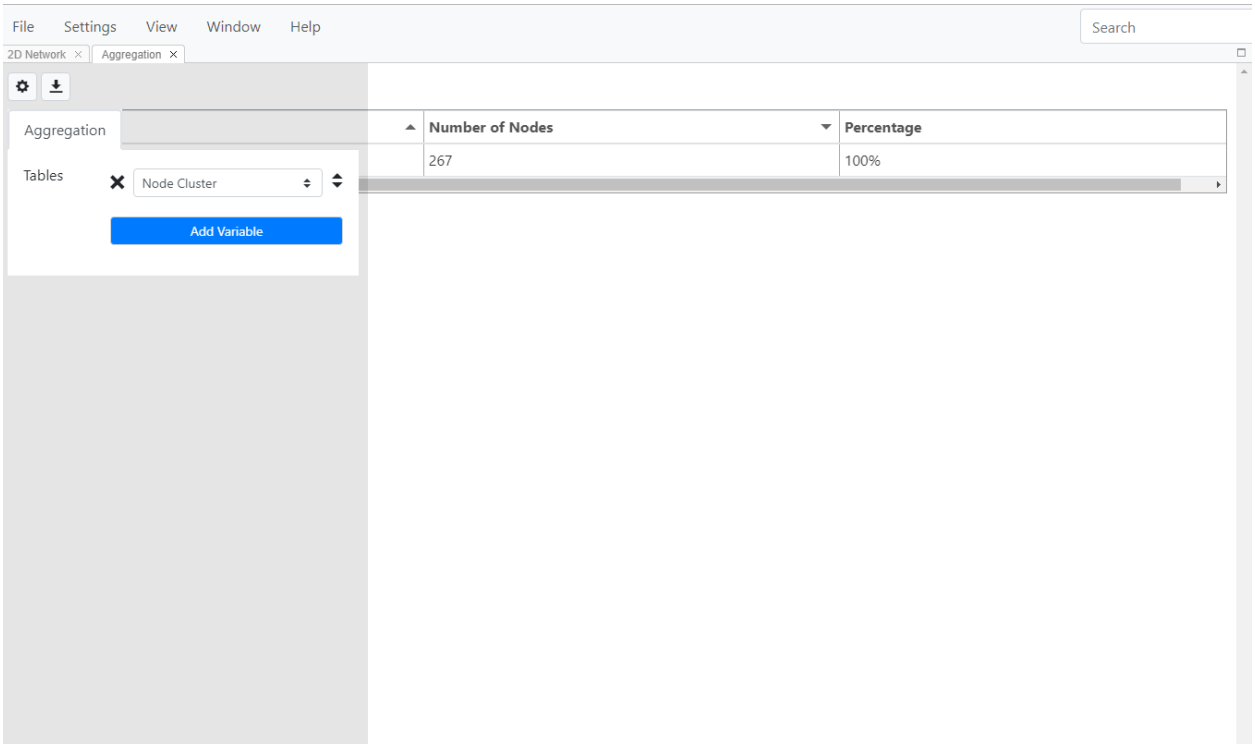


Fig. 52. Aggregation View showing cluster as the variable with the number and percentage of nodes in each cluster.

You can add as many variables as you wish using the **Add Variable** button. In the example below (Fig. 53), we have chosen three variables, risk factor, county and gender. The second and third column list the number and percentage of nodes for each.

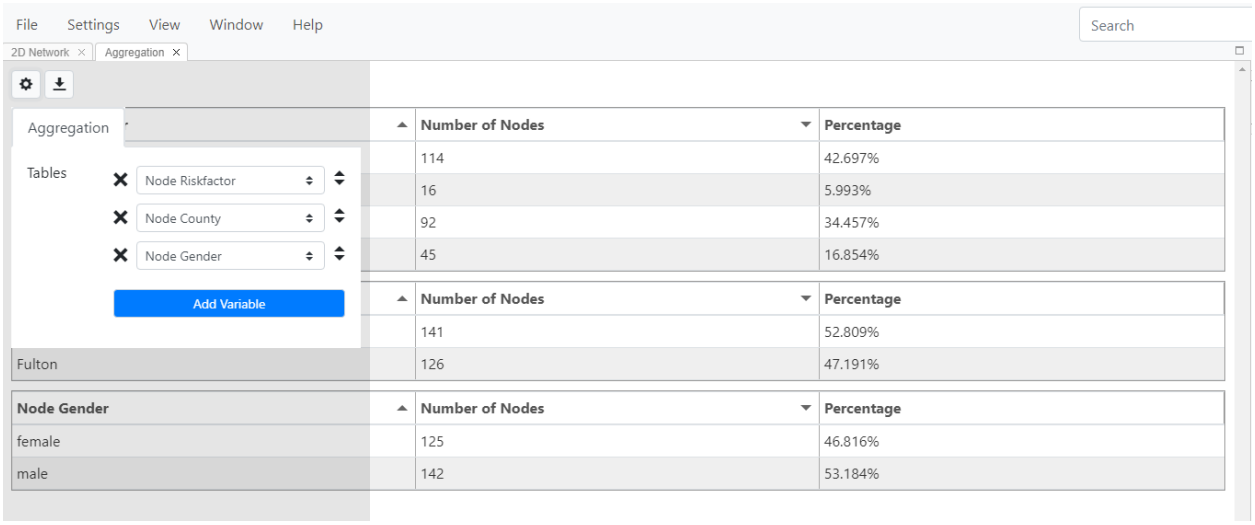
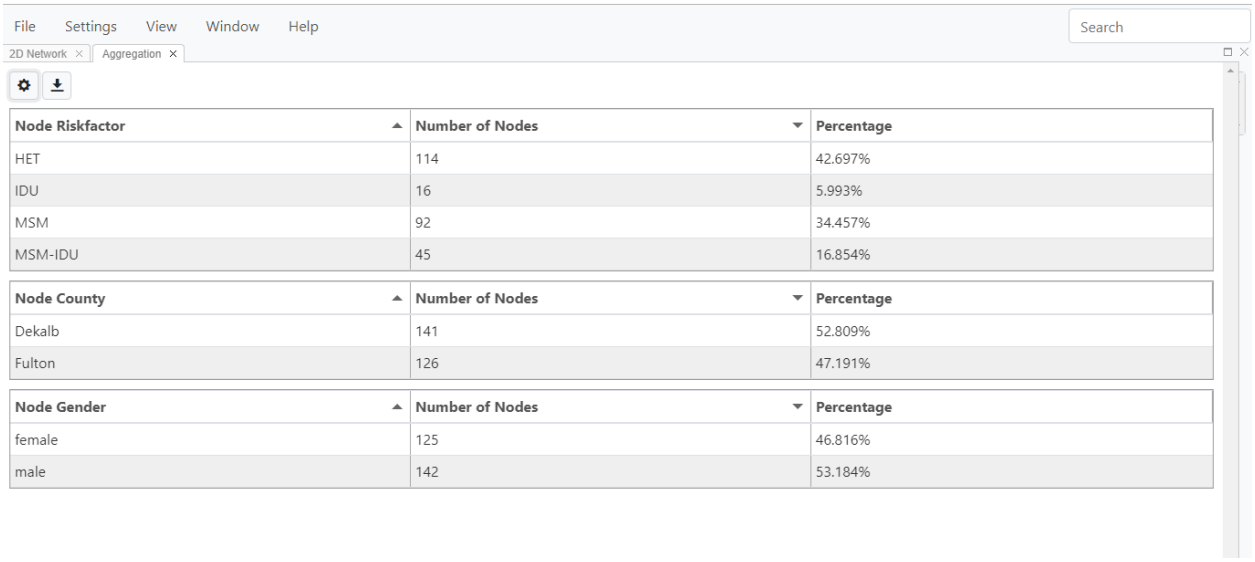


Fig. 53. Selecting multiple variables to create tables in Aggregation View



Node Riskfactor	Number of Nodes	Percentage
HET	114	42.697%
IDU	16	5.993%
MSM	92	34.457%
MSM-IDU	45	16.854%

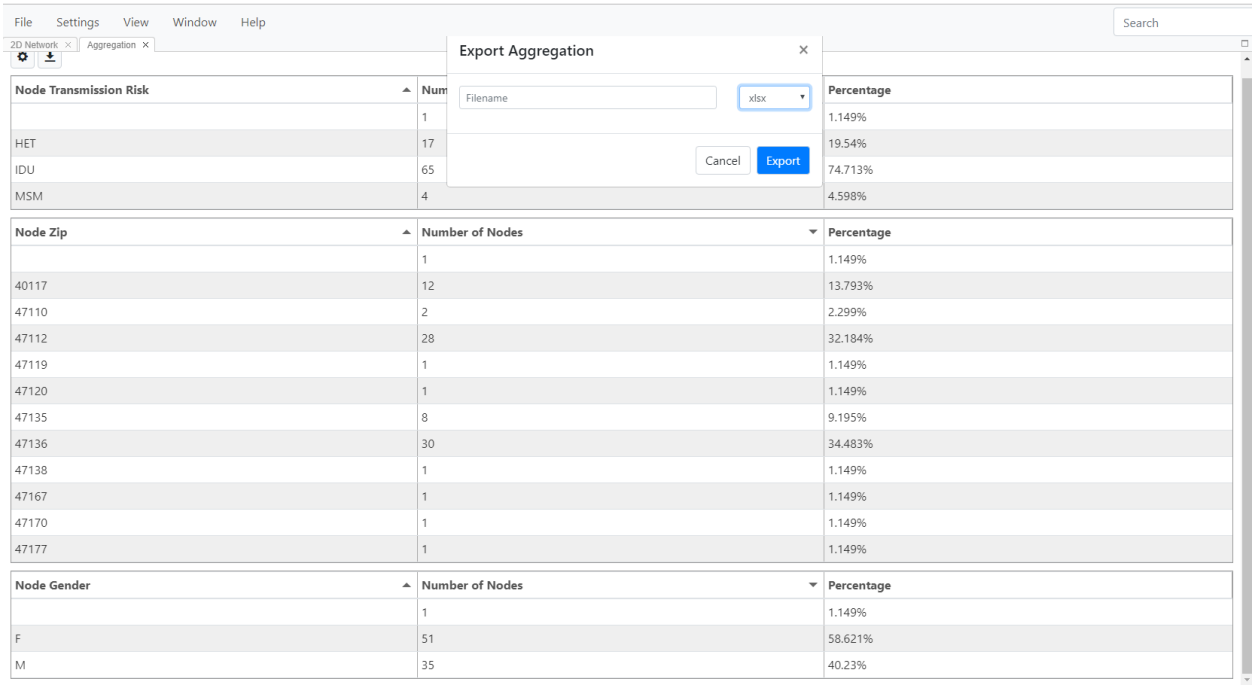
Node County	Number of Nodes	Percentage
Dekalb	141	52.809%
Fulton	126	47.191%

Node Gender	Number of Nodes	Percentage
female	125	46.816%
male	142	53.184%

Fig. 54. Aggregate tables for three variables showing number and percentage of nodes in each category.

This view can then be exported in a format of your choice. Select the download button  to choose the file format from the dropdown menu.

In the figure below (Fig. 55), the .xlsx format is selected. This will export your table as an Excel file with each table in a separate worksheet.



Node Transmission Risk	Num	Percentage
	1	1.149%
HET	17	19.54%
IDU	65	74.713%
MSM	4	4.598%

Node Zip	Number of Nodes	Percentage
	1	1.149%
40117	12	13.793%
47110	2	2.299%
47112	28	32.184%
47119	1	1.149%
47120	1	1.149%
47135	8	9.195%
47136	30	34.483%
47138	1	1.149%
47167	1	1.149%
47170	1	1.149%
47177	1	1.149%

Node Gender	Number of Nodes	Percentage
	1	1.149%
F	51	58.621%
M	35	40.23%

Fig. 55. Exporting aggregate tables

CrossTab View

This view allows bivariate cross-tabulations of your data which can be used to evaluate relationships between categorical variables that may not be readily apparent. It is often used to evaluate multiple response answers to survey questions, typically collected during partner services interviews. Top open, select CrossTab from the dropdown View menu.

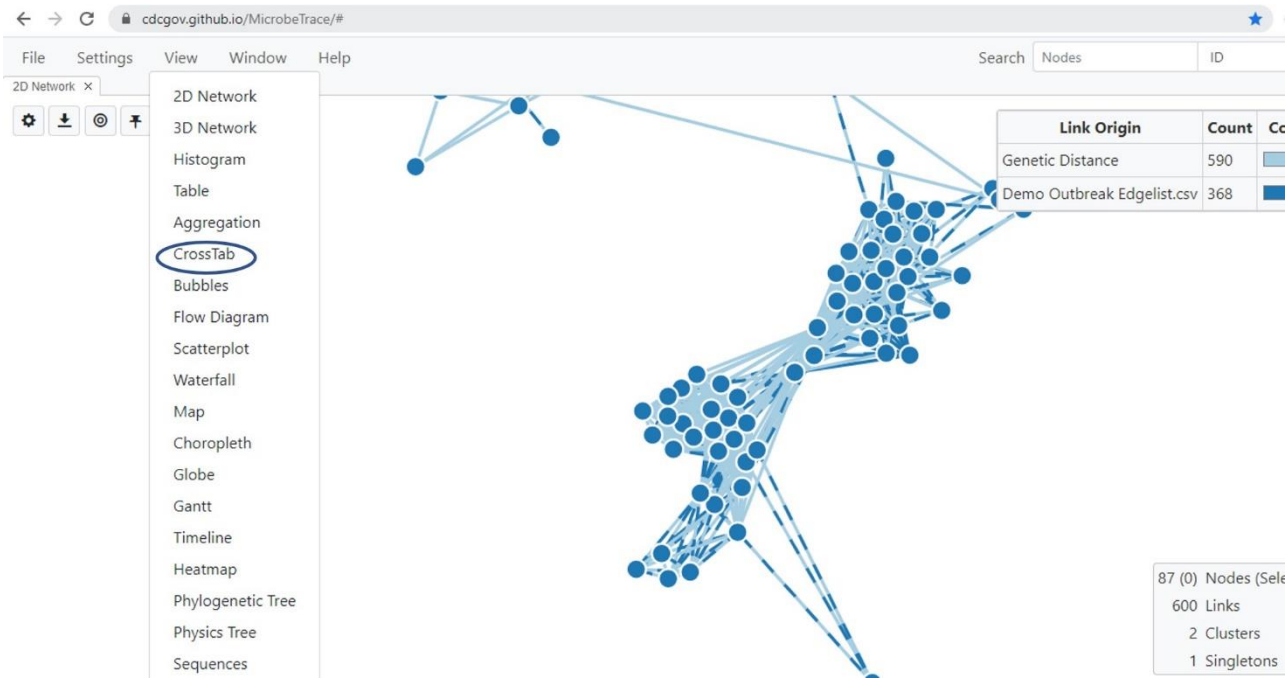
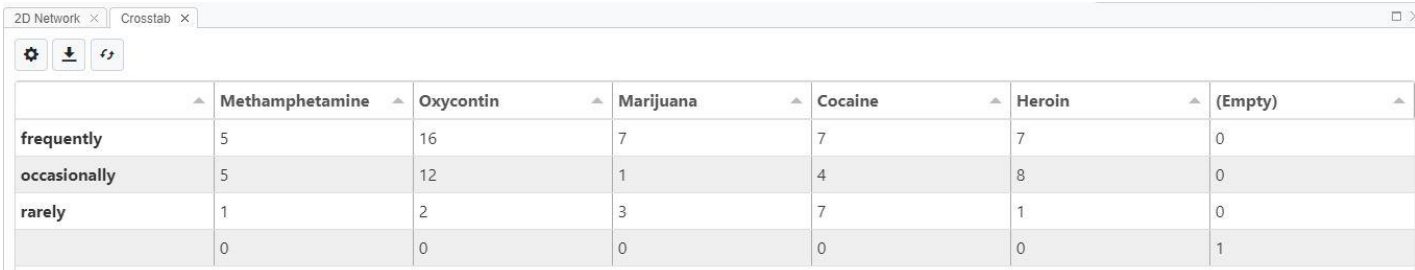


Fig. 56. Selecting CrossTab View

Then, use the settings button to select which entity you would like to cross-tabulate over (nodes, links, or clusters). Next, select two categorical attributes that you would like to evaluate from the X and Y variable dropdown menus (Fig. 57). The resulting counts from the cross-section of the two selected variables will be displayed as a table. This table can be exported as a CSV, XLSX, PDF, or JSON file.



	Methamphetamine	Oxycontin	Marijuana	Cocaine	Heroin	(Empty)
frequently	5	16	7	7	7	0
occasionally	5	12	1	4	8	0
rarely	1	2	3	7	1	0
	0	0	0	0	0	1

Fig. 57. Cross tabulation of primary drug across frequency of use

Bubbles View

The Bubbles View is used for visualizing demographic distributions within the network when you don't want to display the links between nodes in the network. For example, showing linkages between nodes or persons can be misinterpreted with directionality of transmission when that information is unknown or not available (Oster et al., Barré-Sinoussi et al.). Removing the displayed links can then help to eliminate any unnecessary confusion when sharing the network data as an image.

Bubbles View is a force-directed diagram of all the nodes in a MicrobeTrace session, without the links between nodes. Unlike the Network View, you cannot move individual nodes around with the mouse. However, you can move them around the screen to cluster with similar nodes according to variables in your dataset. To select this view, select Bubble from the dropdown menu (Fig. 58).

Fig. 58. Selecting Bubble View

This opens a default view of just the nodes in the data (Fig. 59).

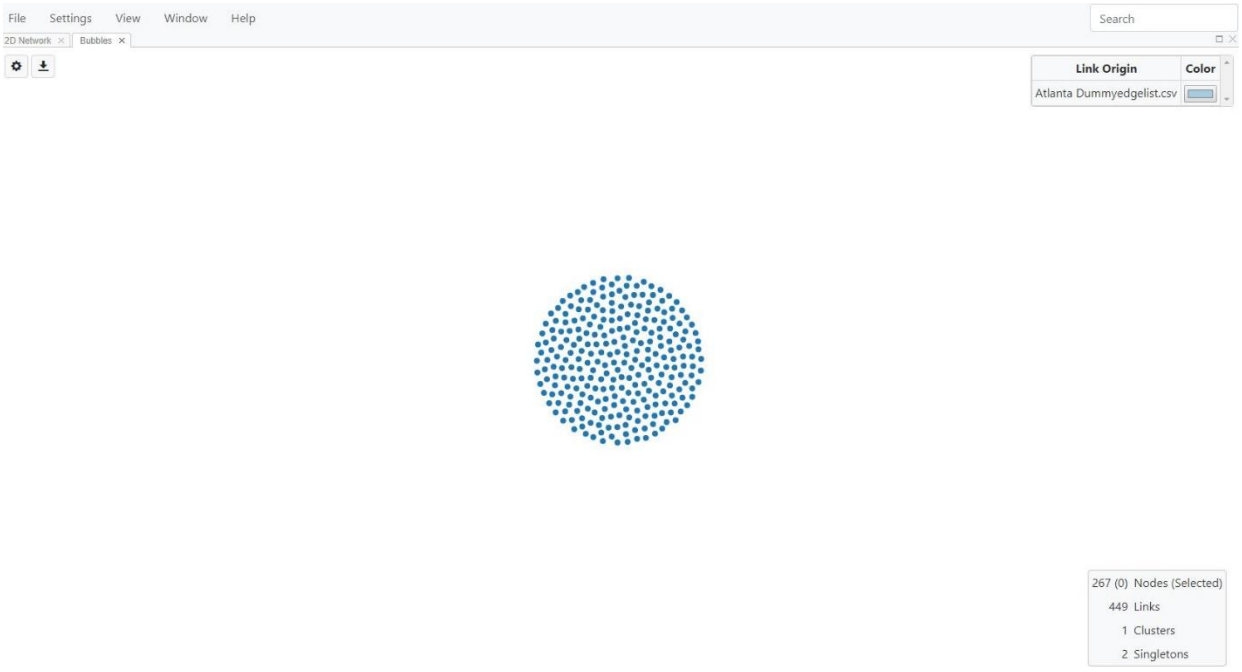


Fig. 59. Default Bubble View

Bubble View Settings

In order explore clustering by a variable in the data, use the **Toggle Bubble Settings** button to open the settings menu This brings up two tabs- **Nodes** and **Physics** (Fig. 60).

Nodes tab lets you change the size of nodes using a slider bar. As in other views, you can use the **Color Options** button to map the node color to a variable of your choice. If you had already mapped node color in the 2D network, then that color mapping will transfer over to other views, including the Bubble View.

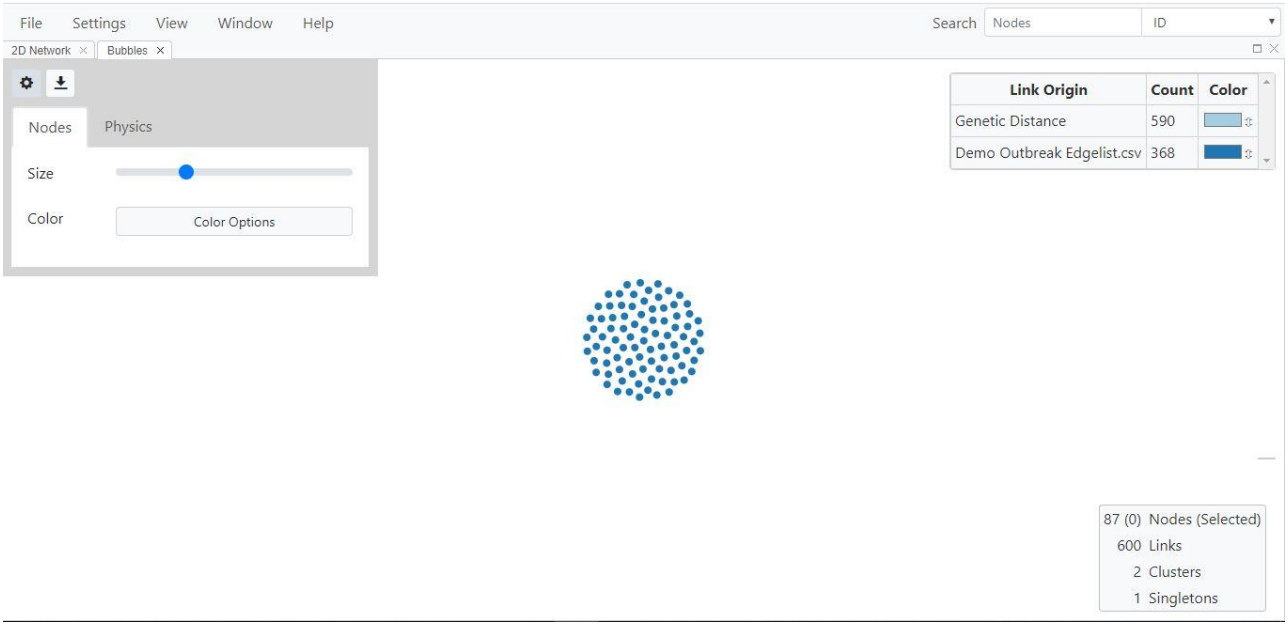


Fig. 60. Toggle settings for Bubble view

The **Physics** tab (Fig. 61) allows you to map nodes to variables on the X-axis and Y-axis. In the example below, we are looking at transmission risk factor versus gender.

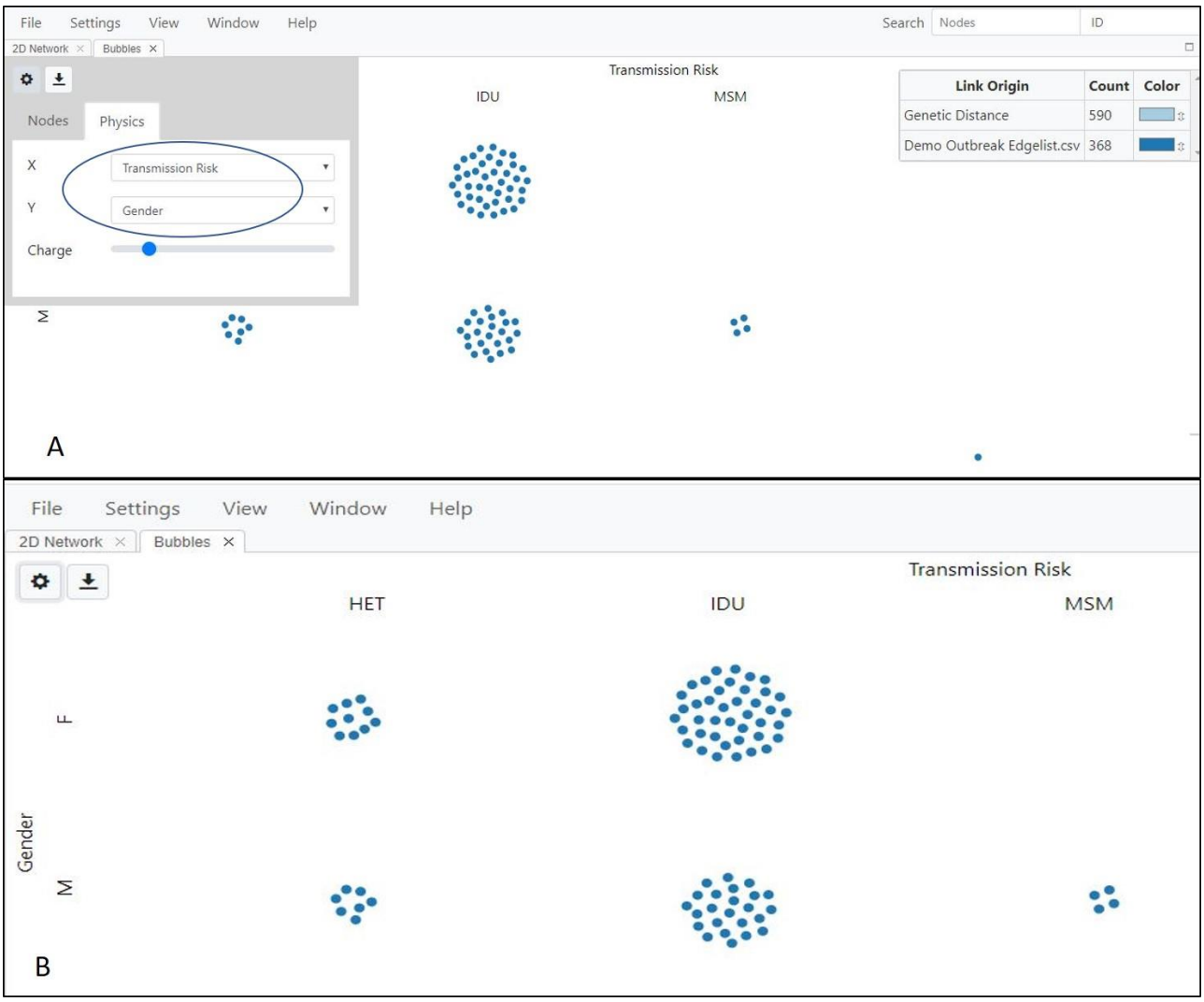


Fig. 61. Bubble view showing risk factor versus county of the nodes in the transmission cluster. A. Choosing variables from the dropdown menu (circled); B. Risk factor versus county. HET-heterosexuals, IDU-intravenous drug user, MSM-Men who have sex with men.

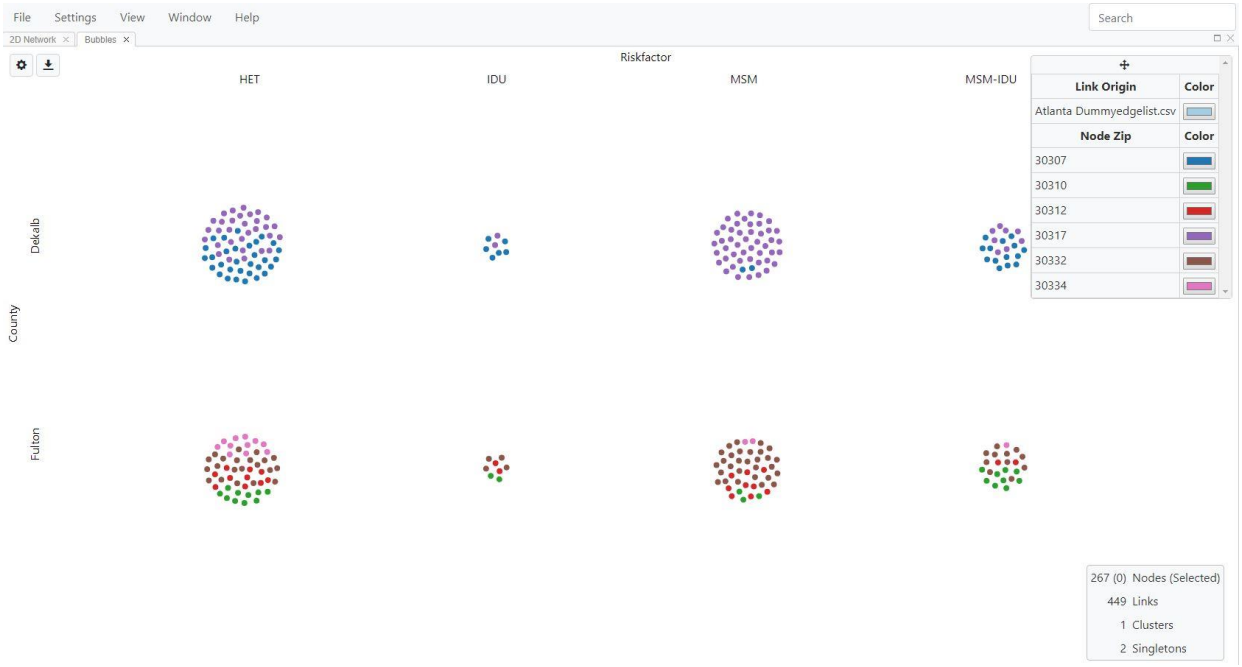


Fig. 61. Bubble view showing risk factor versus county (Dekalb and Fulton in Atlanta, GA) with nodes colored by zip code.

Flow Diagram View

The Flow Diagram View visualizes the data in a form of a flow diagram (specifically, an alluvial or Sankey diagram) and allows for a comparison of variables in the data set. Flow diagrams can be generated from any data variable, as they rely exclusively on node characteristics. Select **Flow Diagram** from the **View** menu (Fig. 62).

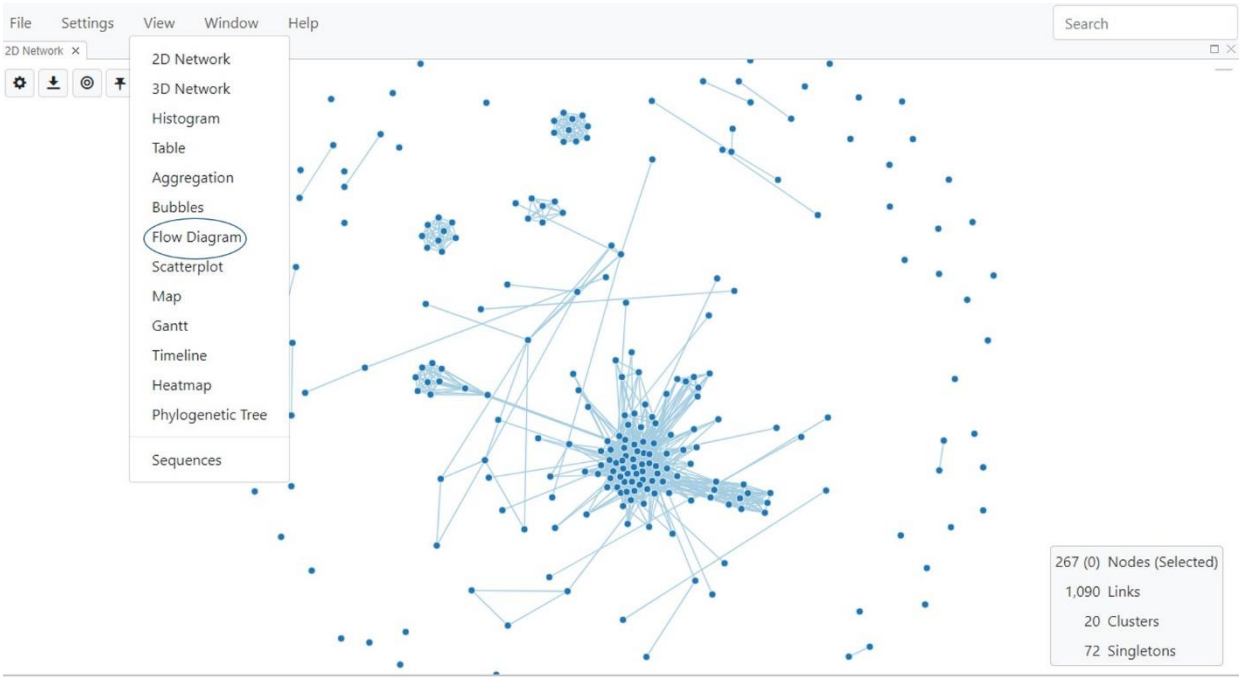


Fig. 62. Selecting Flow Diagram View

Flow Diagram View Settings

Once selected, MicrobeTrace displays the Flow Diagram View in a new window (Fig. 63).

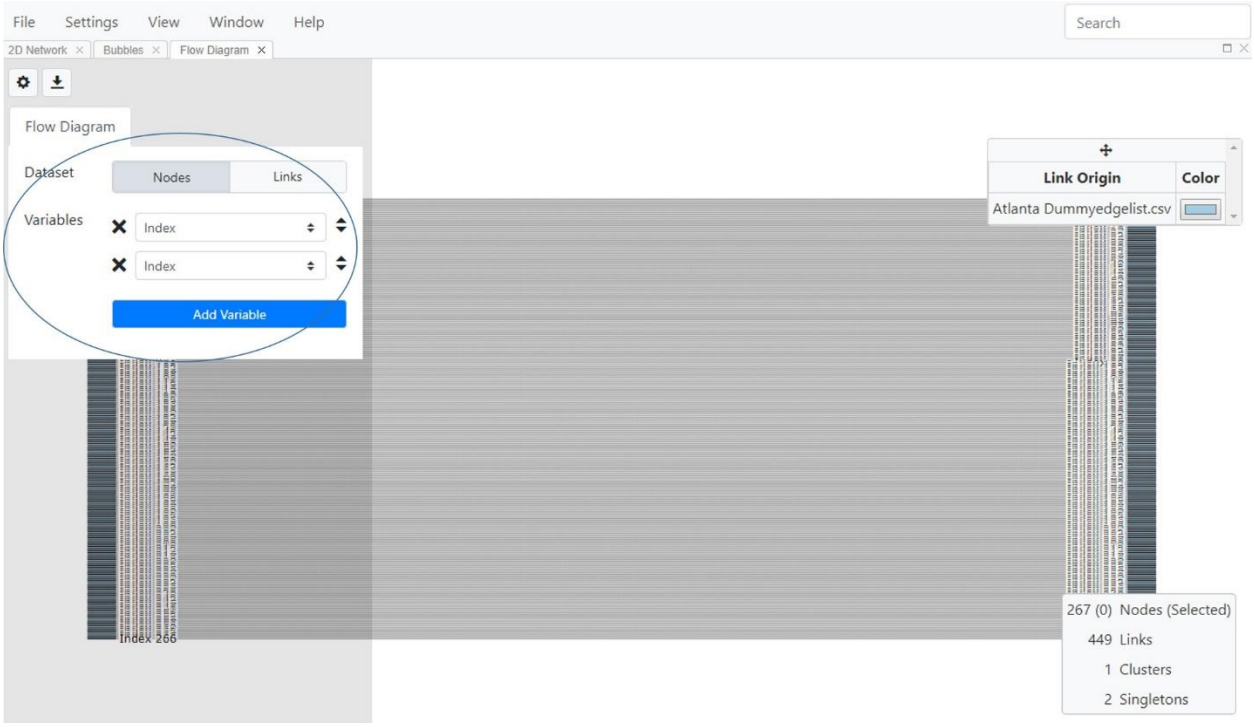


Fig. 63. Flow Diagram View before mapping to specific variables in the dataset.

The Flow Diagram View default is set to index, so the flow diagram will not be displayed until variables are selected from the Flow Diagram dialog box that is opened using the **Toggle Flow Diagram** button  At least two variables are required to explore your data using the Flow Diagram View. Use the dropdown menu to select the two variables you want to visualize in the flow diagram. You can also add more variables using the Add Variable button. In this example, infection risk factors, county and zip codes were examined. Fig. 64 shows mapping of risk factor to county and zip code.

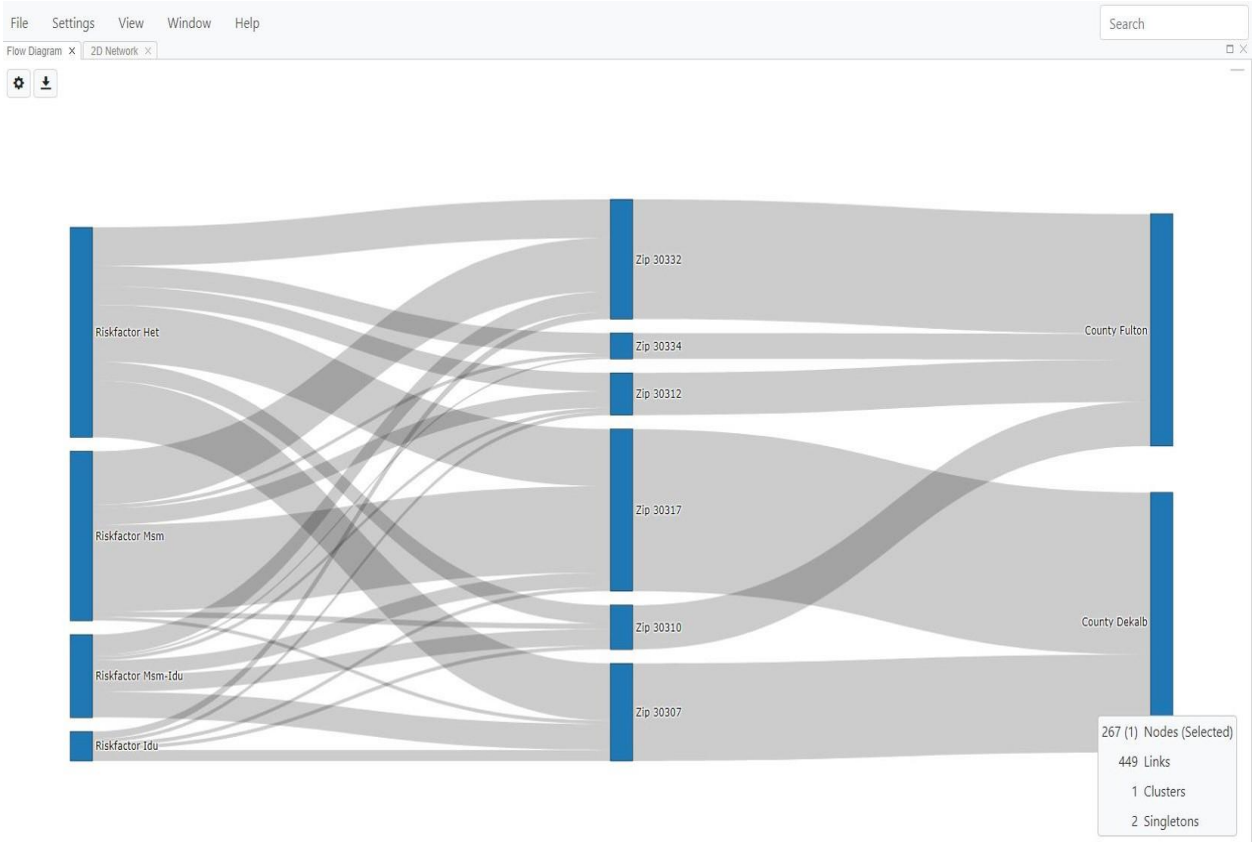


Fig. 64. Flow Diagram View showing a comparison of risk factor across zip codes and counties (Fulton and DeKalb in Atlanta, GA).

The blue blocks in the flow diagram in Fig. 64 represent the relative prevalence of the selected variables and the grey stream fields (curved lines of variable thickness) between the blocks represent associations between the selected variables.

Scatterplot View

Scatterplot View provides visualization of correlations of numeric variables in your data. Select **Scatterplot** from the **View** menu (Fig. 65).

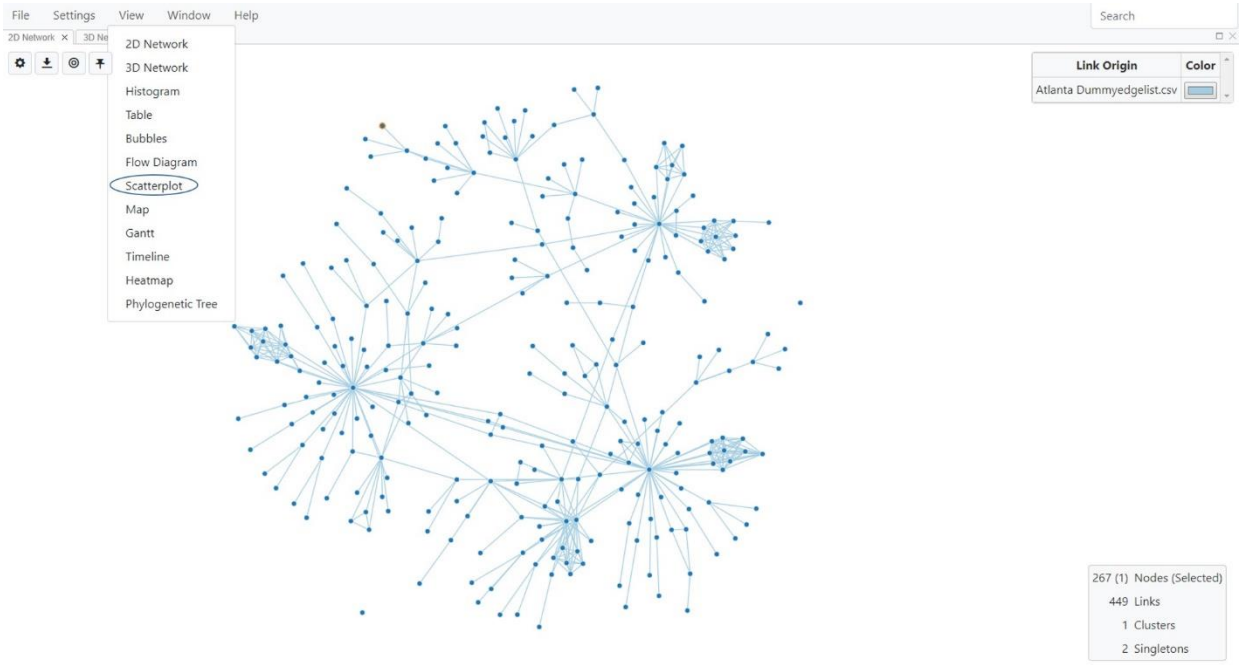


Fig. 65. Selecting Scatterplot View

MicrobeTrace displays the scatterplot in a new tab (Fig. 64). By default, it plots TN93 distance against SNPs. However, you can use the Toggle Scatterplot button  , which opens a dialog box to choose the variables to compare on the X and Y axes using the pull-down menus (Fig. 66). Select Nodes or Links in the **Dataset** option to choose which variables to compare. You can also adjust the numeric scale to either log or linear. Selecting **Color Options** takes you to the Global Settings menu where you can customize colors in this and other views.

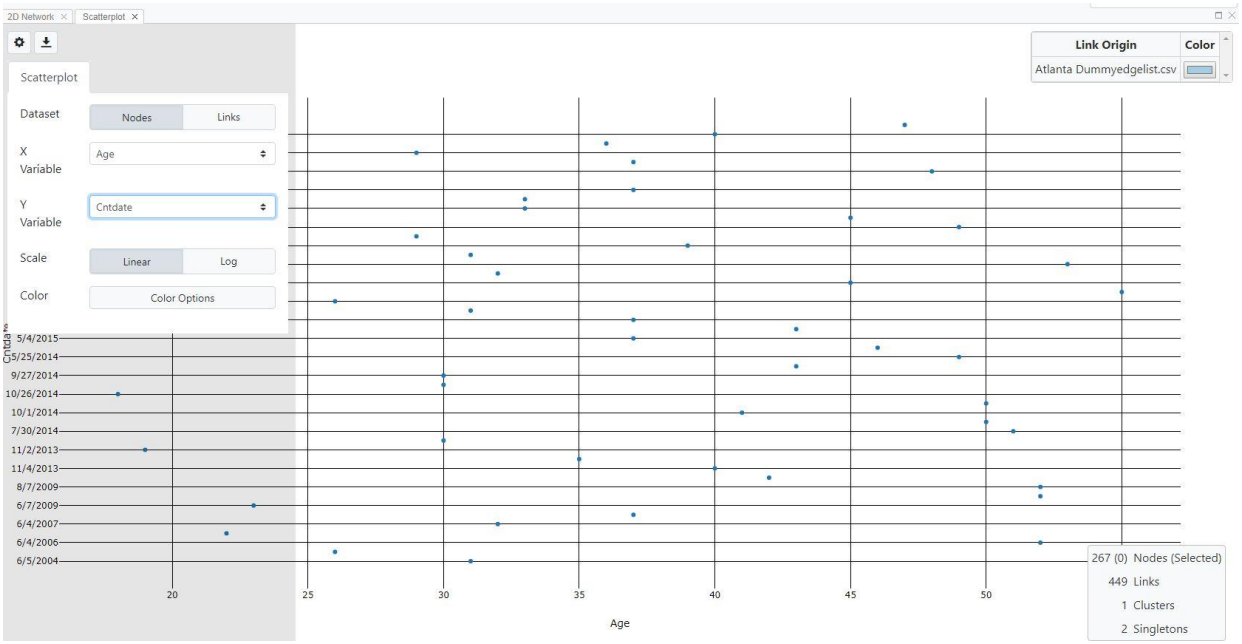


Fig. 66. Scatterplot View and adjusting display options

Waterfall View

This view displays cluster, node and link data in a tabular format rather than a network format. Selecting a cluster from the list shows node and link information for that cluster, and then you can in turn select nodes from the list to view link information. This view gives detailed information about the relationships between these three variables in a textual format rather than a visual, network-based view. To select this view, select Waterfall View from the dropdown View menu (Fig. 67).

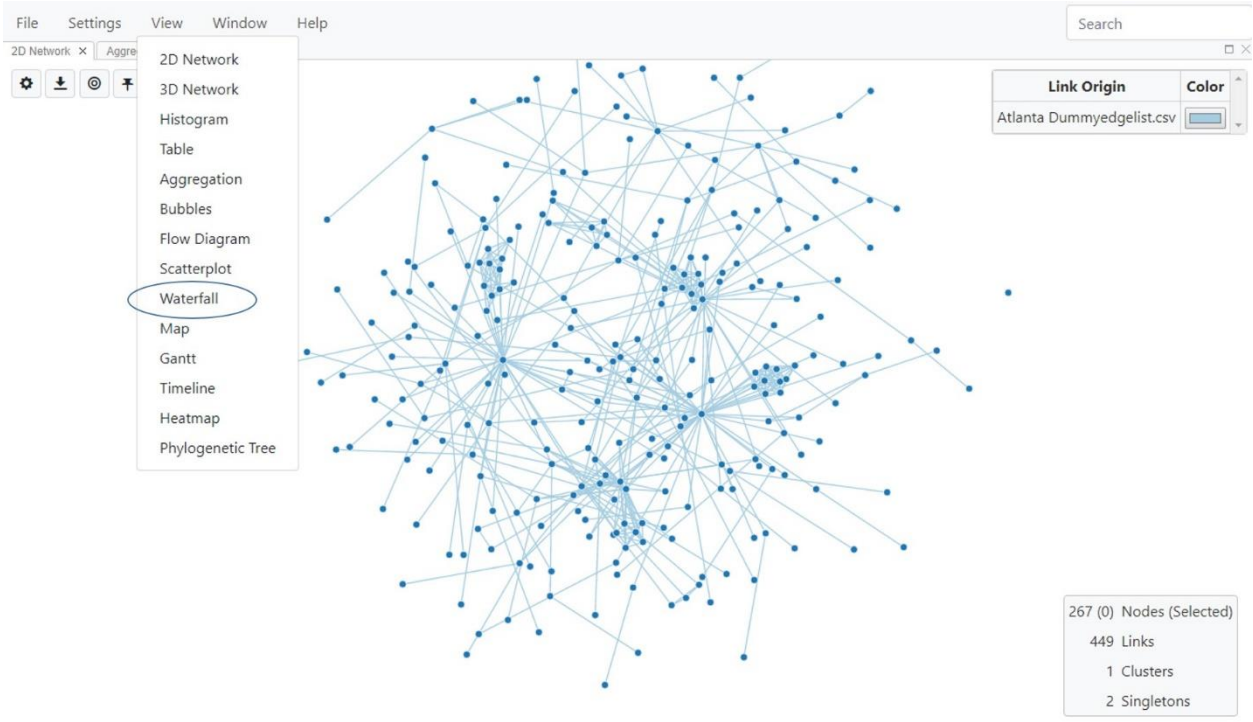


Fig. 67. Selecting Waterfall view

The waterfall view displays three columns, one each for cluster, node and link (Fig. 68).

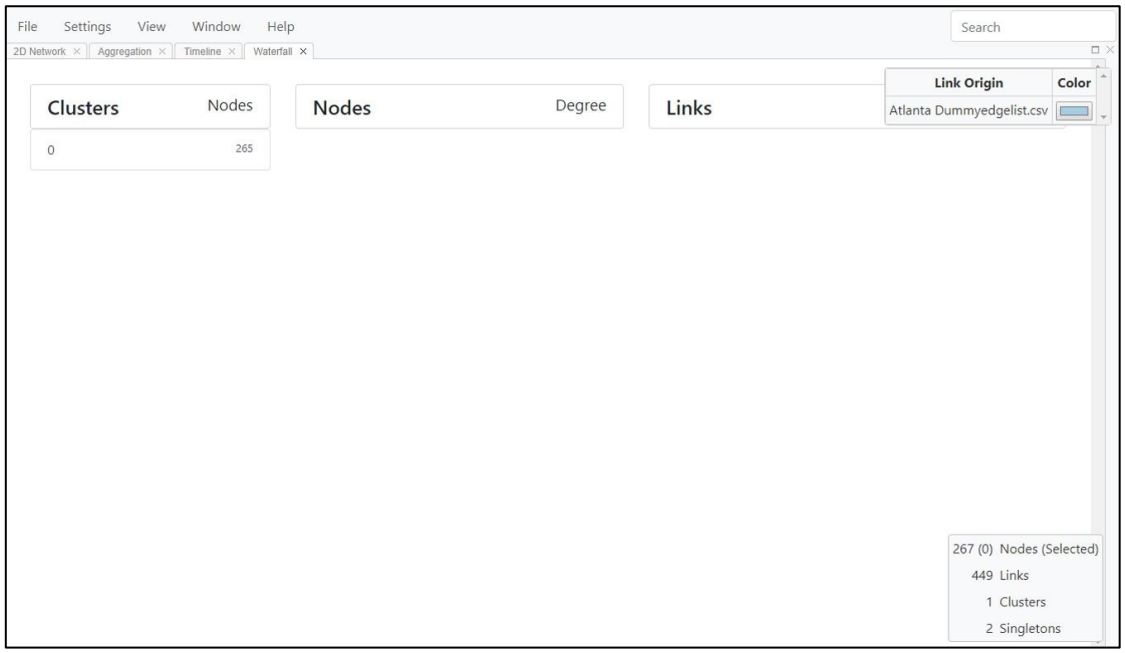


Fig. 68. Waterfall view, opening screen

Clicking on a specific cluster then lists the nodes in that cluster, along with the node degree (Fig. 69).

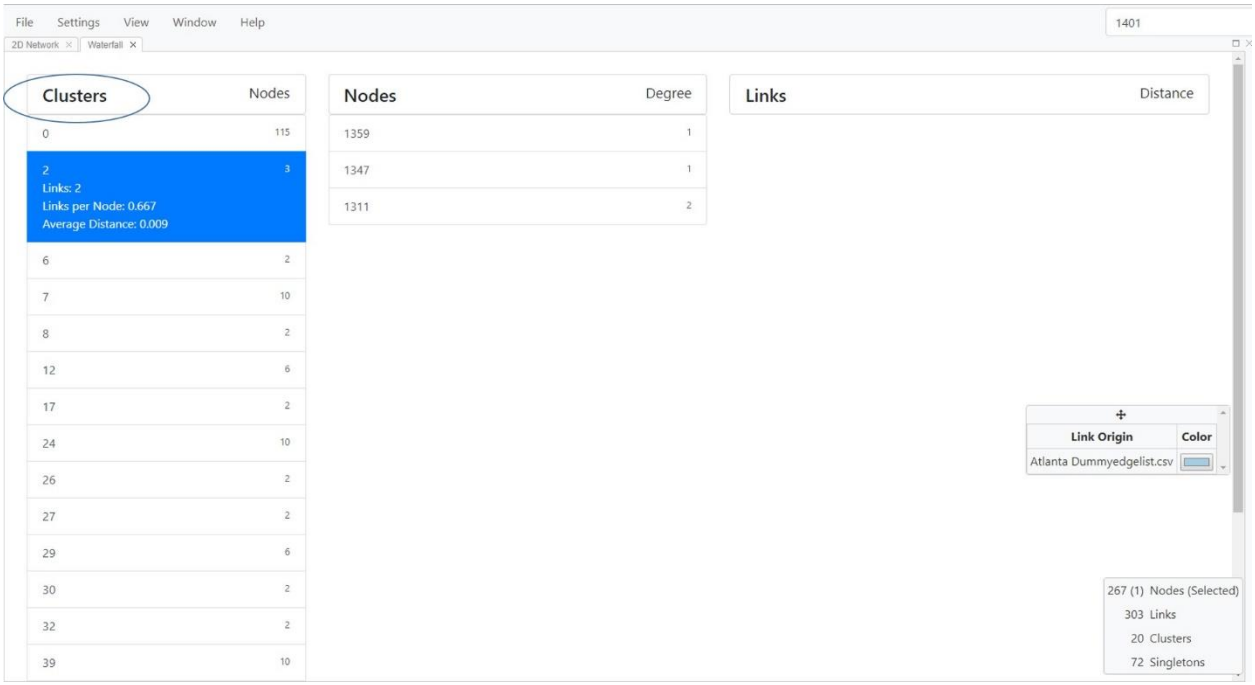


Fig. 69. Waterfall View showing node list by clicking on a cluster of interest.

You can view detailed information for each of these nodes by clicking on them (Fig. 70). MicrobeTrace will then open a column with links associated with the selected node.

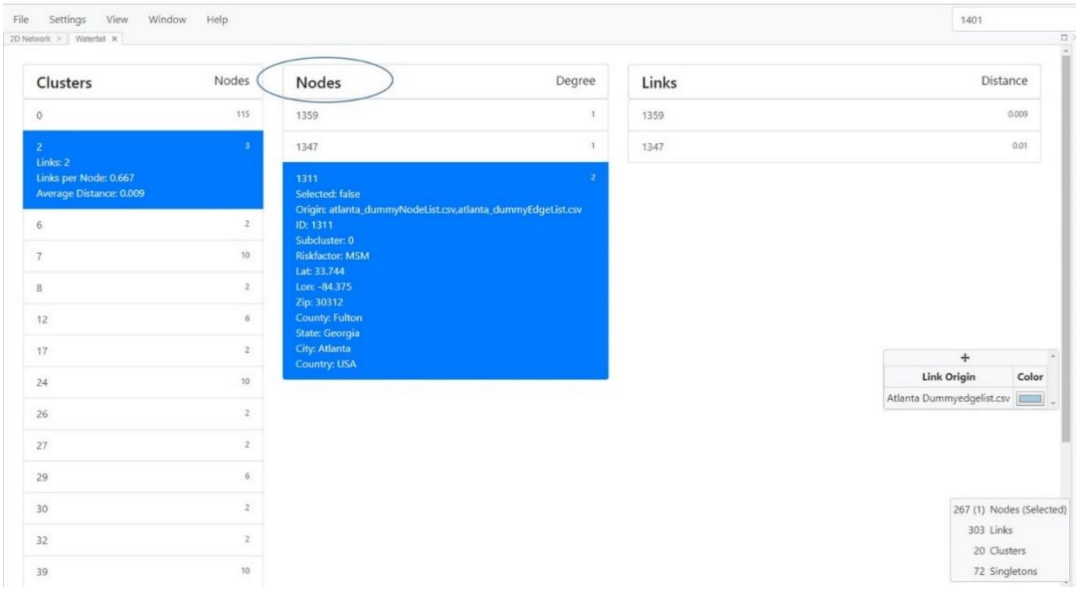


Fig. 70. Viewing node details in dropdown/text format.

Similarly, you can view the link details by clicking on each link ID (Fig. 71).

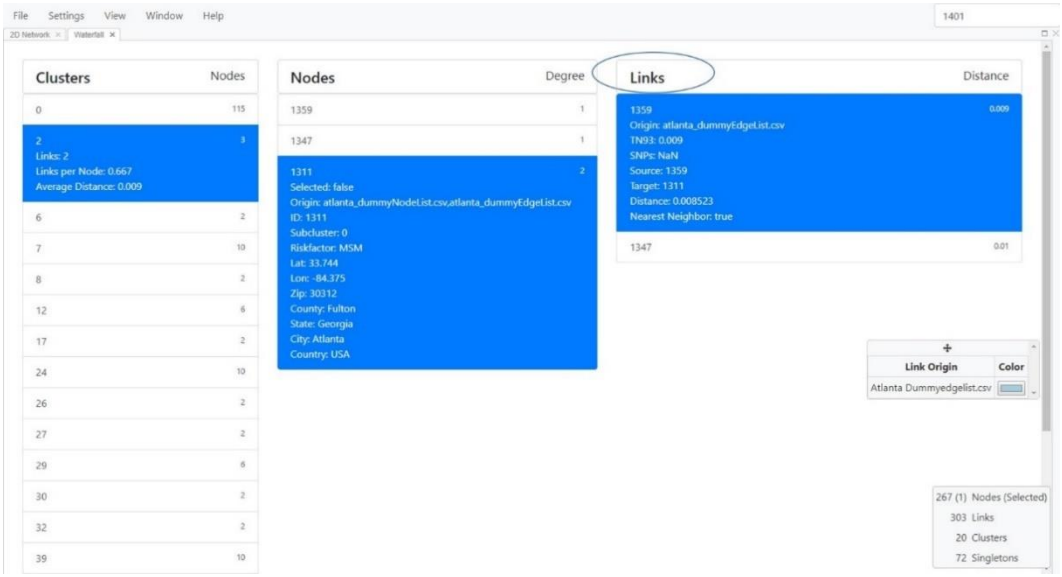


Fig. 71. Waterfall View showing link details

Map View

Nodes in the data can be displayed in a global, geographic map using Map View if geo-coordinates are included in the node file. The map allows you to zoom to the geolocation indicated in the node list. Although latitude and longitude give you the most precise location, zip codes and other geopolitical demarcations (counties, and states) are also rendered on the map if latitude longitude data is unavailable.

Select **View** and select **Map** (Fig. 72) to display the Map View, which by default will just display an empty map (Fig. 73).

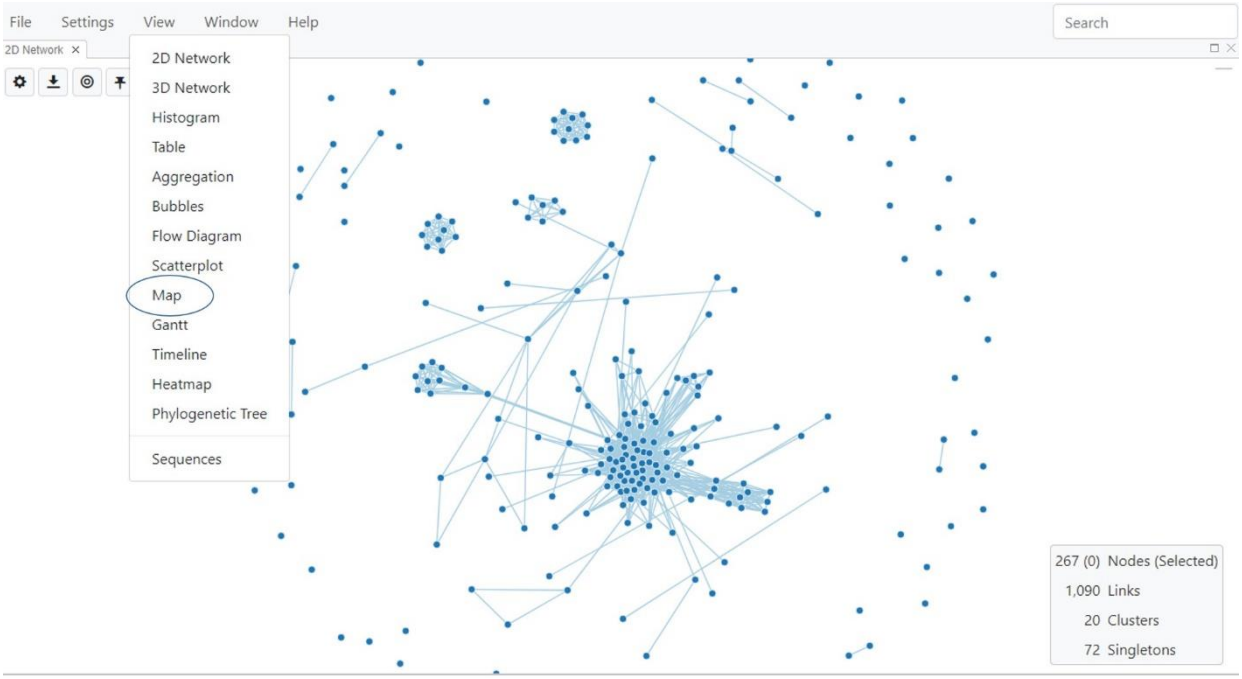


Fig. 72. Selecting Map view

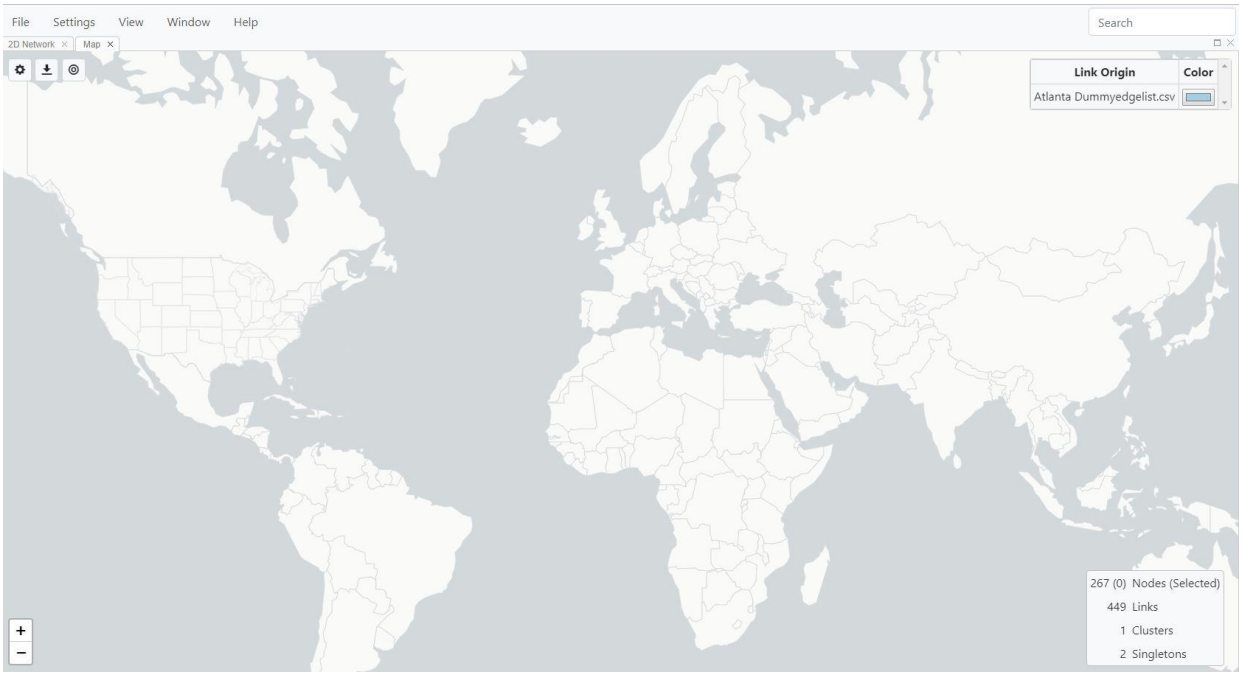


Fig. 73. The default setting for Map View is without the nodes displayed.

Map View Settings

Use the various settings as described in the following sections to customize visualization of your data based on the geo-information included in your node file.

Select the **Toggle Map Settings** button  to open the settings dialog box, which has four tabs: Data, Components, Nodes, and Links.

***Data tab*:** Selecting **Data** displays the pull-down menu options for this feature.

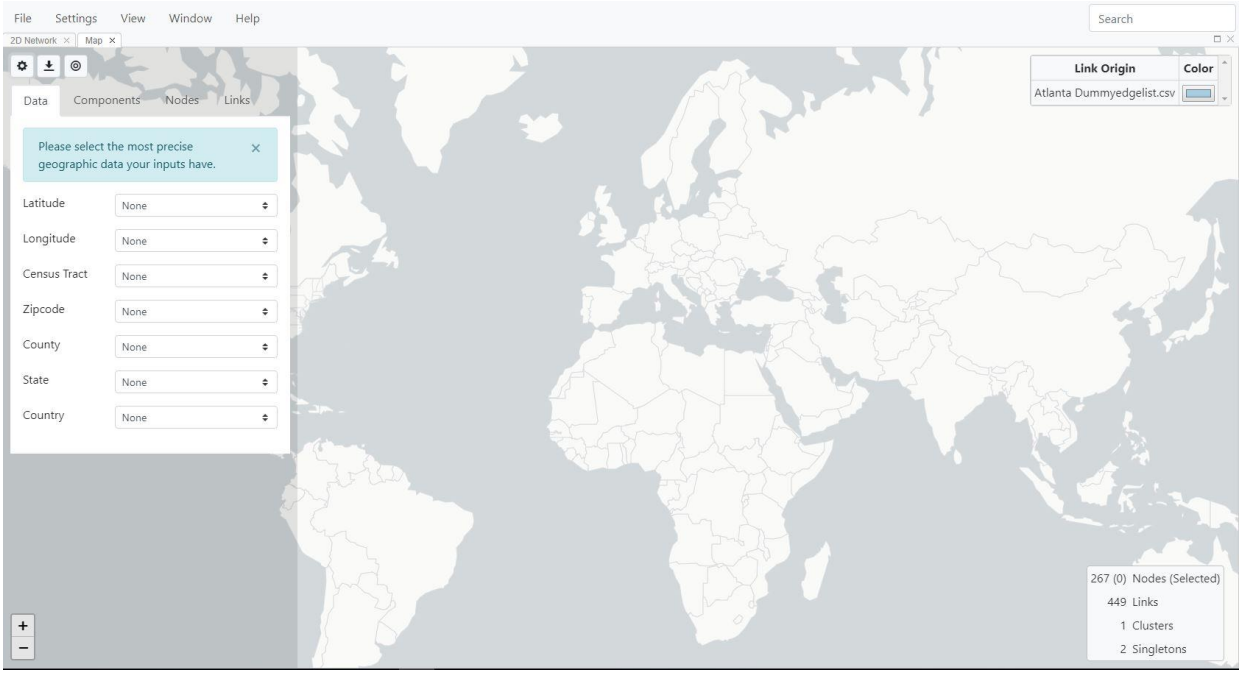


Fig. 74. Changing Map options with the Data tab

In this example, we would like to visualize data by latitude and longitude. ***NOTE*: The map display is hierarchical, so if your data set has all the data columns listed below, and you select multiple properties, the map displayed will default to the highest available level of geographic precision.** Please ensure you select only the variable that works best for your data set, and leave the others as **None**. In this example of a dataset of Atlanta area HIV-1 *pol* sequences, we have latitude and longitude (lat lon) information in the node list, and have selected these as the geographic parameters to use (Fig. 75).

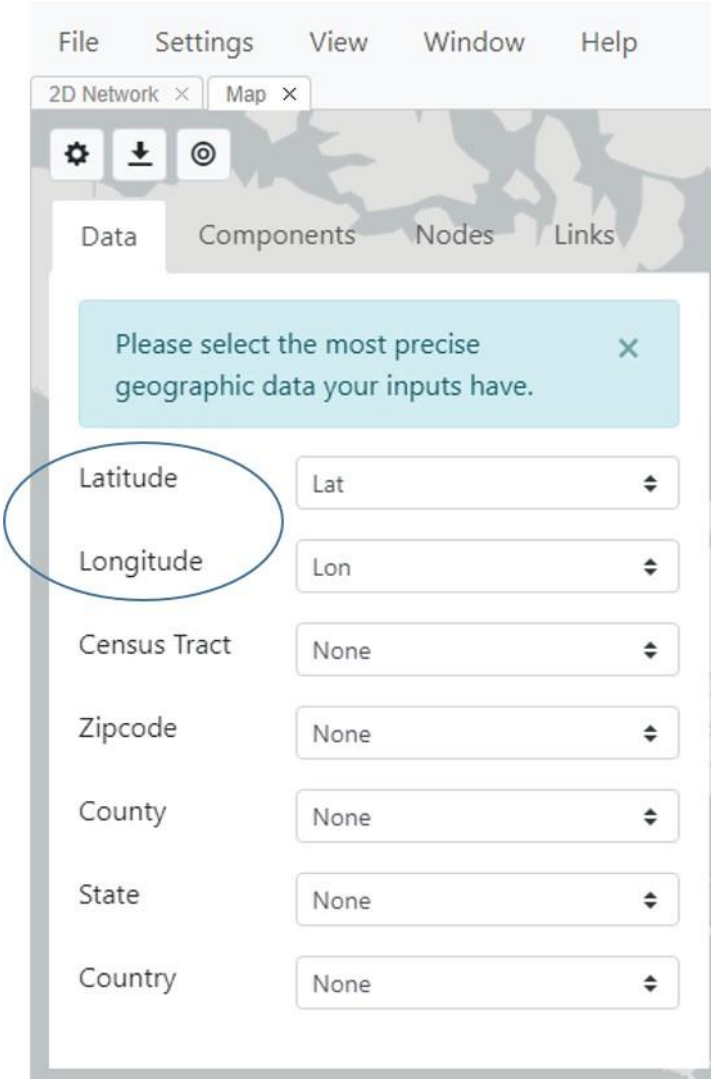


Fig. 75. Map options and selecting latitude and longitude to display geographic location of nodes on a map.

Components tab: Selecting **Components** will enable making changes to the actual network that will be displayed (Fig. 76A), including nodes and links. You can also choose whether to use Map View if you are online or offline (Figs. 76B-C). This option will determine which features are displayed. When offline, you can choose to show or hide countries, states and counties. When online, you can choose to display either base map or satellite view layers that are not available offline in MicrobeTrace. If online and you select these options, MicrobeTrace will download the base map and satellite geographic data (called tiles) from the internet, which are similar to the Google map features. MicrobeTrace also has the capability to load GeoJSON files if you have generated data with specific location information in that file format. To add GeoJSON files into

MicrobeTrace, select **User-Provided,** browse to the location on your computer where the file is stored, and load the file (Fig. 76D).

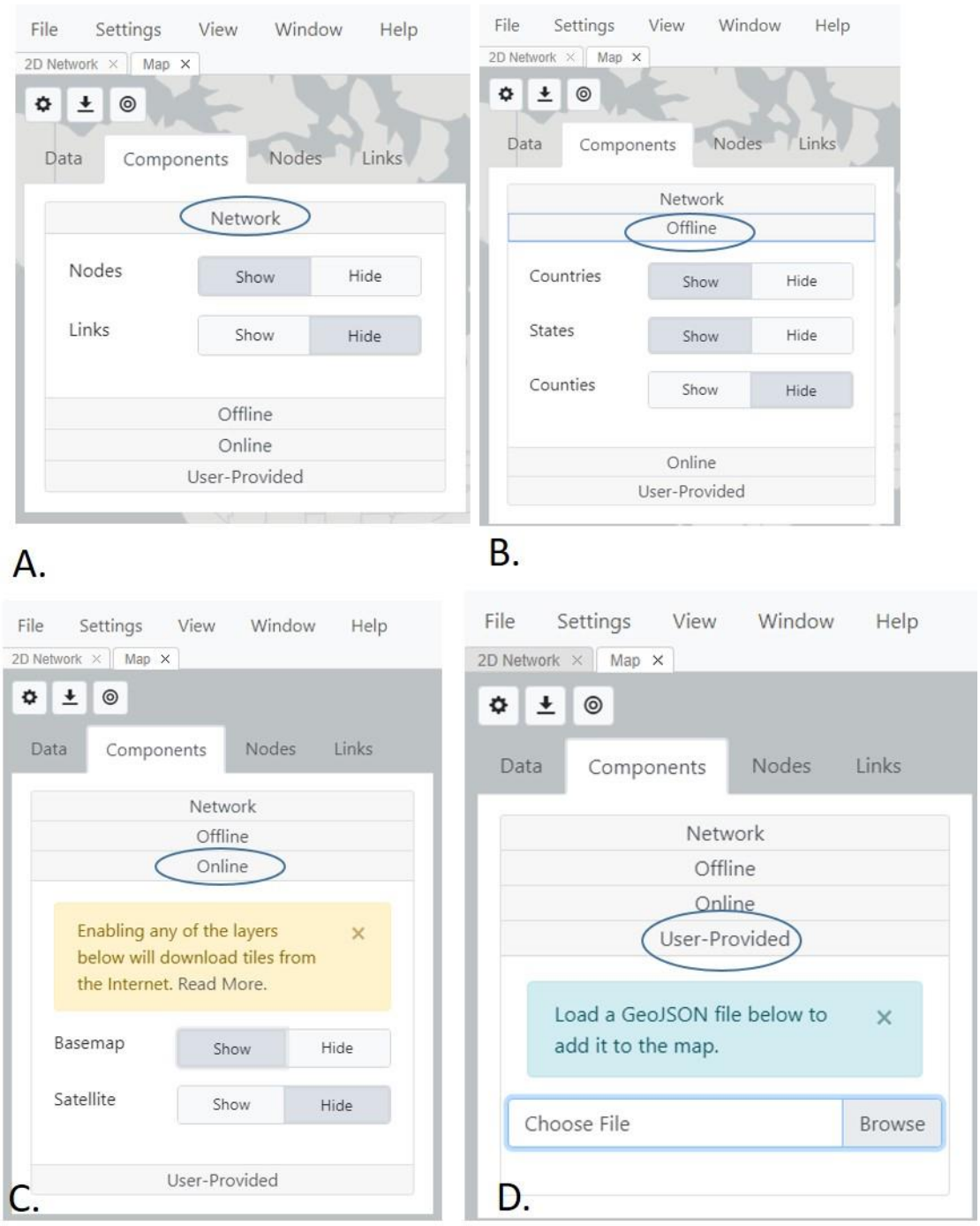


Fig. 76. Map settings, Components tab. A. Network, B. Offline, C. Online, and D. User-Provided tabs.

Nodes tab: Select **Nodes** to change the appearance of nodes on the map (Fig. 77). Nodes can be colored by any variable in your node file. Selecting Color Options opens up the Global Settings menu, and you can customize the Style settings, background, etc. as in other views.

The transparency and jitter speed of the nodes are changed using the respective slider bars. ****PLEASE NOTE: In many datasets, there can be many nodes that share the same geographic co-ordinates causing a very high node density in the map such that these nodes appear as a single large dot. In order to separate nodes with the same geographic coordinates, use the jitter slider bar (Fig. 50, circled) to increase the jitter level to separate or jitter the nodes so they are visible.*** Use **Tooltip** to change which variables are displayed when the mouse pointer is placed ("hovers") over a node. For example, if you choose **ID** from the Tooltip drop-down menu, the node ID will be displayed when the mouse pointer is over that node.

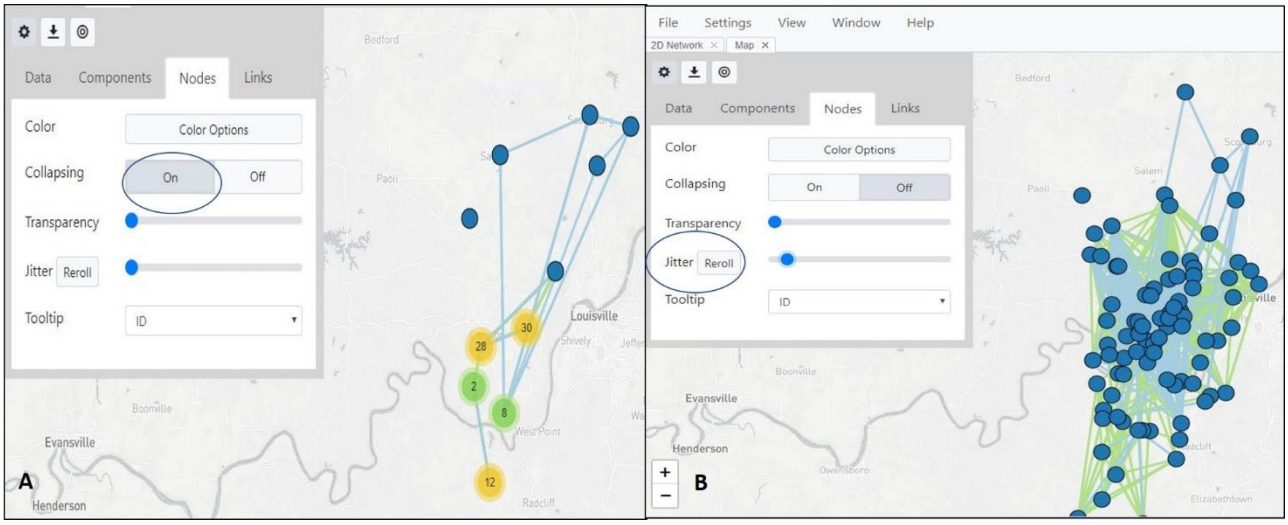


Fig. 77. Map View node settings to change how the node color and tooltip features are displayed.

Node collapsing can be on (panel A), or off (panel B). Increasing jitter spreads out the nodes randomly; clicking on reroll randomly scrambles the nodes around, a useful feature for sensitive data.

Links tab: Select **Links** to change the link settings (Fig. 78). The features are identical to those in the Node tab. You can adjust color, transparency and tooltip settings.

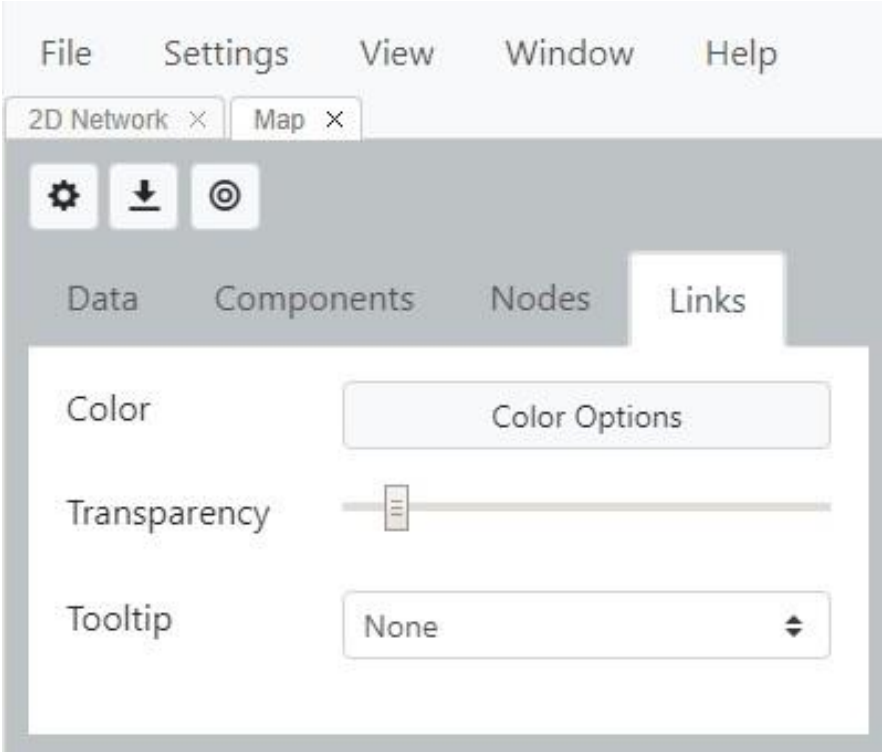


Fig. 78. Map View Links settings

MicrobeTrace displays the node data in the map based on the various options you selected. Fig. 79 shows the map with nodes colored by risk factor, and links hidden. When viewing a map, the scroll bar on your mouse can be used to pan around or zoom in and out. By default, the map is zoomed out, and you see a circle with a number that represents the number of nodes (Fig. 79, top panel). When you zoom in, the nodes pop out to form smaller, more discrete groups. Individual or multiple nodes can be selected or de-selected by using the mouse pointer. These selections will propagate to the Network and Table Views. This enables tracking of particular individuals between multiple visualization windows. As with other views, map images can be exported and saved as .png, .svg, or .jpg image files.

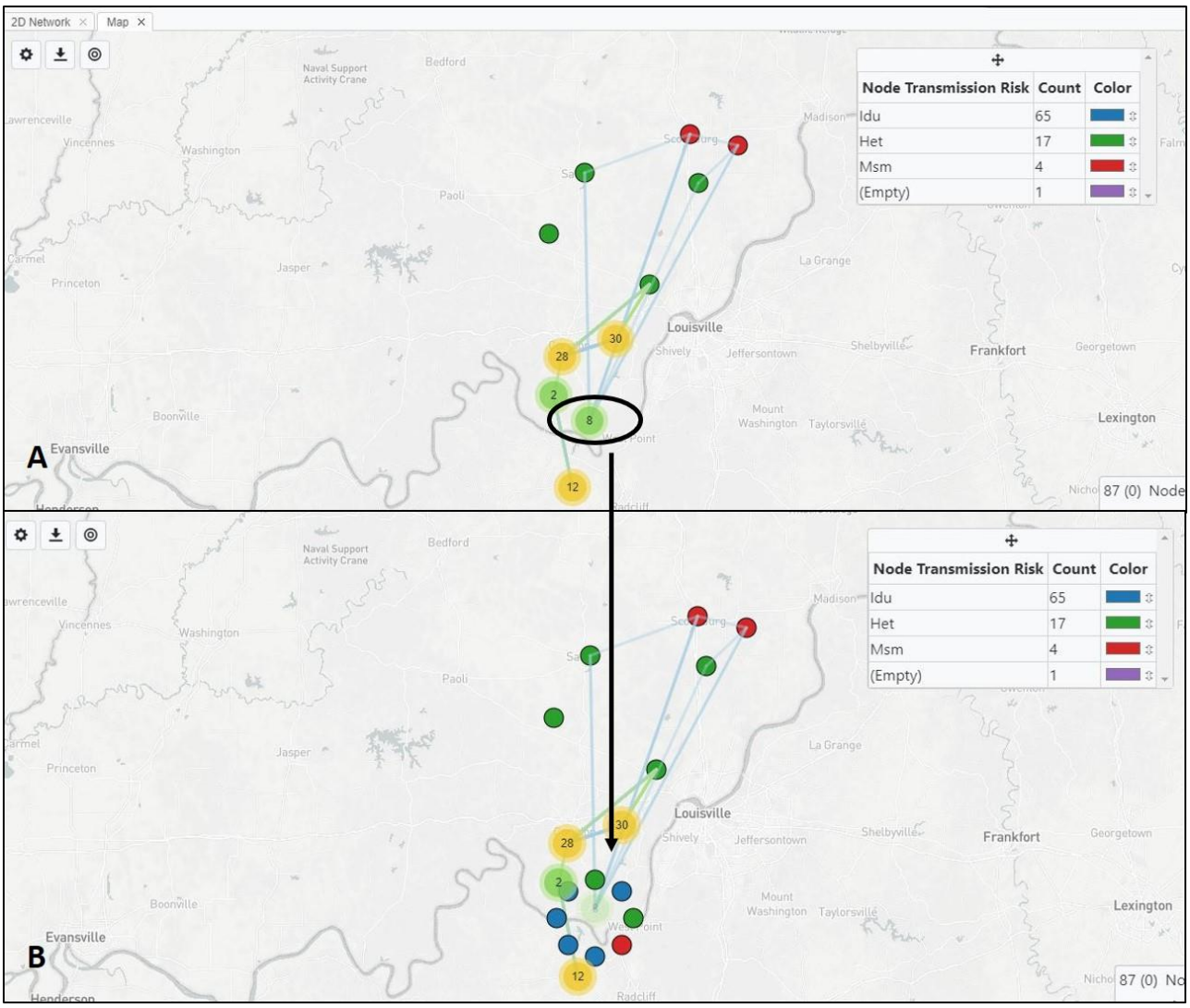


Fig. 79. Map View showing nodes mapped to latitude and longitude coordinates, colored by risk factor, and base map displayed. Nodes are collapsed, jitter is off in Panel A. Panel B shows the node distribution when you hover over a collapsed node; component nodes are displayed as a circle or spiral. *All data are hypothetical, for demonstration purposes only*

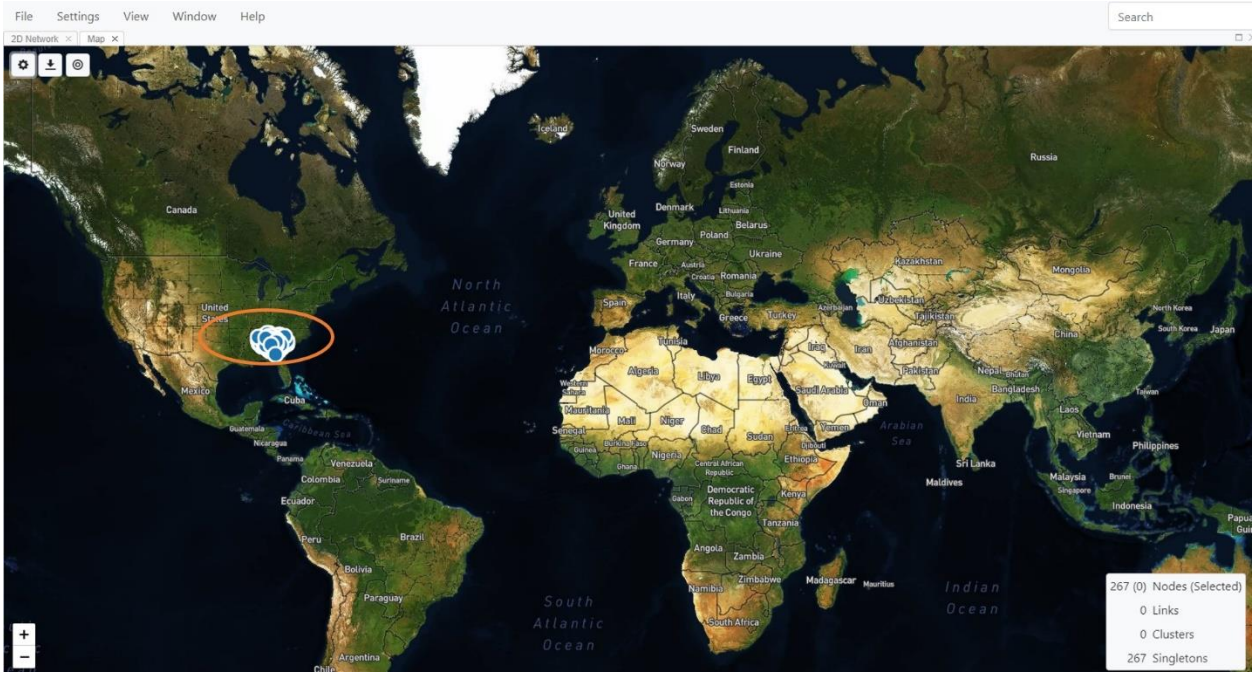


Fig. 80. Map View showing nodes mapped to latitude and longitude coordinates (circled), colored by risk factor, and satellite features displayed.

Globe View

This view is like the map view in that nodes in the data can be displayed in a global format if geo-coordinates are included in the node file. The difference is that in this visual, nodes and links are plotted on a three-dimensional globe which can be rotated. This view is especially useful for international outbreaks or to look at potential connections across countries. The layout and options are very similar to those in the map view. Although latitude and longitude give you the most precise location, zip codes and other geopolitical demarcations (counties, and states) are also rendered on the map if latitude longitude data is unavailable.

Select **View** and select **Globe** (Fig. 81) to display the Globe View.

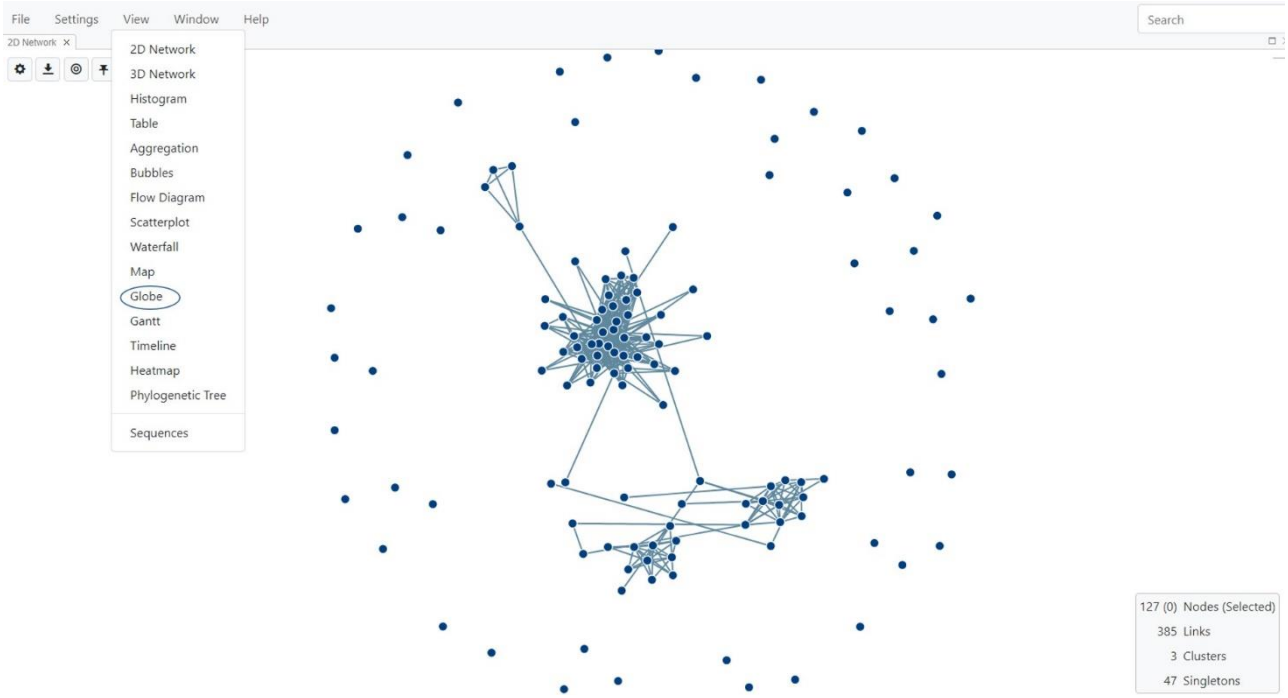


Fig. 81. Selecting Globe view

The opening view is a standard 3-D globe (Fig. 82).

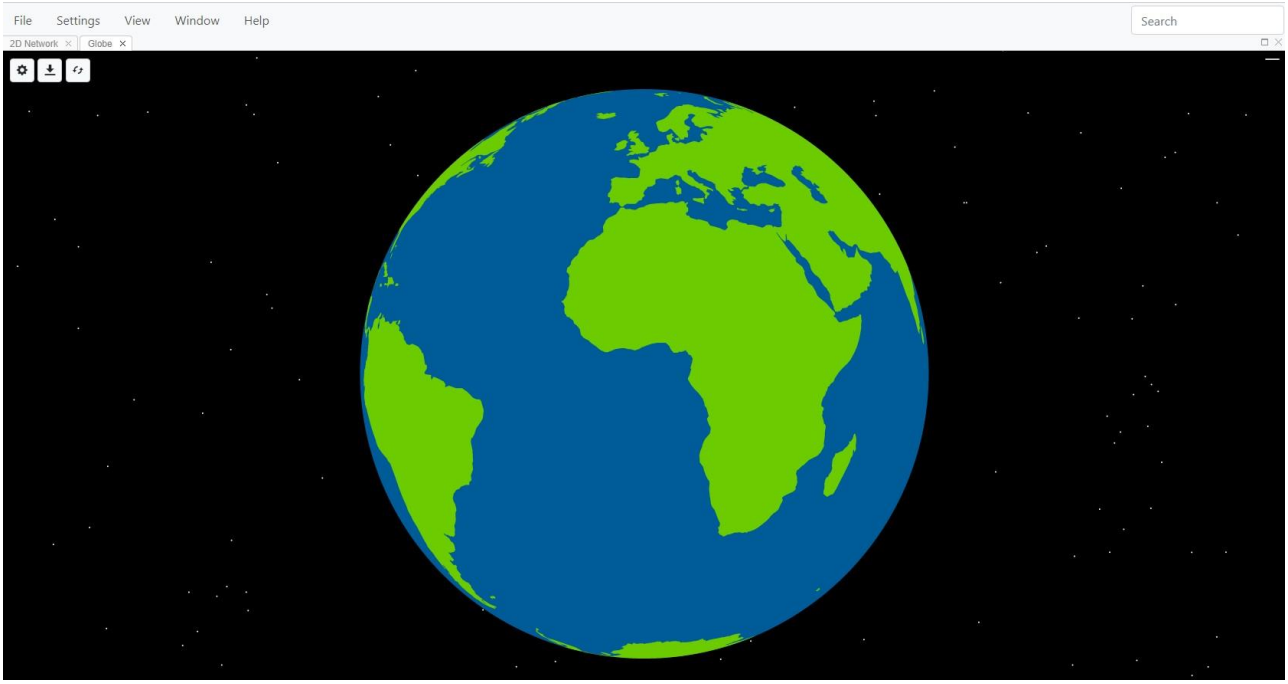


Fig. 82. Globe View - opening window

To display nodes, they need to be mapped to the geographic coordinates provided in your node list.

Changing Globe View Options

Use the various settings as described in the following sections to customize visualization of your data based on the geographic information included in your node file.

Select the **Toggle Map Settings** button  to open the settings dialog box, which has four tabs: Data, Layers, Nodes, and Links (Fig. 83).

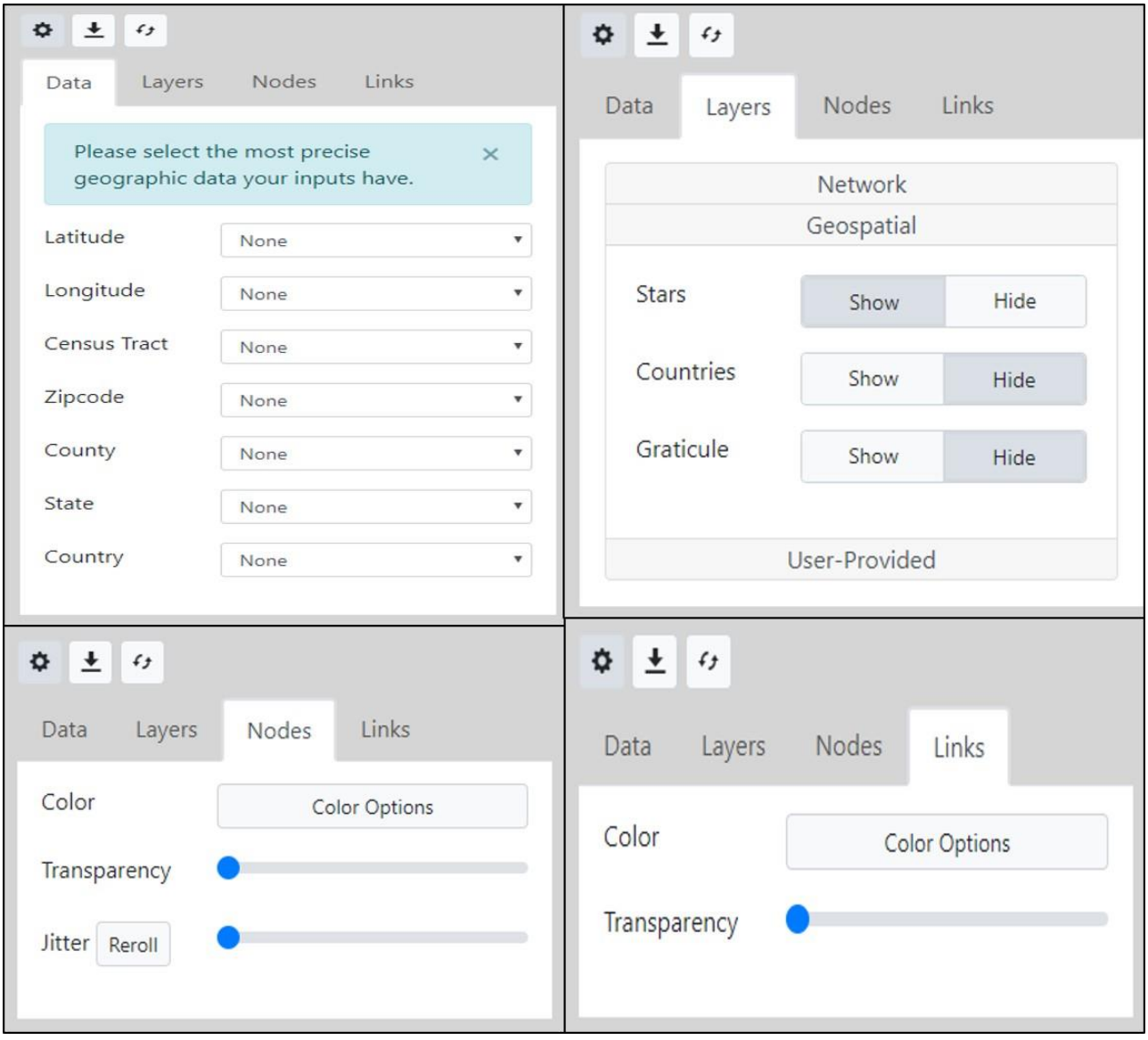


Fig. 83. Globe view showing the available settings using the Data, Layers, Nodes and Links tabs

Data tab

Selecting **Data** displays the pull-down menu options for this feature. In this example, we would like to visualize data by latitude and longitude. ***NOTE*: The map display is hierarchical, so if your data set has all the data columns listed below, and you select multiple properties, the map displayed will default to the highest available level of geographic precision.** Please ensure you select only the variable that works best for your dataset and leave the others as **None**. In this example dataset of European HIV-1 *pol* sequences, latitude and longitude (lat-lon) information is in the node list and are selected as the geographic parameters to use (Fig. 84).

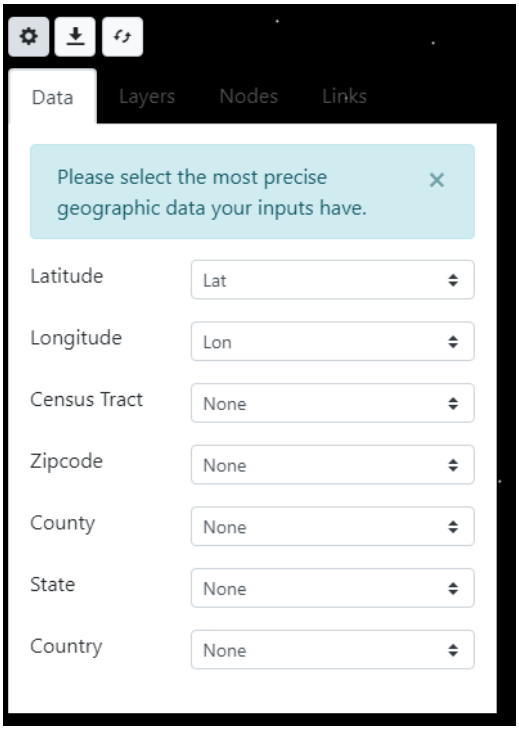


Fig. 84. Data tab for selecting latitude and longitude to display geographic location of nodes on the globe.

Layers tab: Selecting **Layers** will allow you to further customize the display using Network, Geospatial and User-Provided data (Fig. 85).

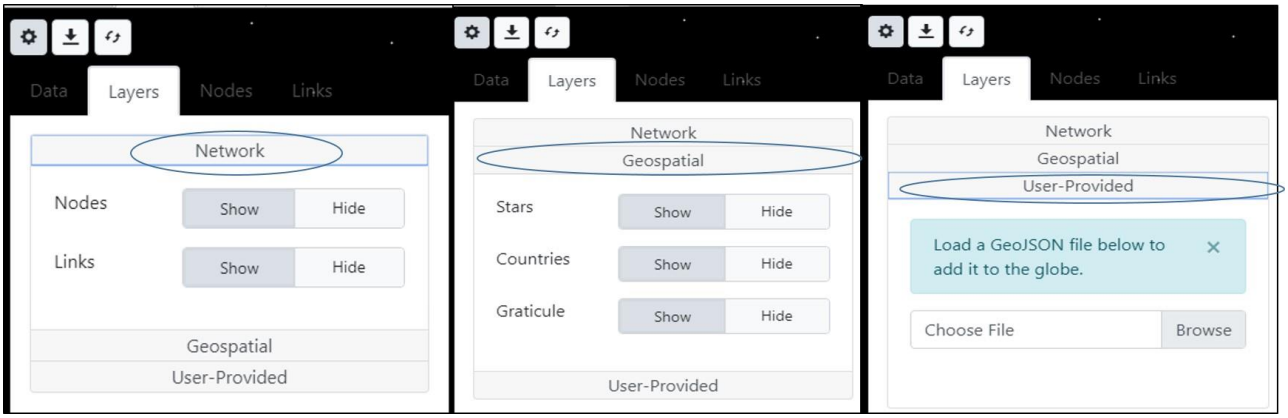


Fig. 85. Customizing Globe view settings with the data Layers tabs

Network: This option allows the network nodes and links to be displayed on the globe.

Geospatial: Allows certain geospatial features to be displayed, including stars, countries, and graticule (lines on the globe indicating longitude and latitude). The night sky displayed in the background was developed to mirror real constellations.

User-provided: MicrobeTrace also has the capability to load GeoJSON files you have generated with specific location information in that file format. To add GeoJSON files into MicrobeTrace, select **User-Provided,** browse to the location on your computer where the file is stored, and load the file.

Nodes tab: Select **Nodes** to change the appearance of the nodes on the map (Fig. 86). Nodes can be colored by any variable in your node file. Selecting Color Options opens up the Global Settings menu and you can customize the style settings, background, etc. as in other views.

The transparency and jitter speed of the nodes are changed using the respective slider bars. ****PLEASE NOTE*: In many datasets, there can be many nodes that share the same geographic co-ordinates causing a very high node density in the map such that these nodes appear as a single large dot. In order to separate nodes with the same geographic coordinates, use the jitter slider bar to increase the jitter level to separate or jitter the nodes so they are visible.***

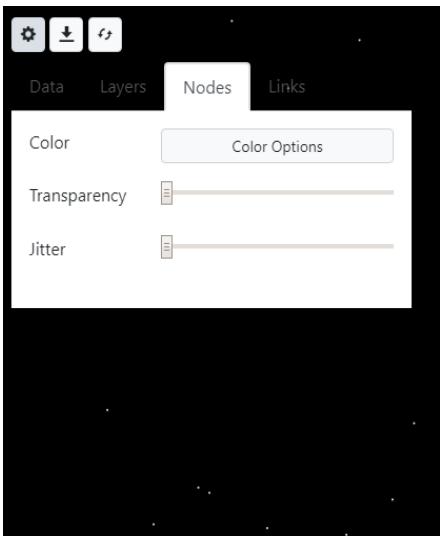


Fig. 86. Changing node settings in Globe View

Links tab: Select **Links** to change the link color and transparency settings. The features are identical to those in the Node tab.

MicrobeTrace displays the node data on the globe based on the various options you selected. Fig. 87 shows the map with node colors changed with links shown. Use the scroll bar on your mouse to pan around or zoom in and out. Moving your mouse while left clicking allows you to rotate the globe and easily view intercontinental network connections. In addition to the usual buttons on the top left corner, the Globe View has an additional rotate button  . Clicking on this icon slowly rotates the globe. As with other views, map images can be exported and saved as .png, .svg, or .jpg image files.

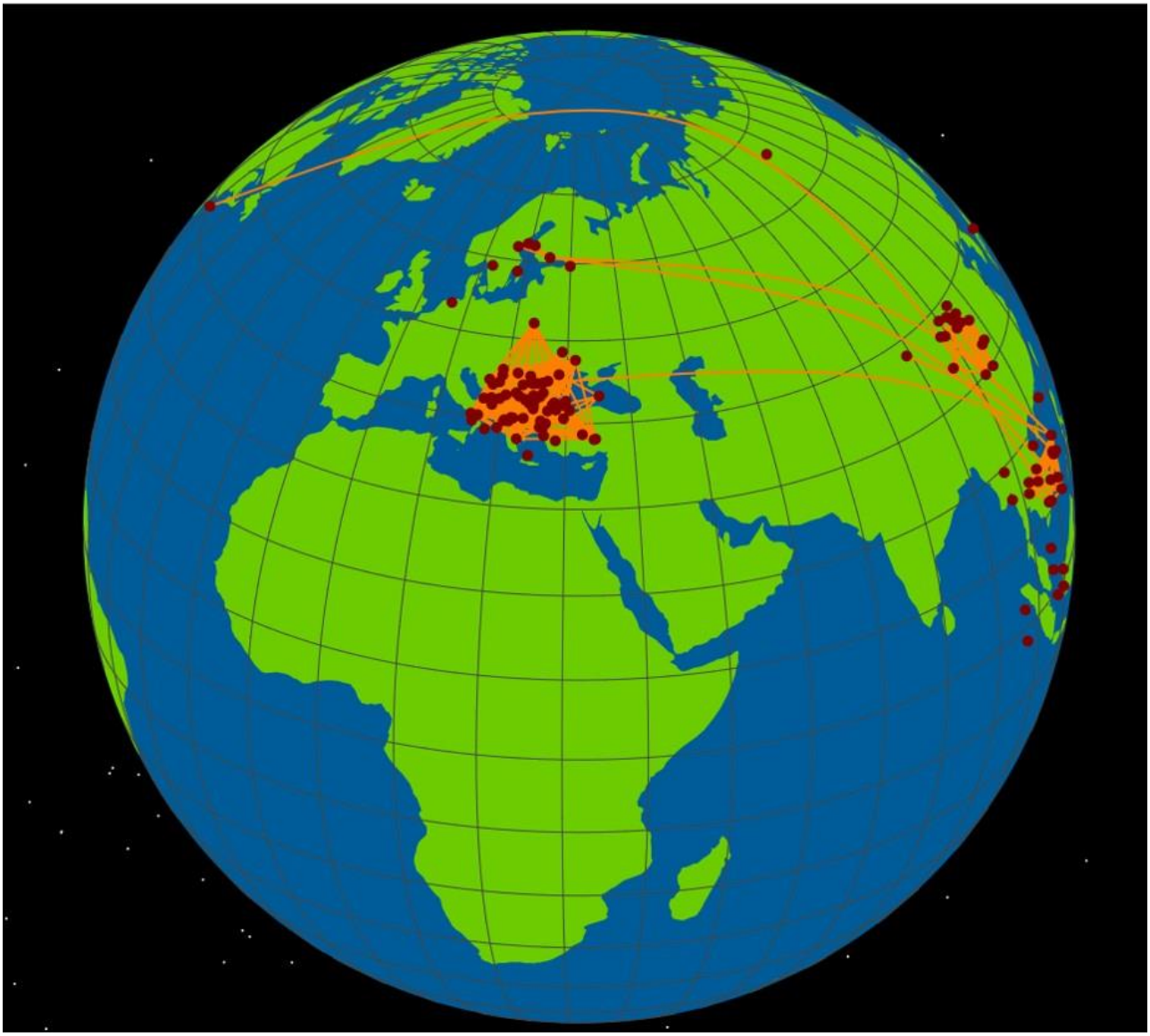


Fig. 87. Globe view with customized settings , nodes mapped to latitude longitude coordinates; node and link colors changed, jitter increased.

Gantt View

This view is used with data that includes time spans or time increments in the form of start date and end date columns, such as infectious periods, lengths of time in a hospital or other institution, or for resource management. The Gantt chart plots the time span on the X-axis against node (individual) IDs on the Y-axis. The chart enables you to look at one or more time spans for an individual and look for overlaps if desired. Select **Gantt** from the **View** menu (Fig. 88).

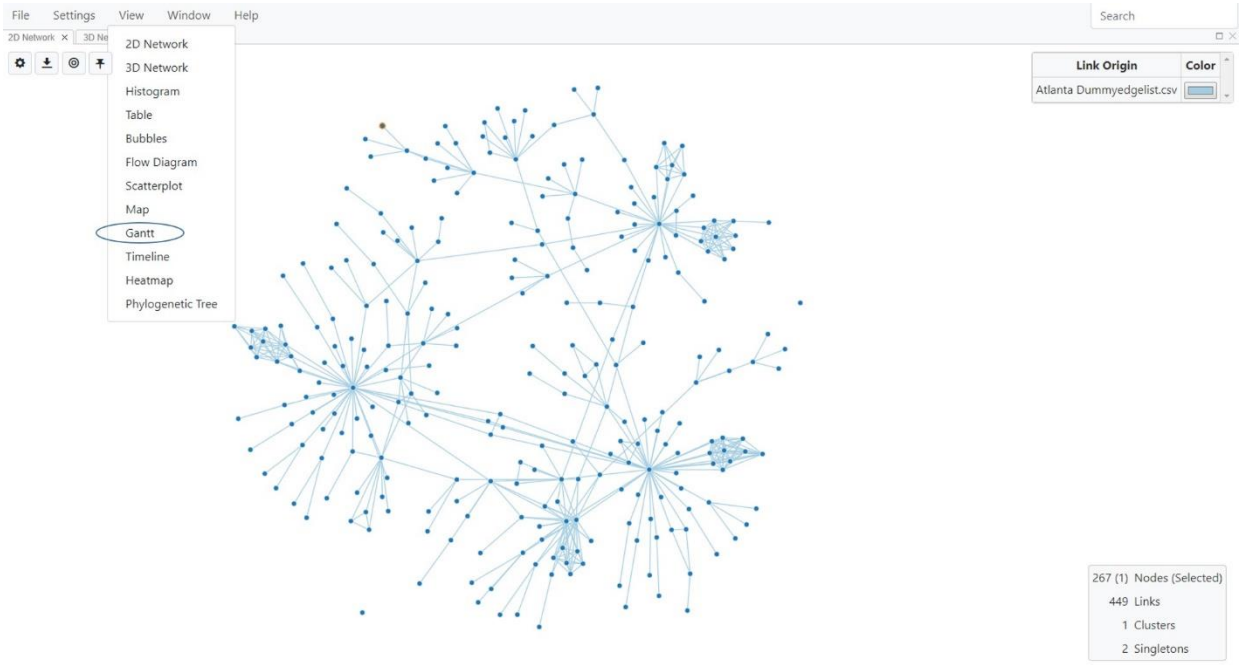


Fig. 88. Selecting Gantt chart view

The Gantt chart is initially blank but selecting the **Toggle Gantt Settings** button  opens a menu to add dates that are included in the node list (Fig. 89). Select **Add Date** and select a start and end date from the dropdown menu. If the data has multiple columns for start and end date, you can select **Add Date** again to add more dates. You can also set colors for each time span.

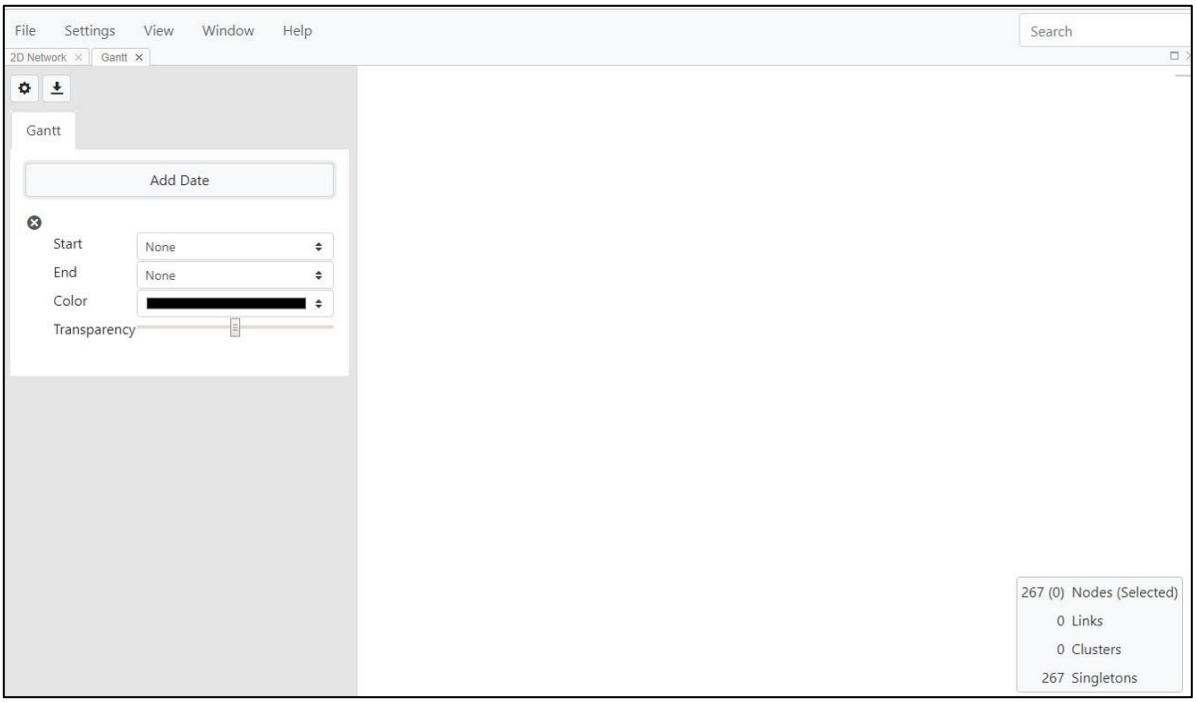


Fig. 89. Menu to add one or more date ranges in the Gantt chart view and select color scheme for

each range.

Changes will be immediately reflected in the displayed Gantt chart (Fig. 90).

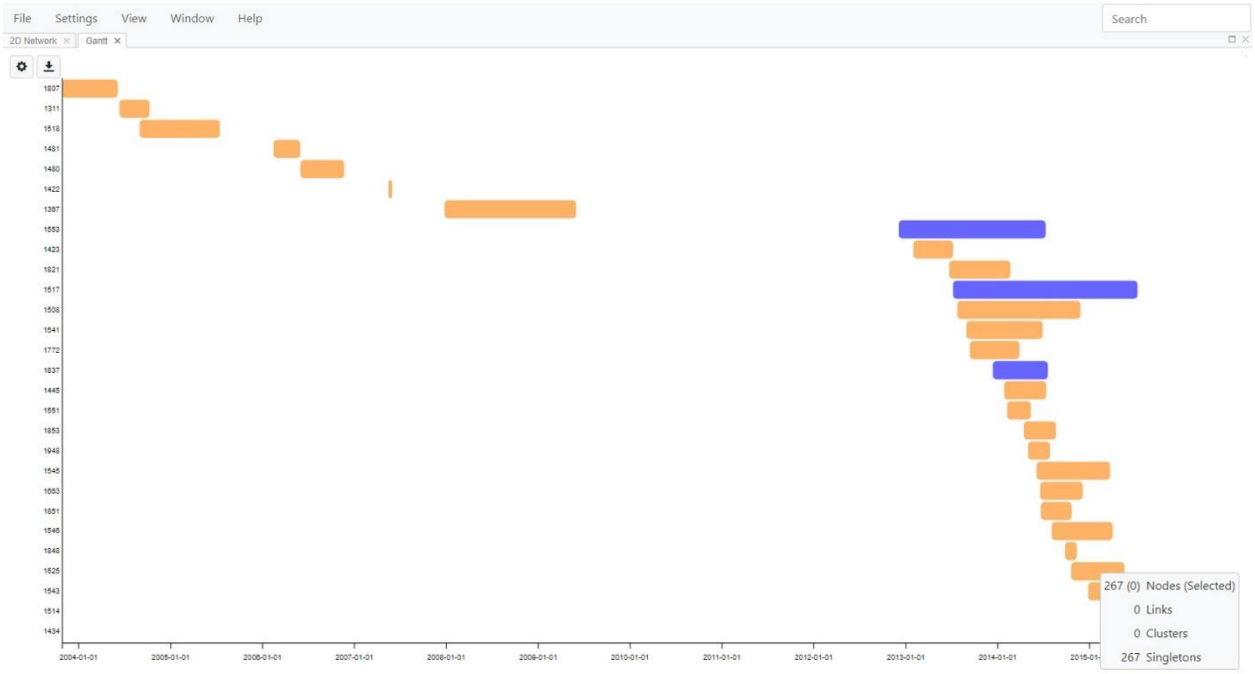


Fig. 90. Gantt chart with two timespans selected and plotted against node ID on Y-axis. Each time period is indicated with a different color.

Epi Curve

Graphical representation of dates plotted on the X-axis can be visualized using the Epi Curve View, which graphs your data on a timeline to create either a cumulative or non-cumulative epi curve. Select **Epi Curve** from the View menu (Fig. 91). This opens a window without any selection for the date field (Fig. 92).

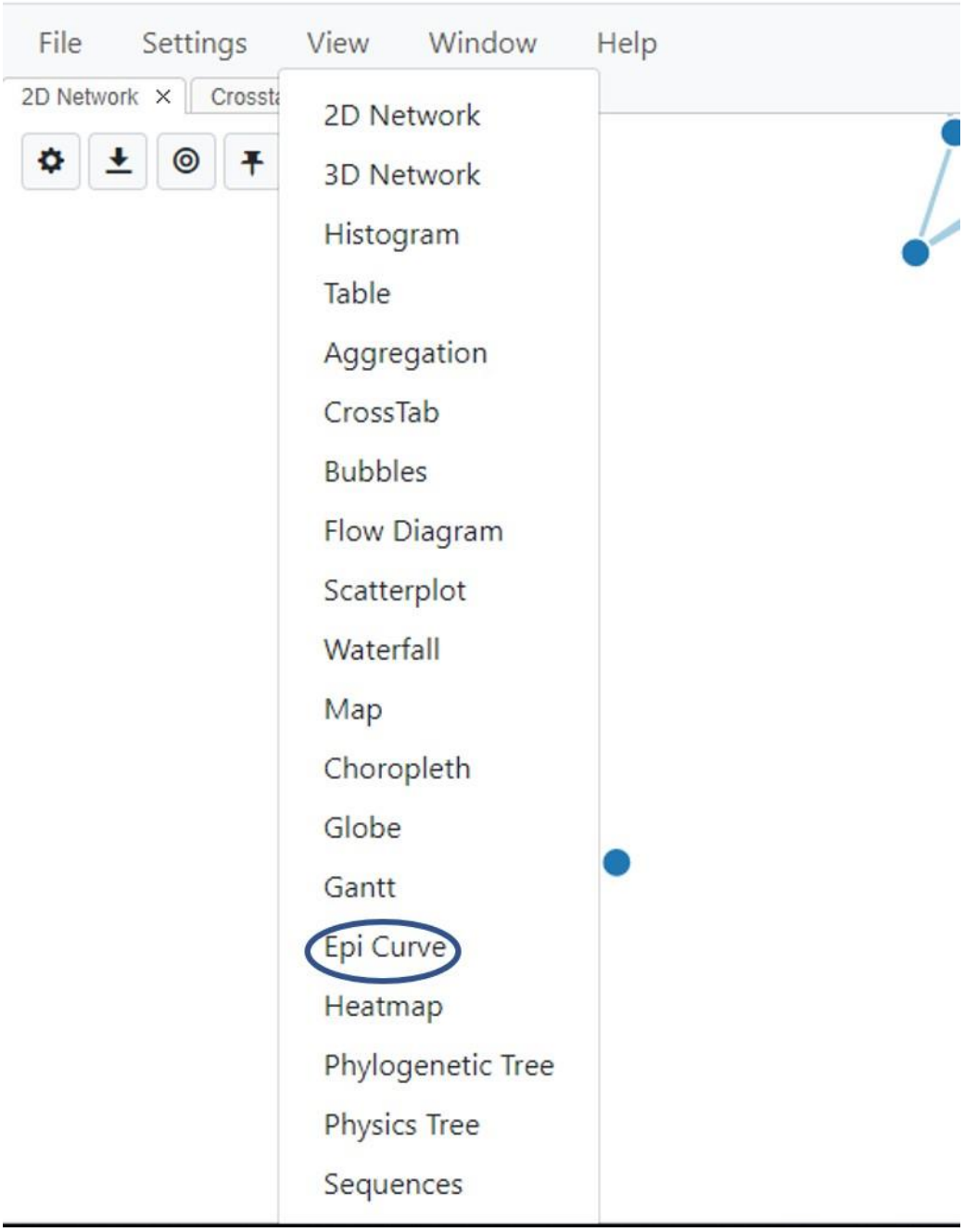


Fig. 91. Selecting Epi Curve View

Select the relevant date field from the dropdown menu (Fig. 92). You can select either a cumulative or non-cumulative epi curve.

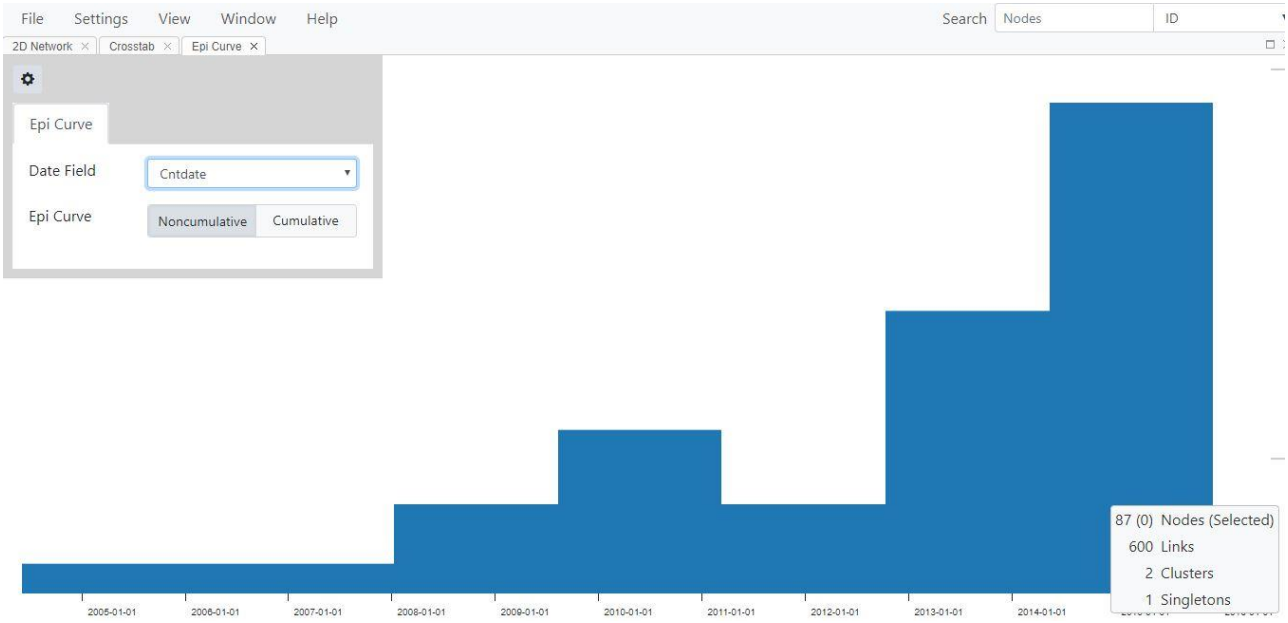


Fig. 92. Epi Curve view showing the default view (no dates selected).

This feature is typically used in conjunction with the Network View. See Tiling Views for instructions on how to arrange one or more views side by side or in a tiled pattern for simultaneous analysis.

Heatmap View

If the analysis includes a FASTA file with nucleotide sequences, then MicrobeTrace offers a heatmap visualization of the calculated nucleotide genetic distance matrix.

To display the heatmap, select **View** on the Menu bar and then select **Heatmap** (Fig. 93) to open a new window showing the genetic distances in a heatmap matrix (Fig. 94).

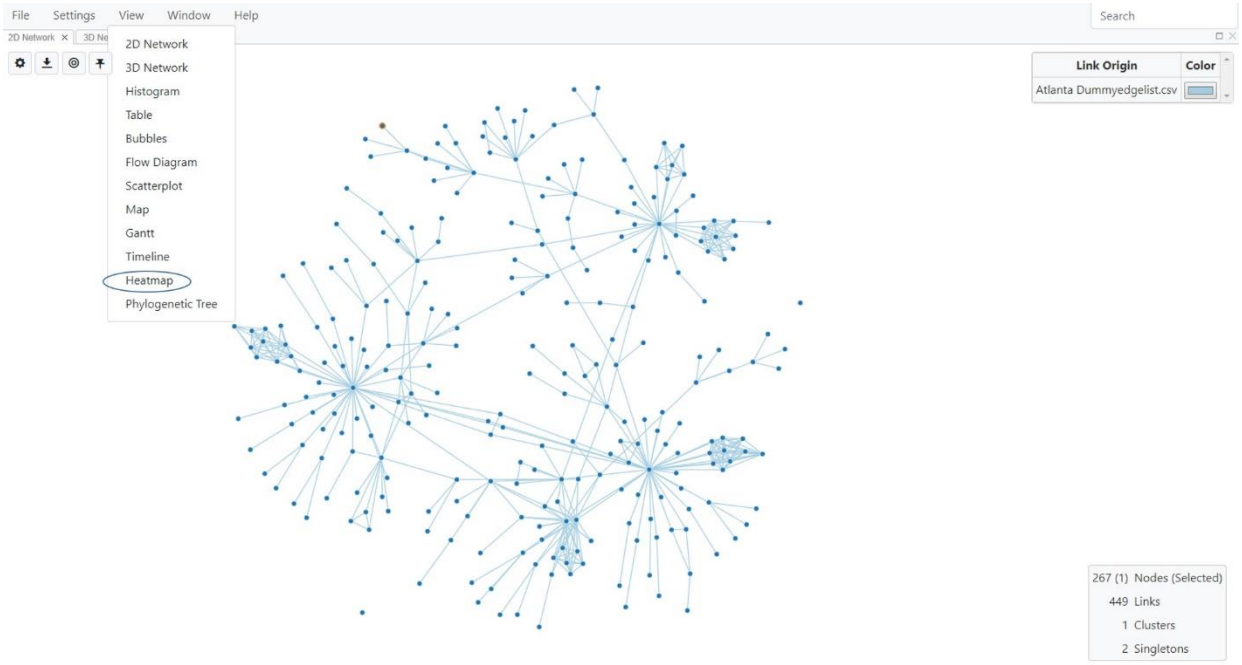


Fig. 93. Selecting Heatmap View

Note that rendering the genetic distance matrix takes a moment, so the window will appear blank until the computation is completed, and the genetic distance heatmap loads.

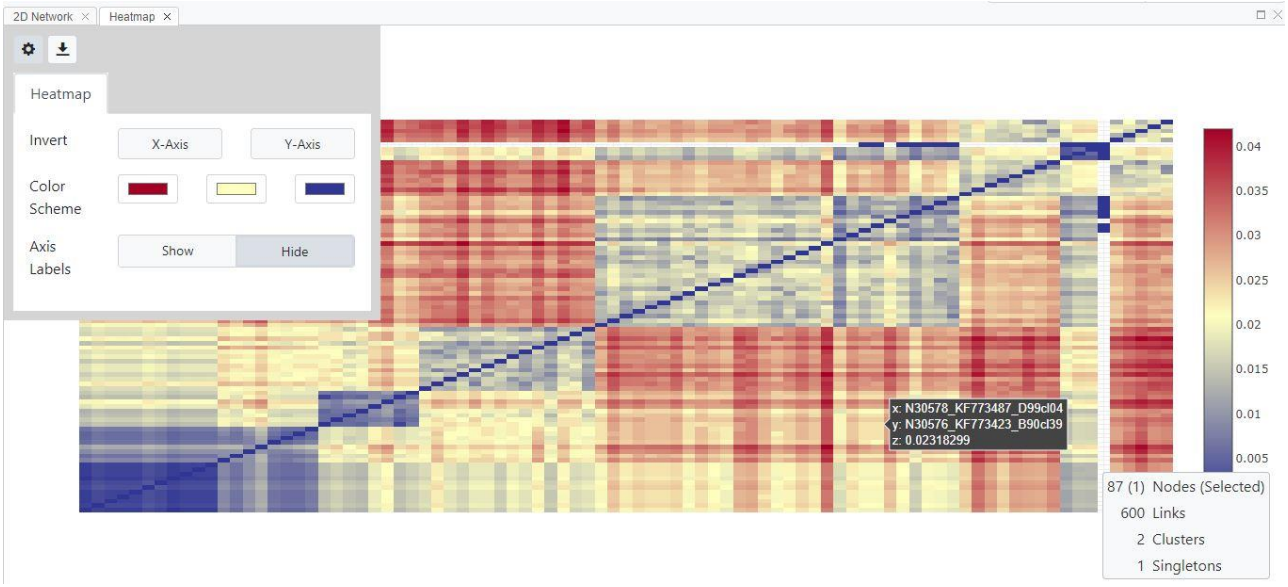


Fig. 94. Heatmap View of the genetic distance matrix of sequences in the FASTA file

Each cell in the matrix represents the genetic distance (substitutions/site) between two sequences in the dataset. The cells across the diagonal (bottom-left to top-right) represent a comparison of a sequence to itself (i.e., will have a genetic distance of zero, resulting in the visibly distinct diagonal line). The scale bar on the right of the heatmap indicates the color scheme used for the range of genetic distances in the dataset analyzed. For example, dark blue cells indicate sequences that are more closely related genetically than those in yellow or red. The actual genetic distance value for two sequences can be viewed by hovering the mouse pointer over the desired cell in the matrix. A pop-up bubble will show the IDs of the two sequences and their calculated genetic distance.

Settings can be changed by selecting the Toggle Heatmap Settings button  . This lets you change the distance metric, invert axes, change the color scheme, or to show or hide axis labels.

The heatmap graphic can be saved as .png or .jpg image files by selecting the **Export Heatmap button on the top left corner** (Fig. 95). Use the Export dialog to navigate to the desired destination on your computer and type in a filename for the image to be saved. Select **Save** to save the image file. You can also export the actual distances as a .csv file, which can then be viewed in Excel.

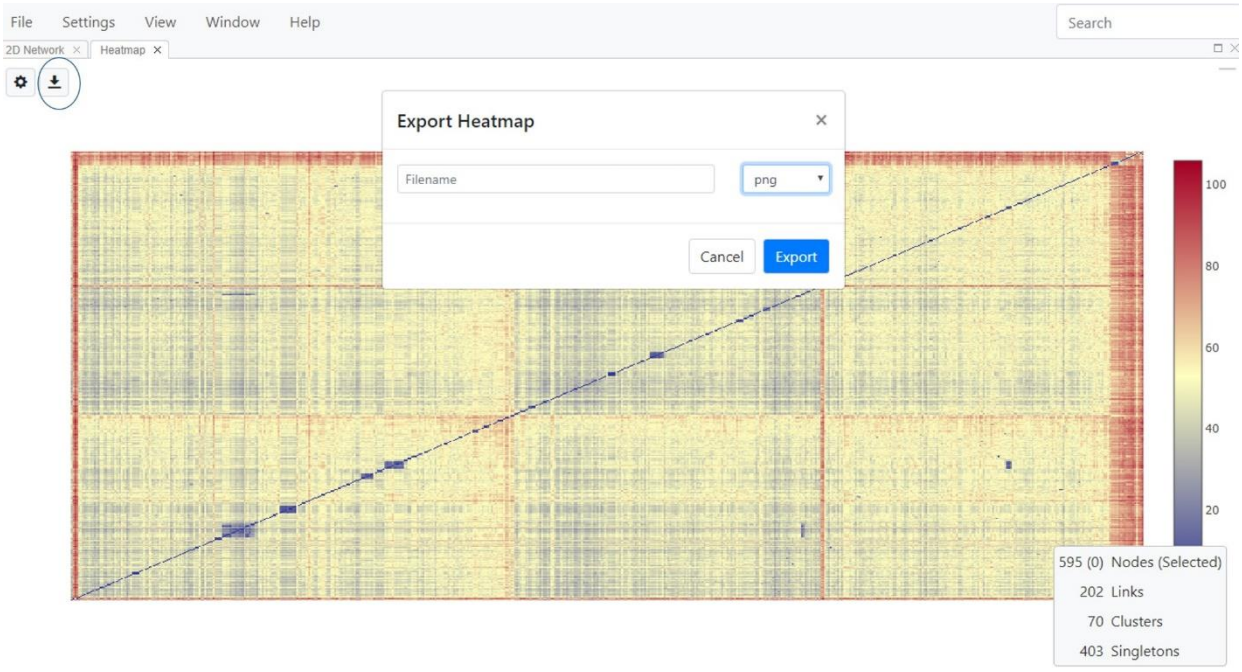


Fig. 95. Exporting a heatmap as either an image file or as a .csv file containing the calculated distances.

Sequence View

The primary function of the Sequence View is to permit a visual inspection of the quality of the final alignment to ensure that there are no unexpected gaps, insertions, etc. in the alignment. An improper alignment will greatly impact determination of genetic distances and thus also the inferred network or phylogenetic tree. This may be especially true for non-HIV sequences, since the aligner in MicrobeTrace is not configured to handle sequences from pathogens other than HIV-1. Non-HIV-1 sequences may not always properly align using the TN93 nucleotide substitution model that is commonly used for HIV-1. The Sequence View may have limited value if a pre-aligned FASTA file is used in the analysis unless the Sequence View is used to re-check that alignment.

IMPORTANT NOTE ***We strongly recommend checking the quality of all pre-made alignments prior to using them in MicrobeTrace. Also, please do not use MicrobeTrace to align a pre-aligned sequence, which*** **may change the alignment.** Please note that any edits made to the alignment in the sequence viewer will not automatically be rendered in the inferred network. The edited sequence alignment file must first be saved and can then be used from the beginning of the analysis in MicrobeTrace at the file loading step.

Select **Sequence** under the **View** menu (Fig. 96) to display the sequences as an alignment in a new window (Fig. 97). Use the scroll bars to maneuver the displayed sequence view.

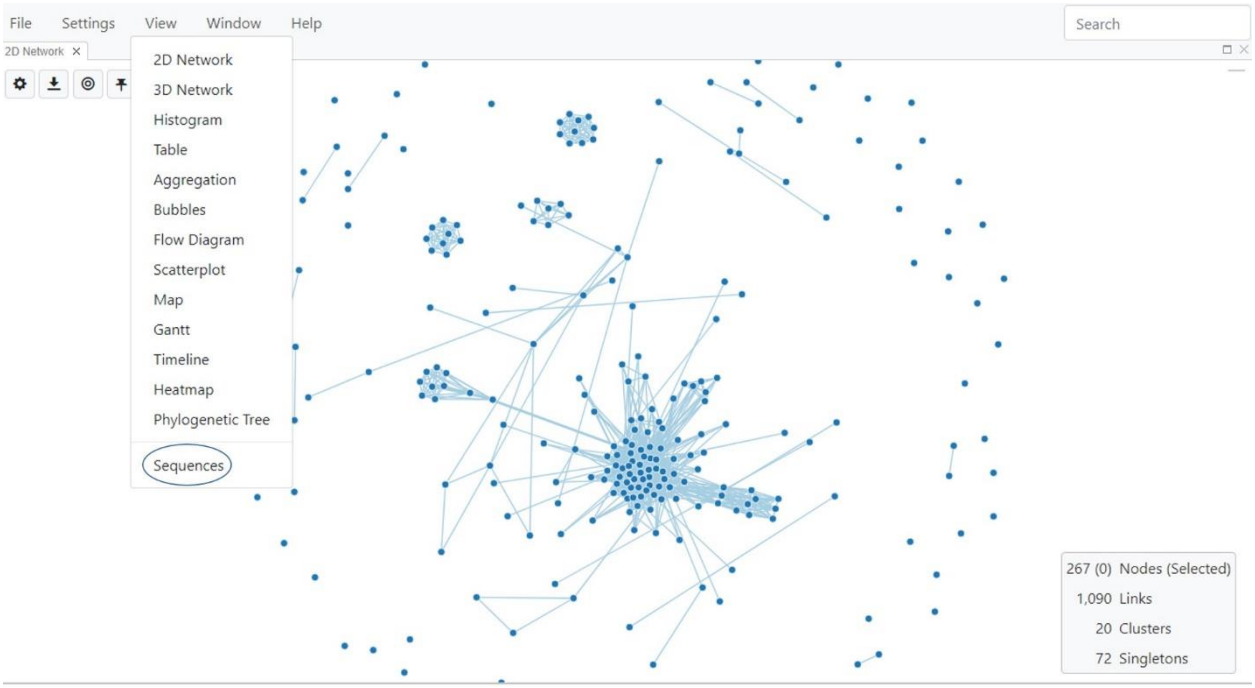


Fig. 96. Selecting Sequence View

You can use the roller on your mouse to scroll horizontally and vertically to check your alignment.

You can change the appearance of the nucleotide characters using the settings button (Fig. 97, circled). You can also export the sequence in multiple formats (png- Portable Network Graphic, FASTA or MEGA, which is a file format used in the Molecular Genetics Evolutionary Analysis (MEGA) program) using the dialog box that appears when you select the download button . The sequence ID will be appended to the sequence in the file format you choose.

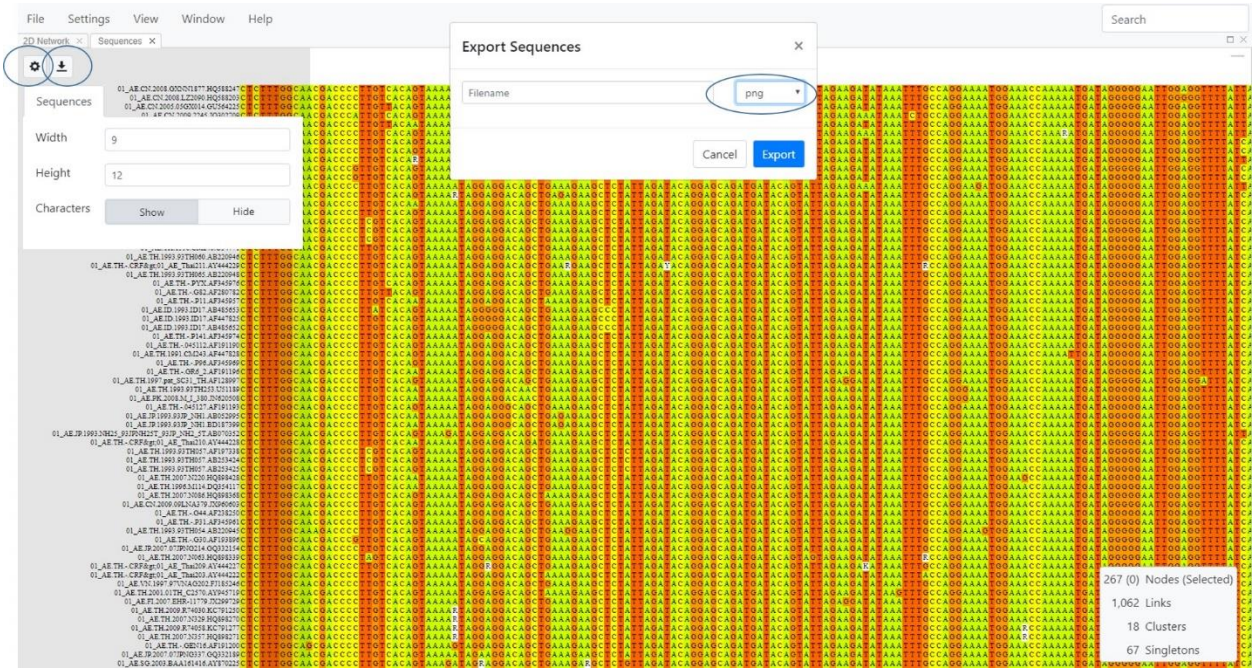


Fig. 97. Sequence View showing the aligned nucleotide sequences. Dialog boxes allow you to change the alignment appearance and to export the alignment.

Phylogeny View

The Phylogeny View is used to generate a phylogenetic tree using the sequences in your FASTA file and the genetic distance-based neighbor-joining (NJ) method. Like genetic distance networks, a phylogenetic tree represents the evolutionary relationships among a set of sequences from a group of organisms. For example, phylogenetic analysis has shown that HIV-1 is composed of four main groups, M, N, O, and P. Group M is the most common group that has spread globally. Group M is further divided into multiple subtypes which represent closely related but distinct virus genotypes. Subtype B is the most common subtype in the United States. In a rectangular phylogenetic tree (Fig. 93), the horizontal lines are called branches or tips of the tree and represent each taxon or descendant, such as a descendant in a “family tree”. The nodes on the tree represent the inferred common ancestor for one or more taxa (plural of taxon). Clusters of sequences or taxa are called clades, which represent closely genetically related sequences. A clade is a group of taxa that includes an ancestor and all descendants or taxa of that ancestor. For example, HIV-1 subtype B is considered a clade consisting of all subtype B sequences.

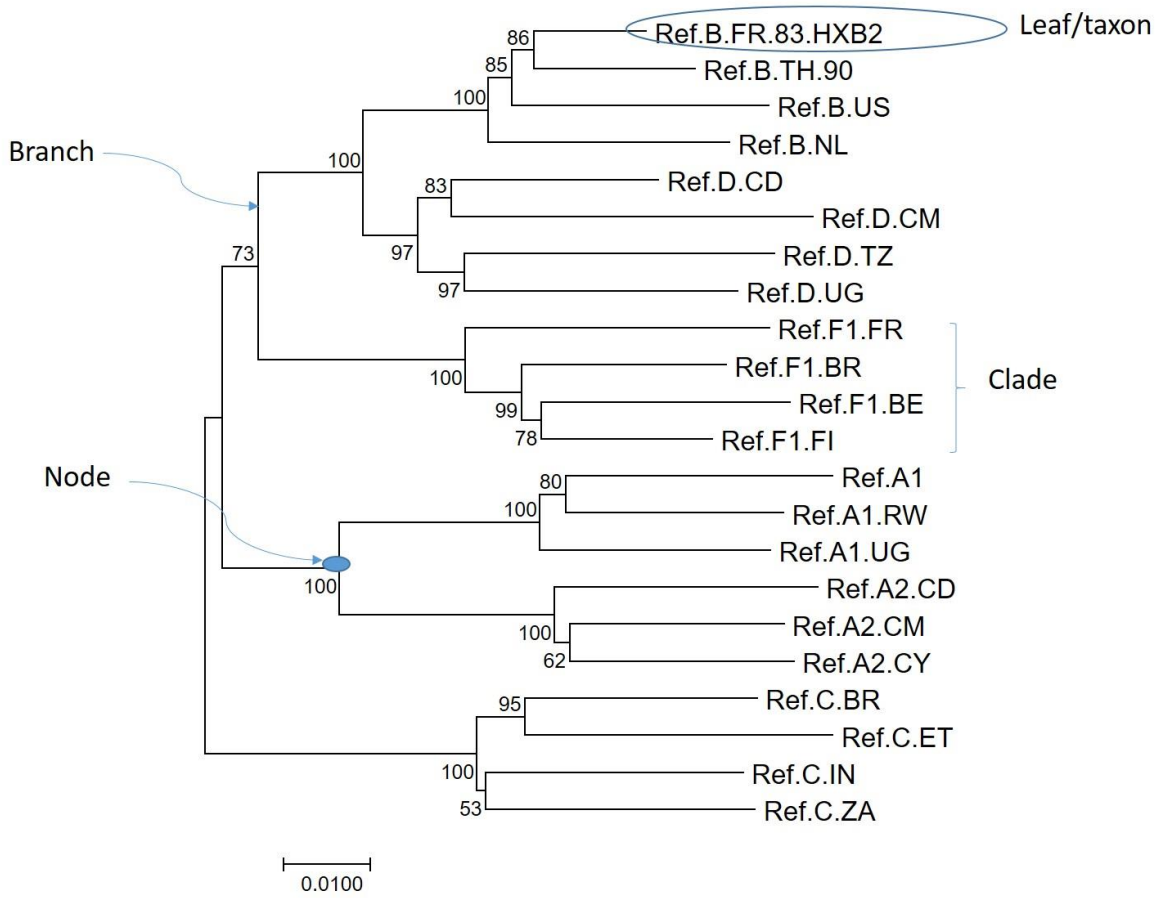


Fig. 98. Example of a rectangular neighbor-joining tree with components labeled

Phylogenetic trees can be rooted or unrooted. Rooted trees provide information about the order of nodes in the tree. The root of the tree is the oldest ancestral lineage of the dataset examined. Unrooted trees show the relationships of the taxa without making assumptions about ancestry. The NJ method used for Phylogeny View infers a rooted tree.

The length of the horizontal branch in a tree is directly proportional to the amount of genetic change in your dataset. The scale bar in the Phylogeny View provides the number of nucleotide substitutions/site in the dataset for the branch lengths in the inferred tree. The vertical lines have no meaning but are used to evenly display the taxa in the tree.

To display the phylogenetic tree, select **Phylogenetic Tree** from the **View** menu (Fig. 99).

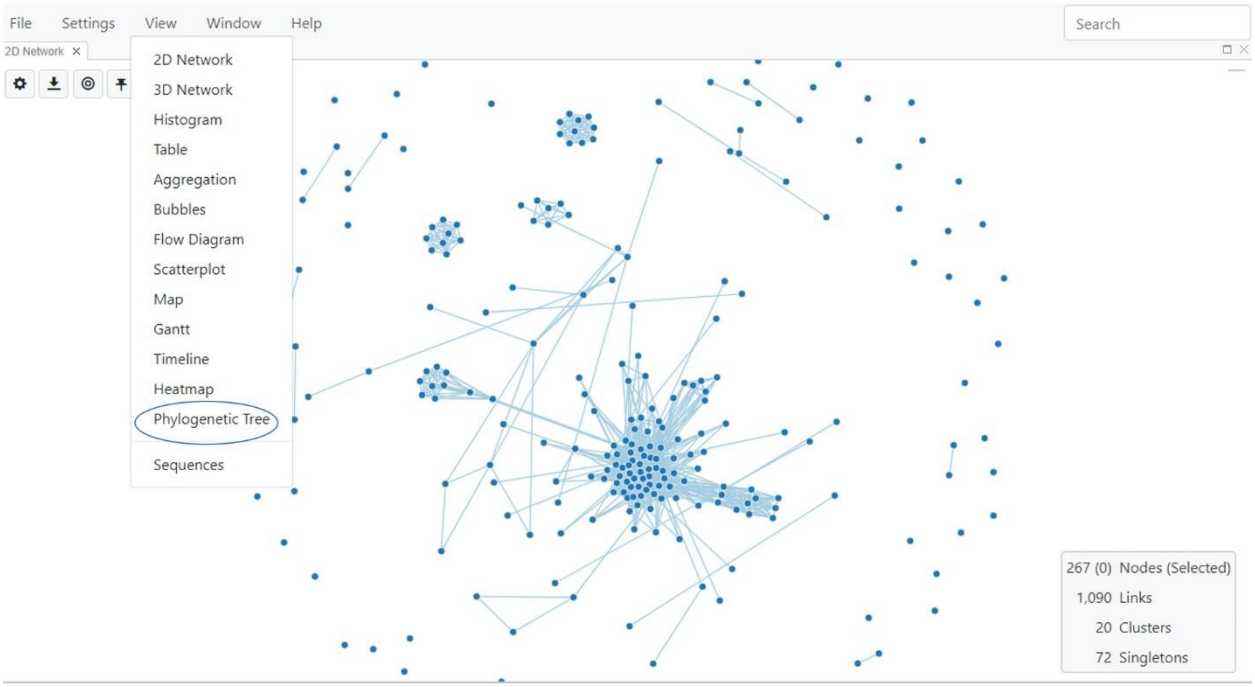


Fig. 99. Selecting Phylogenetic Tree View

The NJ tree will then be displayed in a new window (Fig. 100). You can use the mouse to zoom in and out and use the Center and Scale icon  to re-orient the tree to its default size and position.

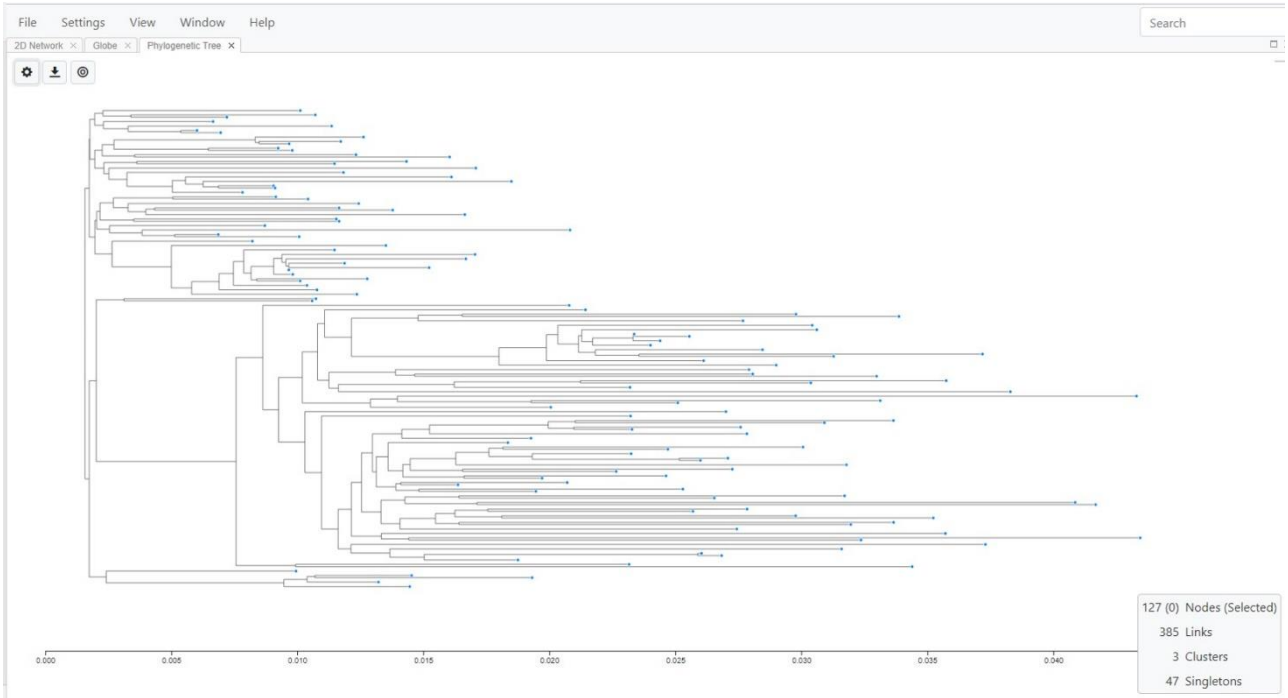


Fig. 100. Default neighbor joining Tree View

Phylogeny Settings

Tree settings can be customized using the Toggle Phylogeny Settings button  to access the menu. There are three tabs within this menu: Tree, Branches and Leaves (Fig. 101).

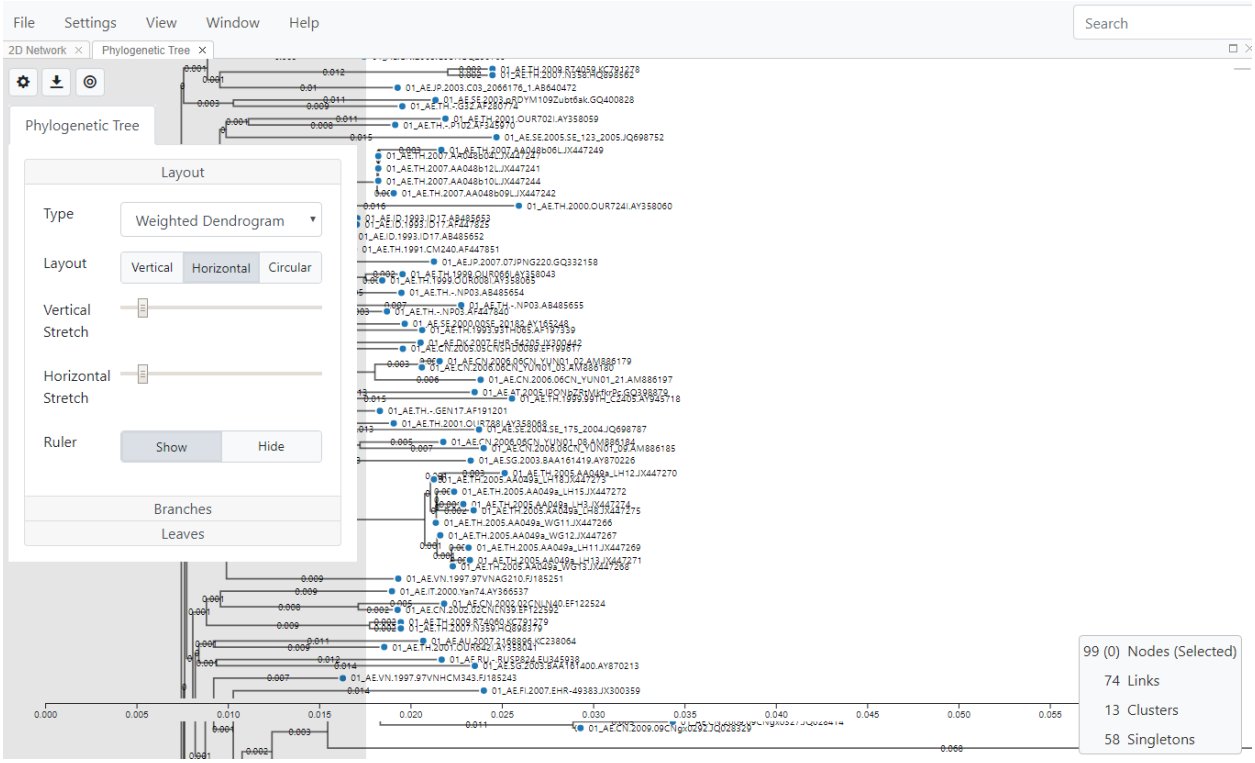


Fig. 101. Phylogenetic Tree View showing the settings menu tabs

Tree tab:

Modify tree settings using three tabs: Layout, Actions and Meta

Layout: Select the Layout tab to change the type of tree, layout, and whether you would like the ruler (scale bar) displayed. You can also stretch the tree vertically or horizontally. This feature is helpful when you have a large tree with many taxa. The stretch feature spreads the branches out and allows for greater clarity.

The default is a vertical weighted dendrogram, which is a tree with the branch lengths scaled. The length of the branch is proportional to the number of nucleotide substitutions (Fig. 102). A

dendrogram is a diagram representing a phylogenetic tree that shows how ancestors are related to descendants. A weighted dendrogram infers a rooted tree that reflects the structure present in the genetic distance matrix.

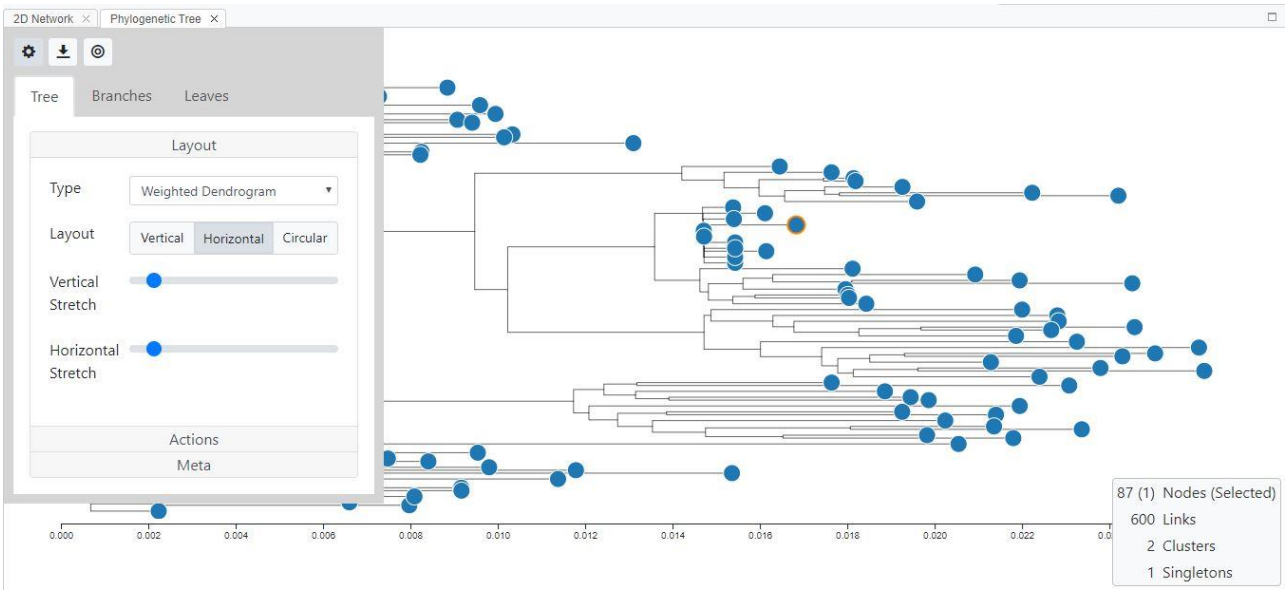


Fig. 102. Selecting a horizontal, weighted dendrogram with the scale bar displayed

If you select "**Tree**" from the dropdown menu, you will see that the branches are not scaled, i.e., branch lengths are even, and not proportional to the nucleotide substitutions (Fig. 103).

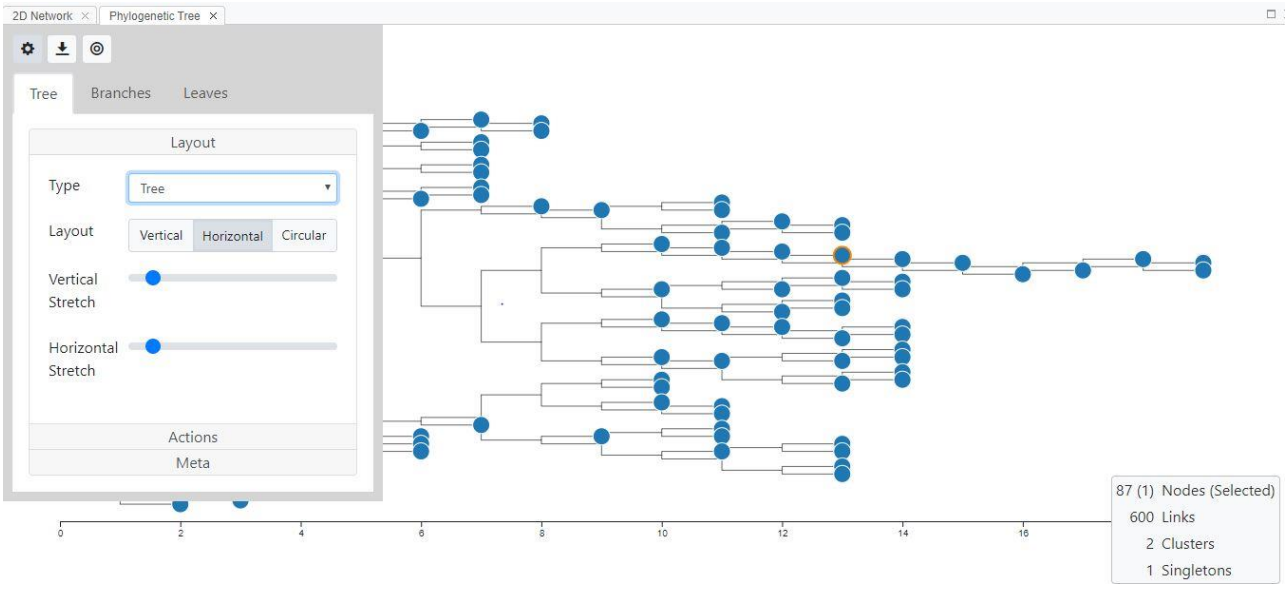


Fig. 103. Horizontal tree with the scale bar displayed

If you select "**Unweighted Dendrogram**" from the pull down menu, the branch lengths are not scaled but are rather the distances contribute to each branch length average that is computed and the length of the branches are adjusted so the taxa (or leaves as they are sometimes called) are aligned on the right (Fig. 104).

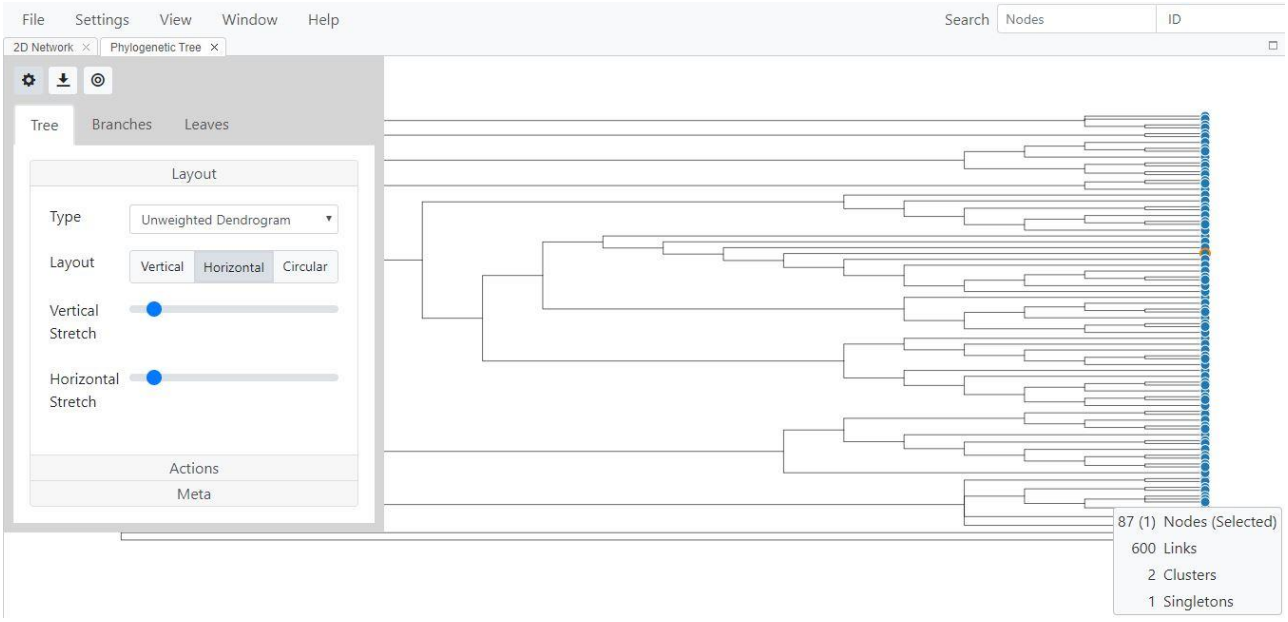


Fig. 104. Unweighted dendrogram

You can change the tree layout to either vertical (Fig. 105) or circular (Fig. 106) by selecting the respective tree Layout options.

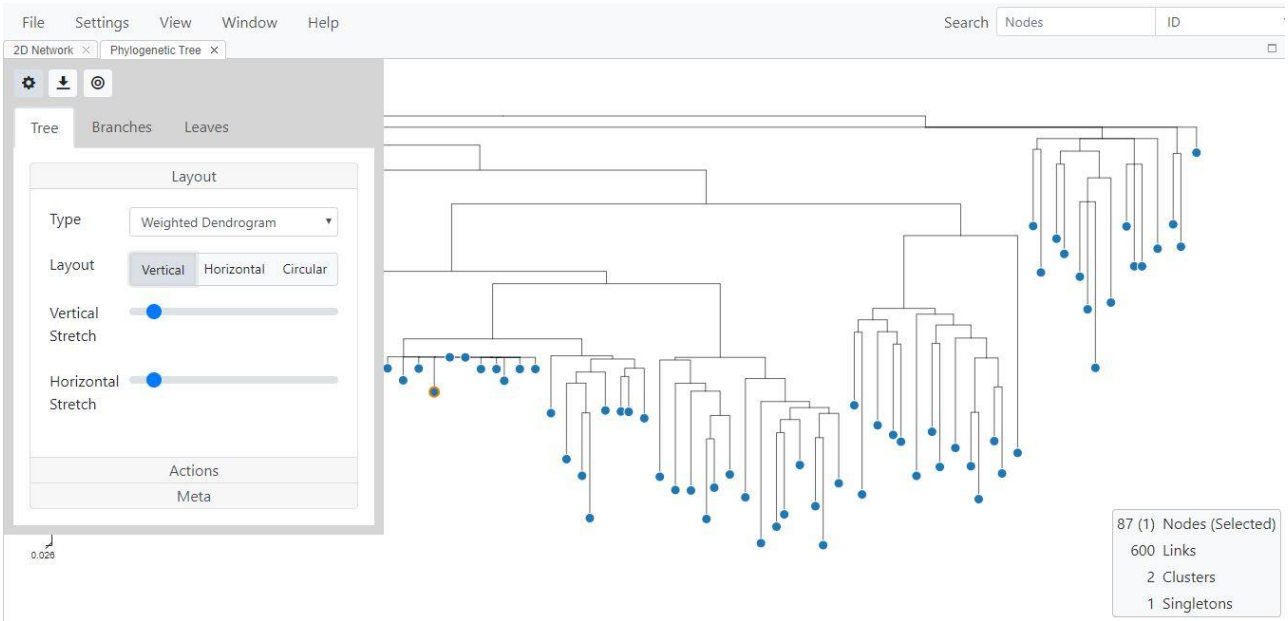


Fig. 105. Vertical weighted dendrogram

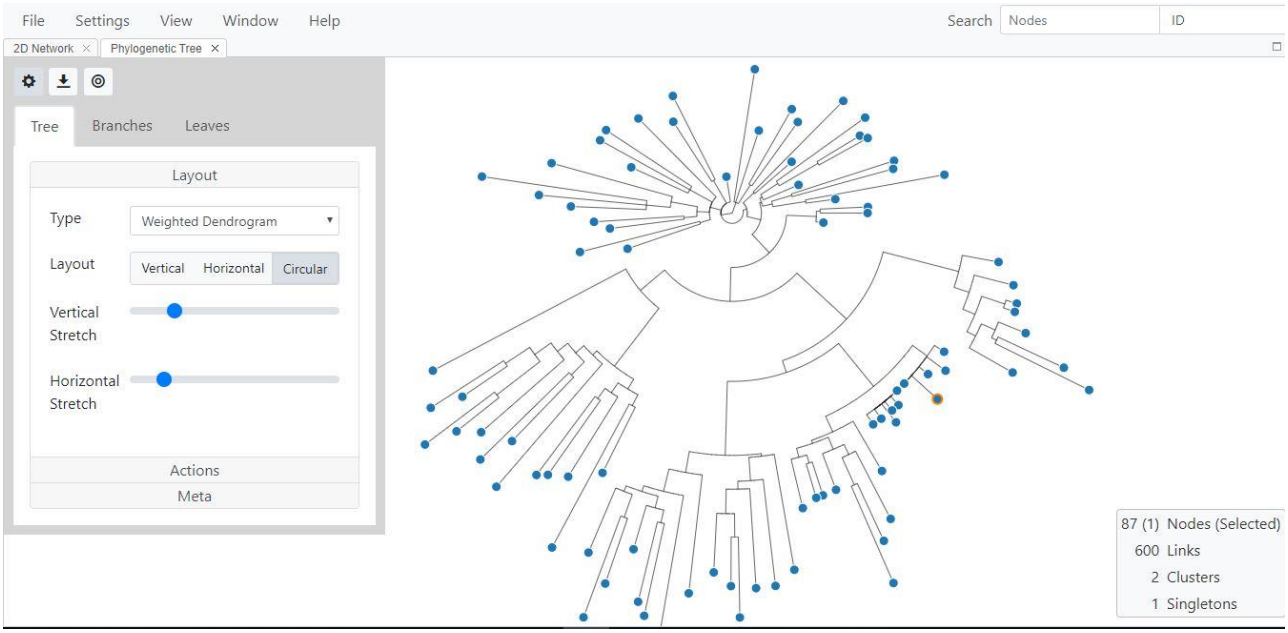


Fig. 106. Circular weighted dendrogram

Actions subtab:

This tab lets you modify the tree branches for improved visualization (Fig. 107).

Simplify: Collapses branches with single descendants into a long, continuous branch.

Consolidate: Collapses branches with zero distance into combined branches.

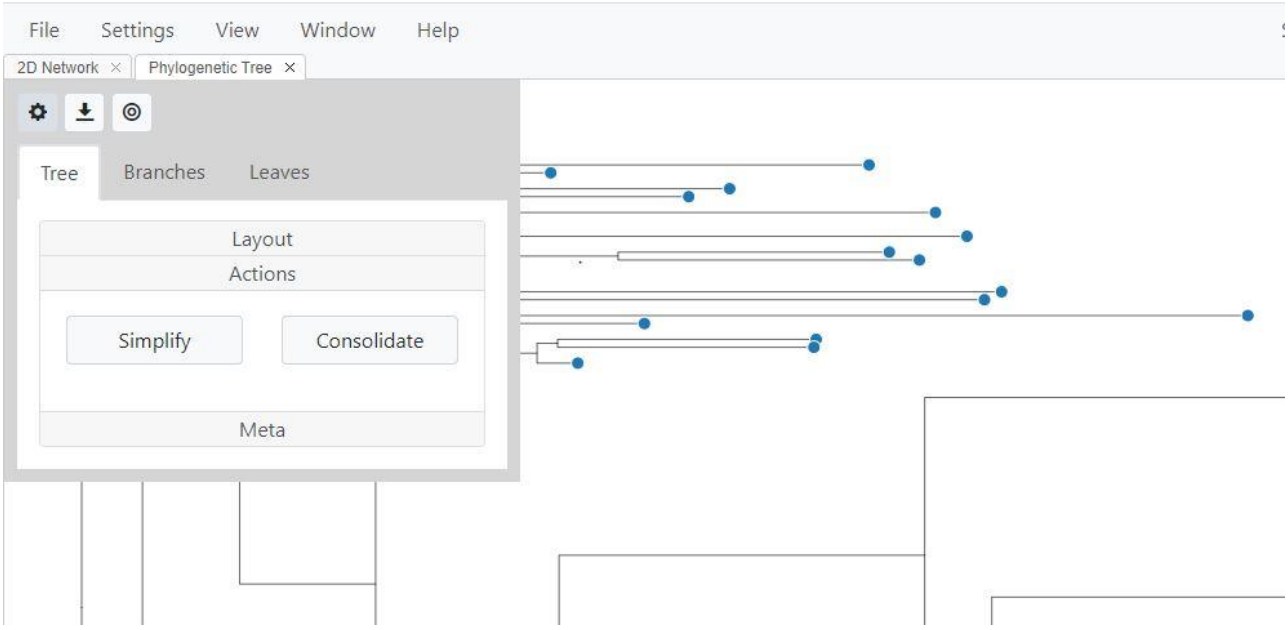


Fig. 107. The **Actions** tab allows you to modify how descendant branches are displayed.

***Meta*:**

Use the buttons (Fig. 108) to select ruler display and animation of tree configuration changes.

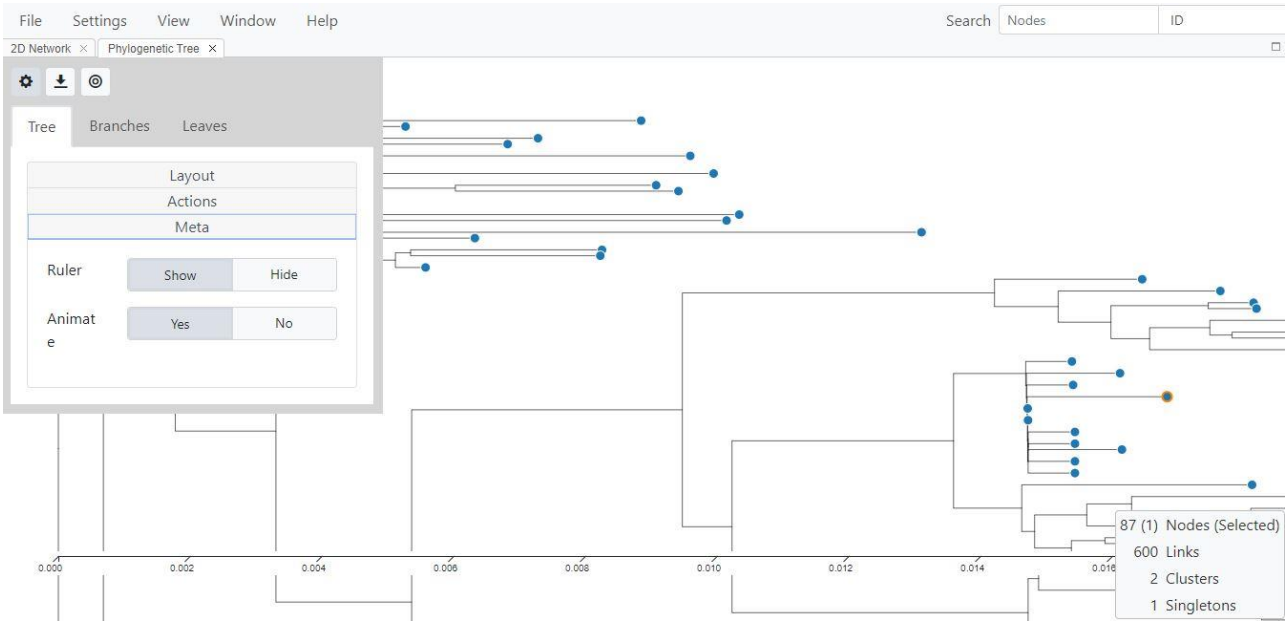


Fig. 108. Meta tab allows you to change ruler display and animation

Branches tab:

This option allows appearance of the branches to change to square, smooth or straight (Fig. 109). You can also choose whether to show or hide nodes as well as branch lengths. These two parameters are hidden by default.

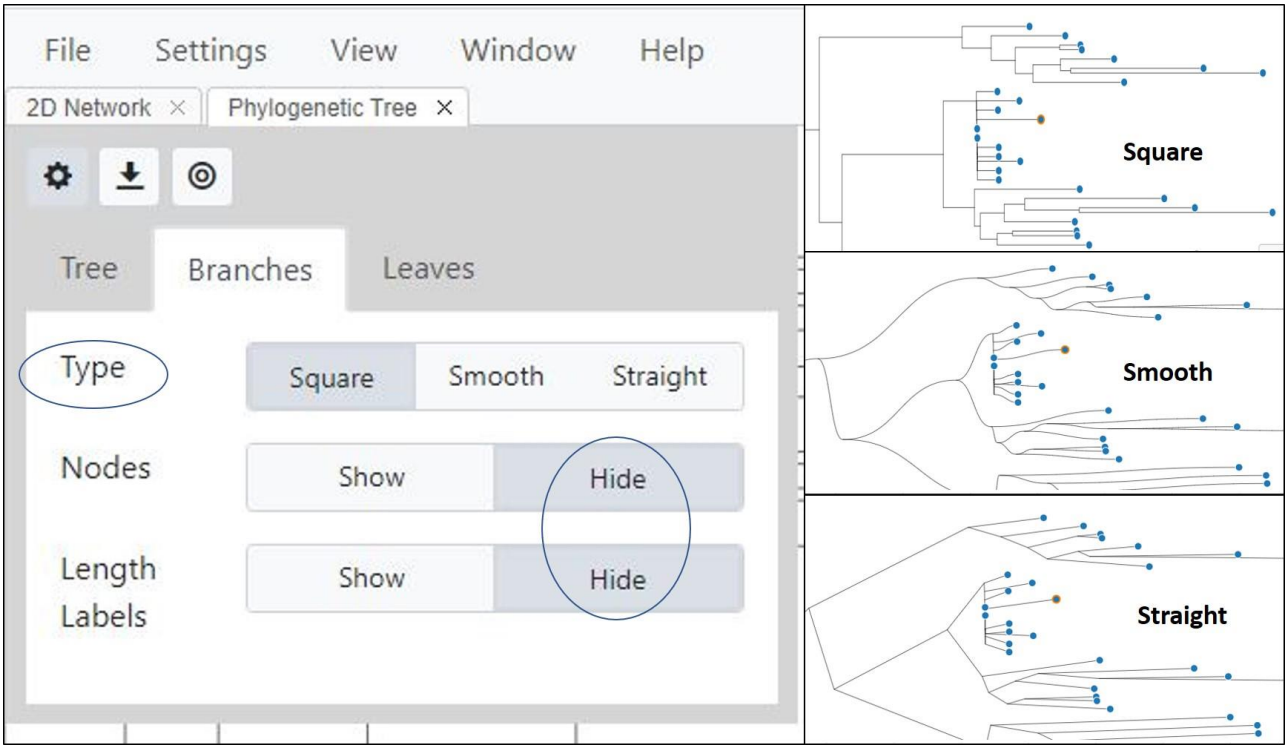


Fig.109. Branch tab options

Leaves tab:

Labels and Tooltips: This option allows you to select whether to display labels and tooltips, and lets you change label sizes (Fig. 110).

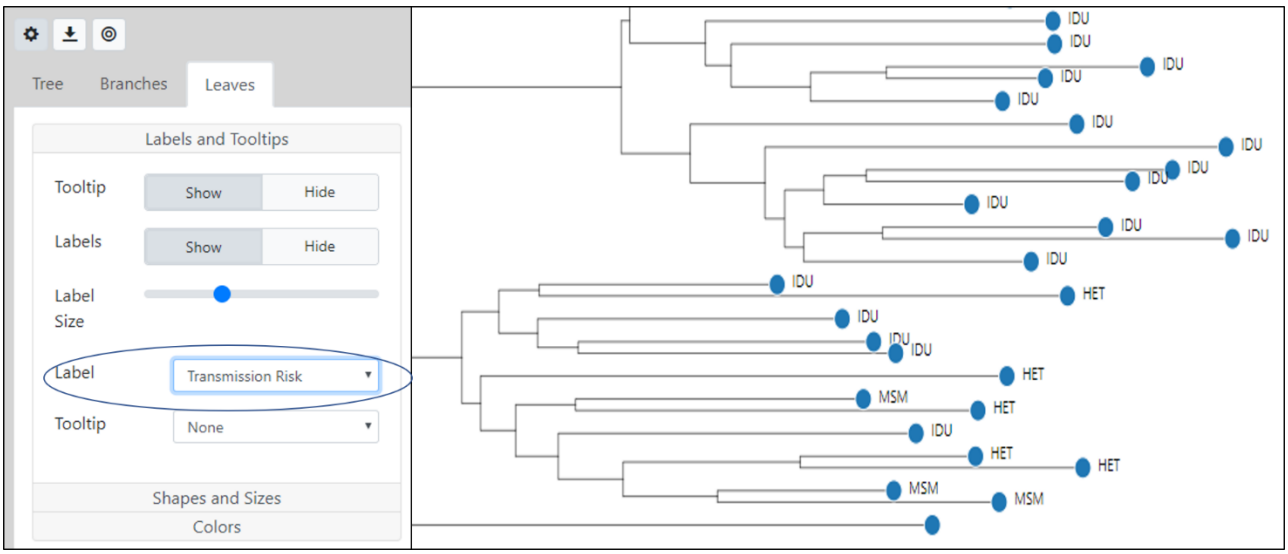


Fig. 110. Leaves tab, Labels and Tooltips. Leaf nodes are labelled with transmission risk.

Shapes and Sizes: This tab allows you to size nodes by any variable in your node list, available from the dropdown menu (Fig. 111).

Colors: Takes you to the styling tab of Global Settings where you can color by the variable of your choice.

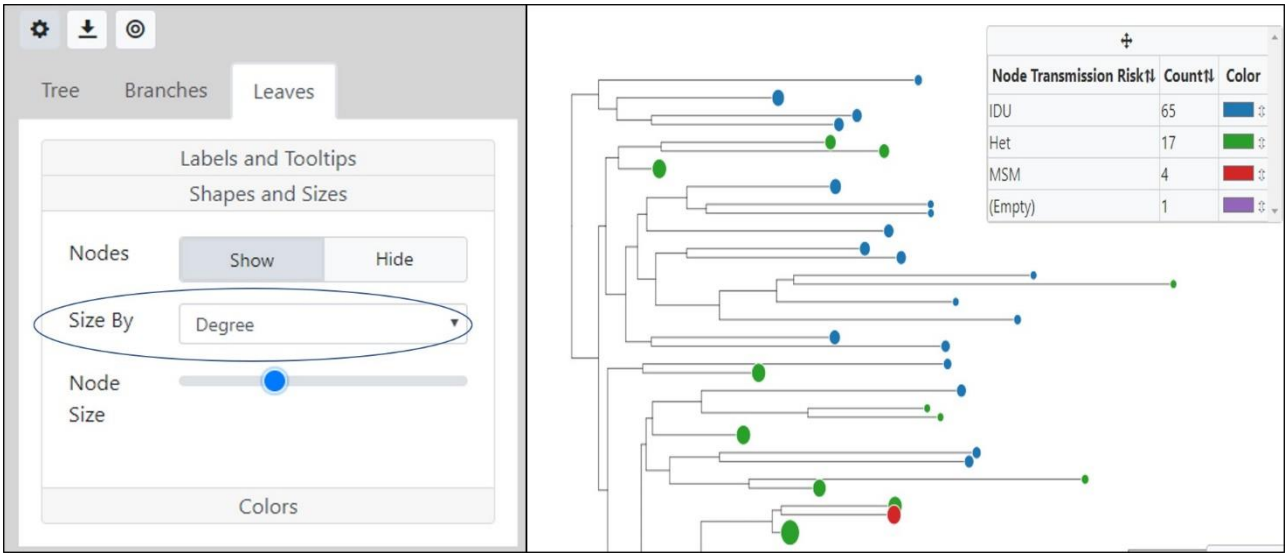


Fig. 111. Nodes are sized by degree and colored by transmission risk; node labels are hidden

Colors: Takes you to the styling tab of Global Settings where you can color by the variable of your choice. A search box on the top right of the Phylogenetic View allows the tree to be searched by

taxon (sequence) name or ID. You can search the tree for any variable available in the dropdown menu. The taxon name is then highlighted in blue on the tree. This feature is useful if you have many taxa in a large tree and need to locate an individual taxon.

Node options

Right-click on a node in the tree to see various options available to you, including rotate, re-rooting, and removing a branch. See the figures below for examples of these features.

Rotate: This option is used to rotate a branch along the axis of a selected node in a tree.

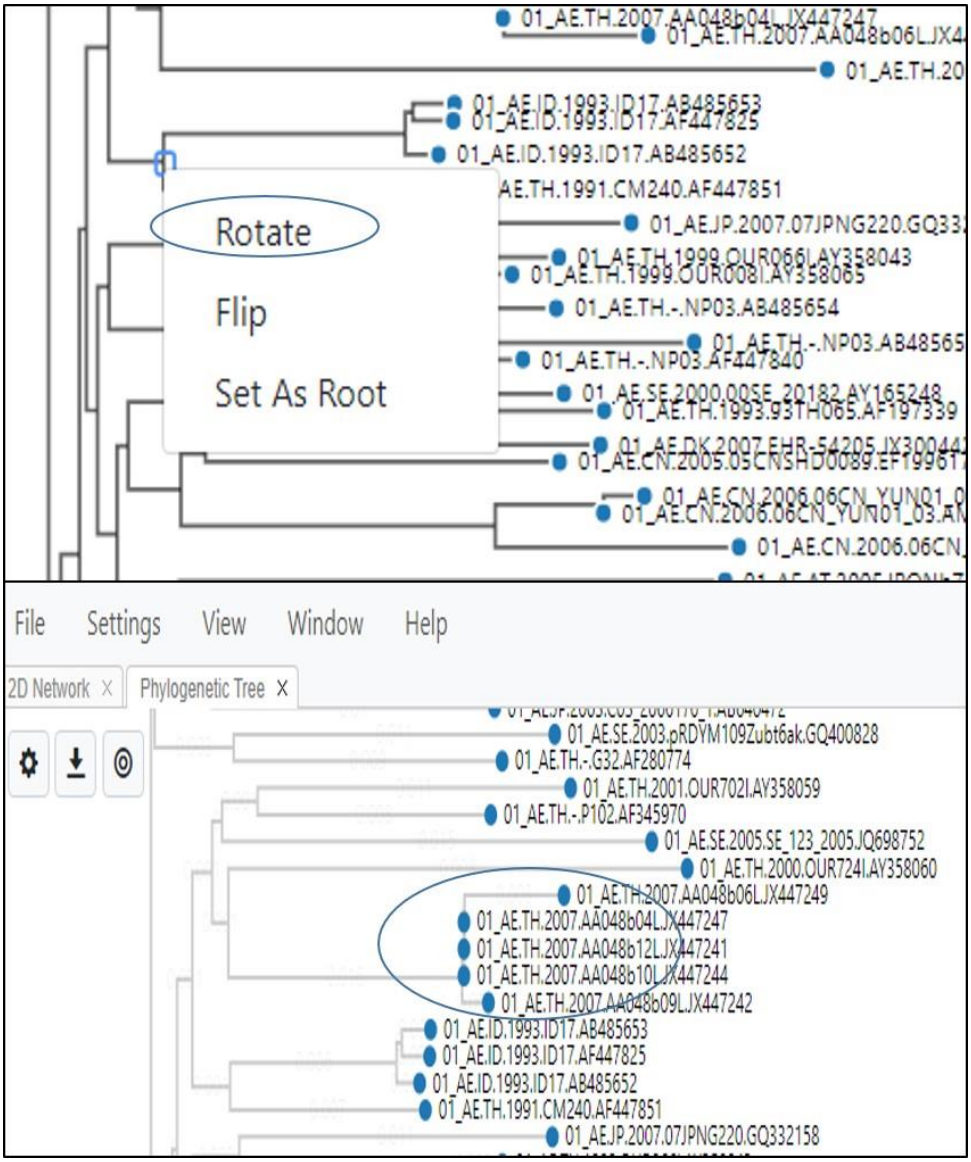


Fig. 112. Phylogenetic Tree View showing how to rotate a clade (cluster of sequences) along the axis of the selected node.

Flip: This flips the clade along the axis of the node, so a clade which was at the bottom will now be flipped to be at the top.

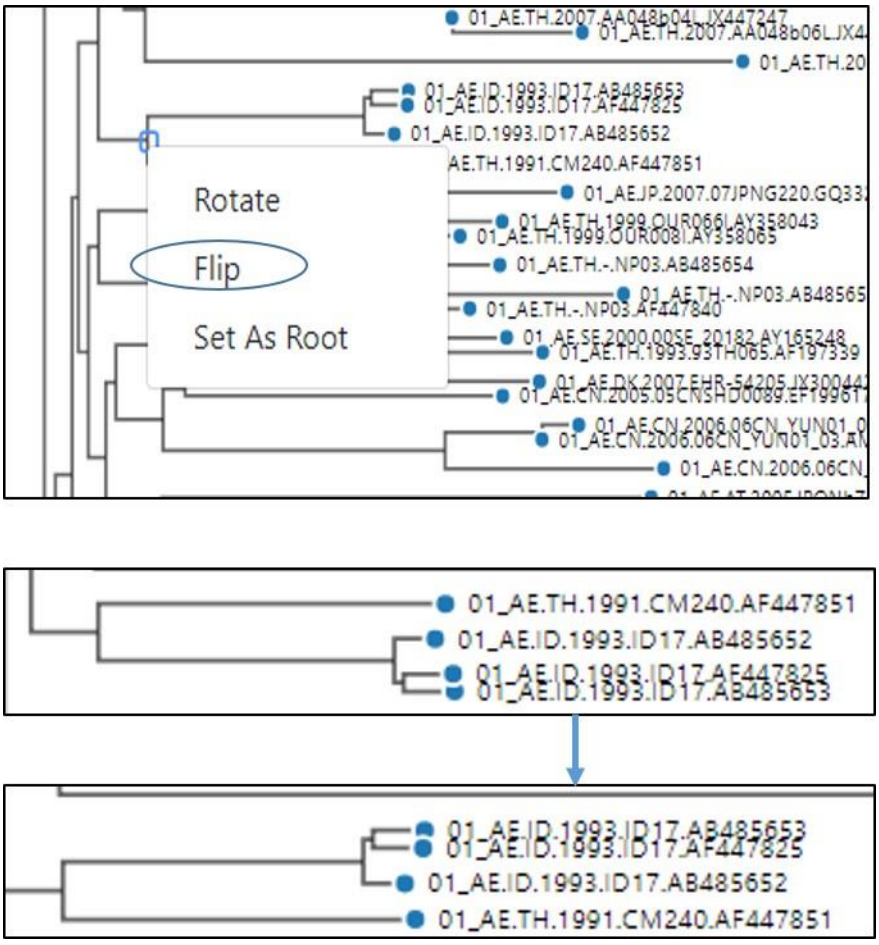


Fig. 112. Phylogenetic Tree View showing how to flip a clade (cluster of sequences) along the axis of the selected node.

Set As Root: This option is only used if you have prior information about the potential evolutionary history of the taxa or if an outgroup was included in the analysis. The outgroup is used to "root" the tree. An outgroup is a set of taxa that are close but distinct from the taxa you wish to analyze. For example, HIV-1 subtype J could be used as an outgroup for an analysis of HIV-1 subtype CRF01_AE sequences (Fig. 113). Please note that there is no undo option for re-rooting. If you re-

root on a node and would like to go back to the original tree structure, please close the window, select **Phylogenetic Tree** from the **View** menu to display the original tree.

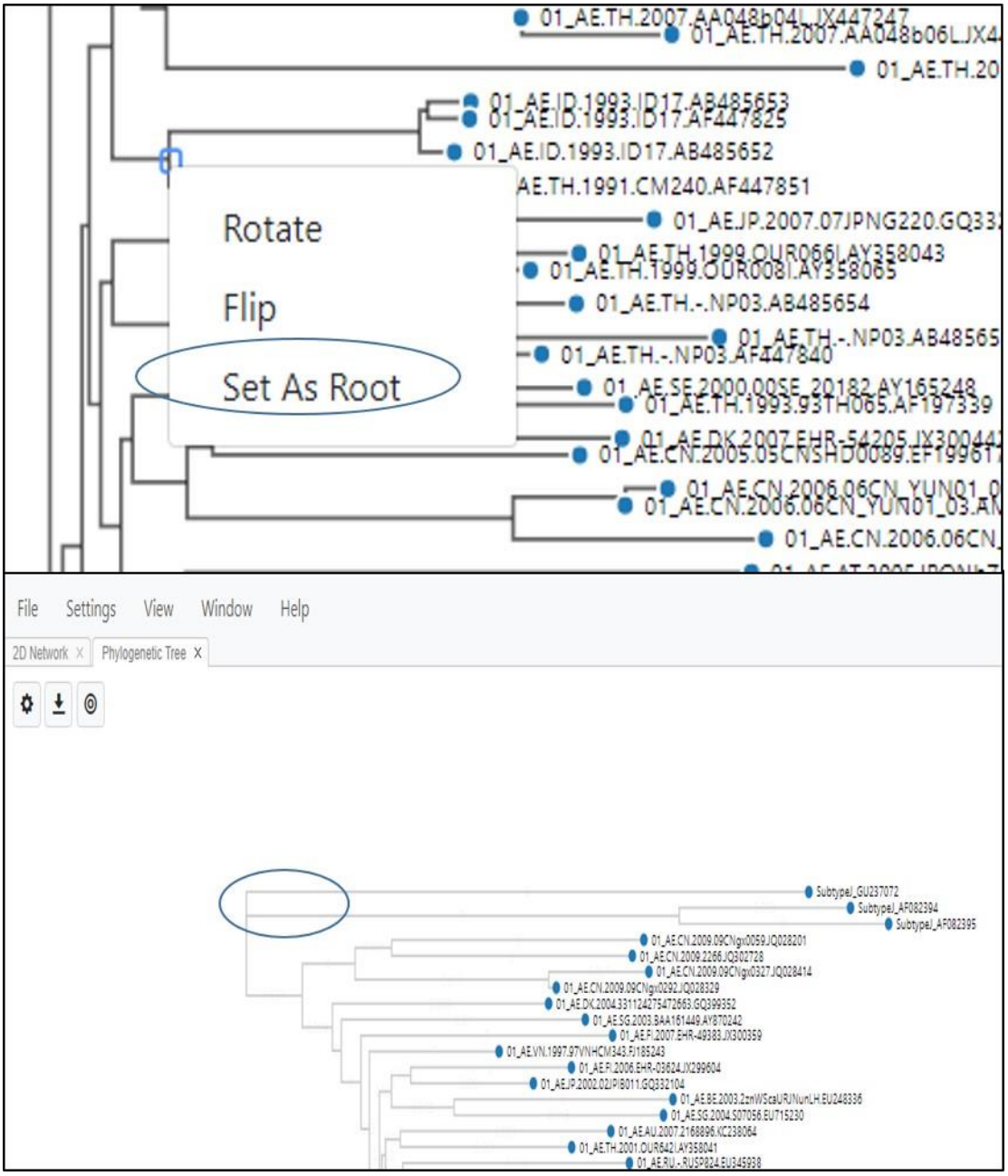


Fig. 113. Phylogenetic Tree View showing how to re-root a tree at a selected node.

Troubleshooting

If you are having trouble with loading a data file, if you observe unusual behavior, or if MicrobeTrace freezes or hangs, you may need to clear the browser cache using the following steps.

In Chrome, open developer tools (Fig. 114) by clicking on the three vertical dots; select **More** tools, then Developer tools.

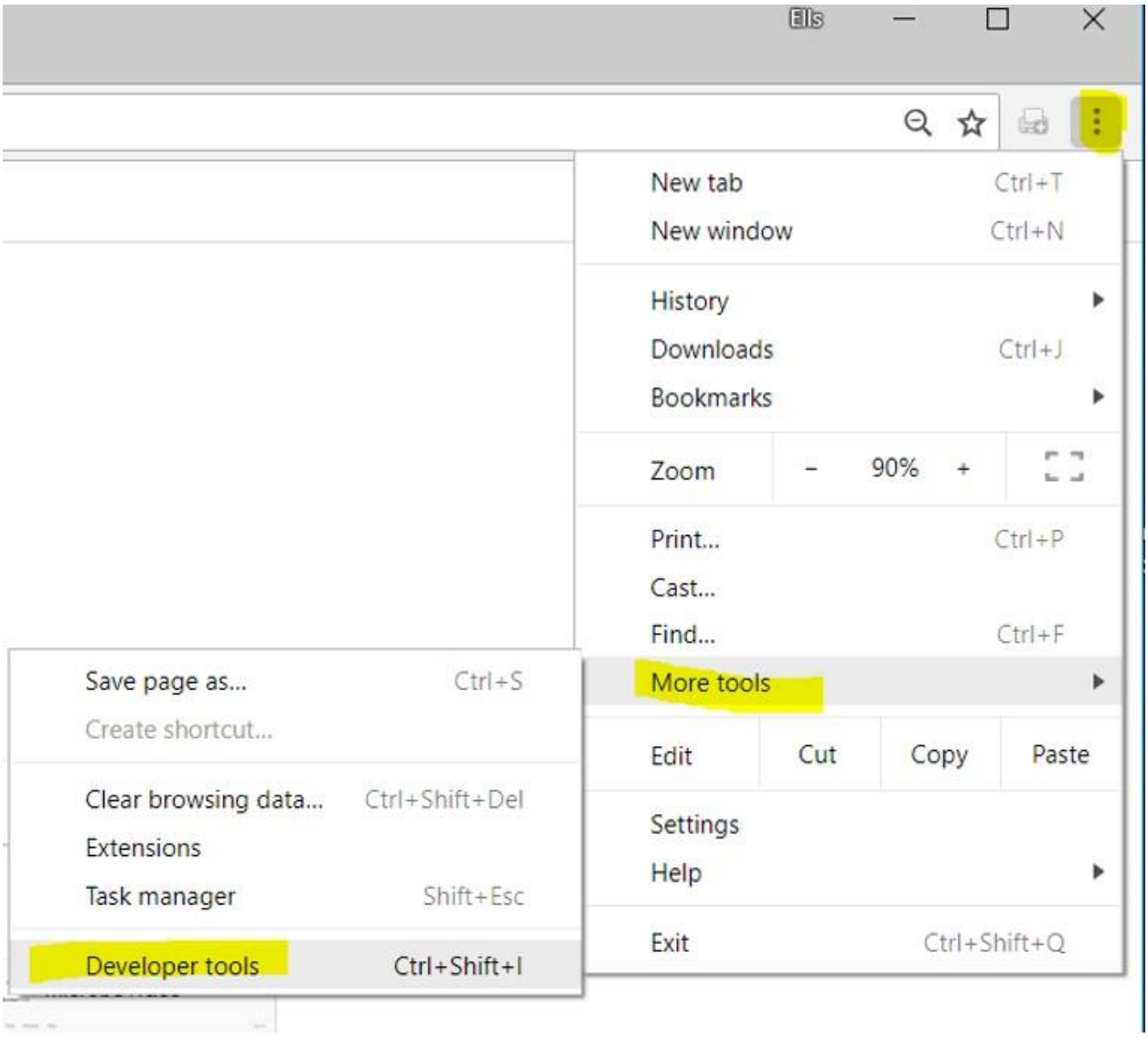


Fig. 114. Clearing the browser cache in the Developer tools window

Select the >> (Fig. 115) and then select Application from the resulting pull-down menu.

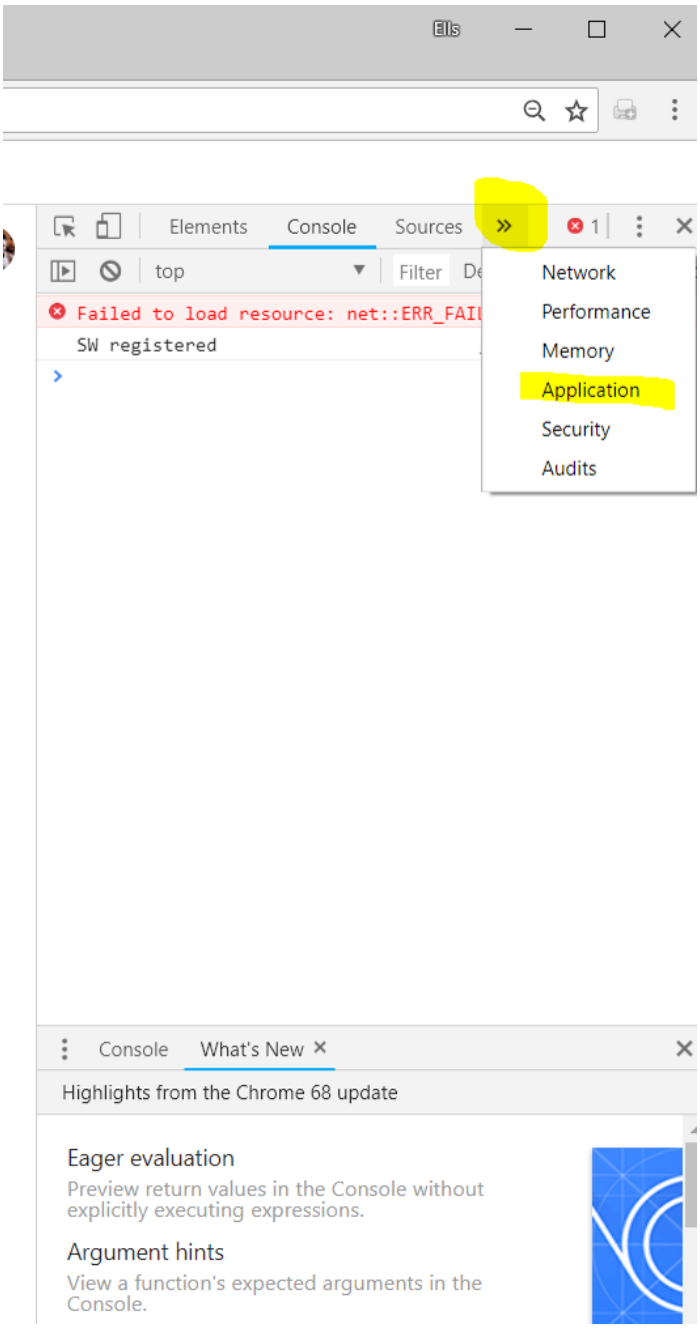


Fig. 115. Clearing the browser cache via the Application option

Once you are in the application tab, select Clear Storage (Fig. 116)

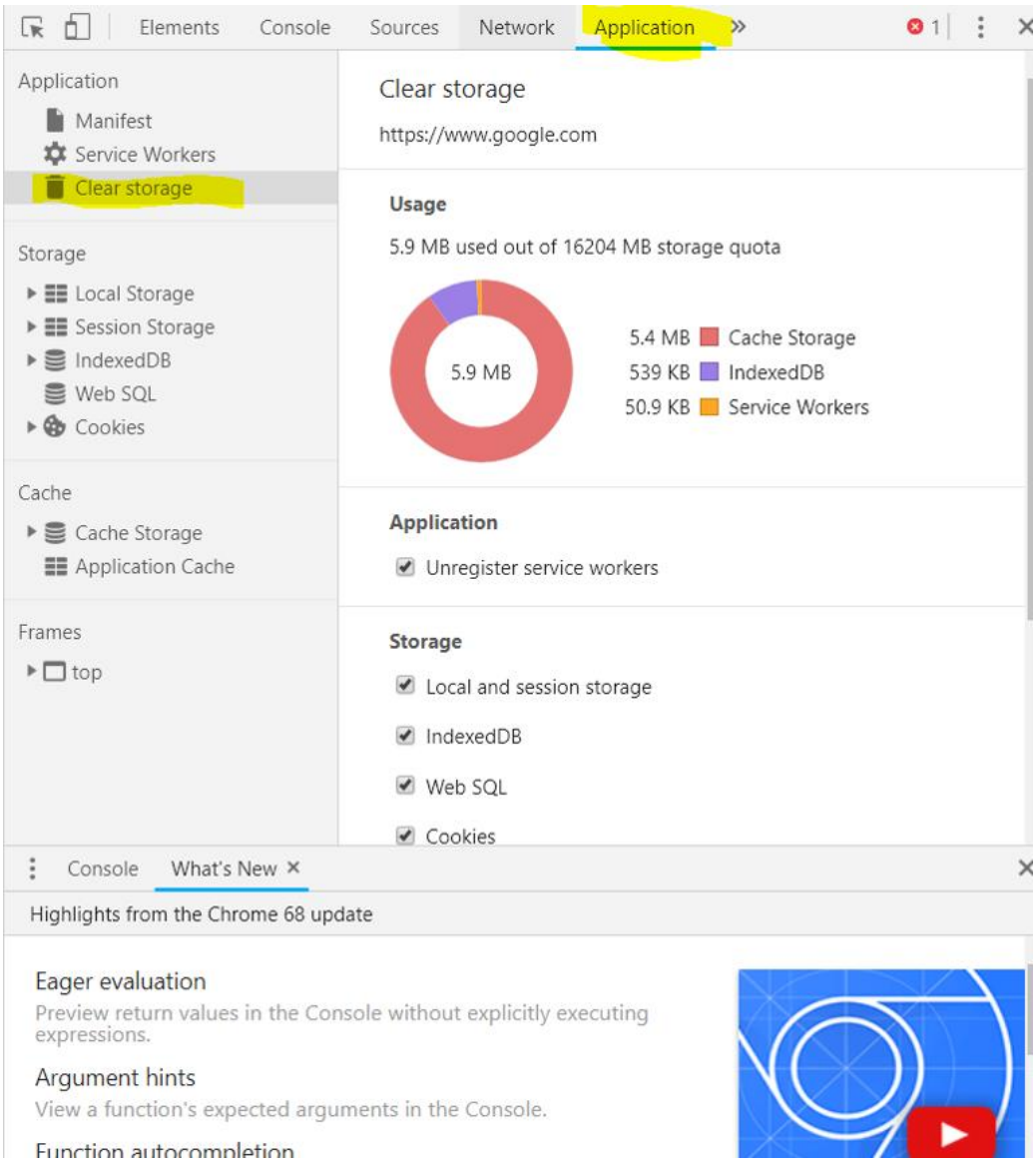


Fig. 116. Clearing browser cache using the Clear storage option

Finally, scroll to the bottom of this subtab, and select Clear Site Data (Fig. 117).

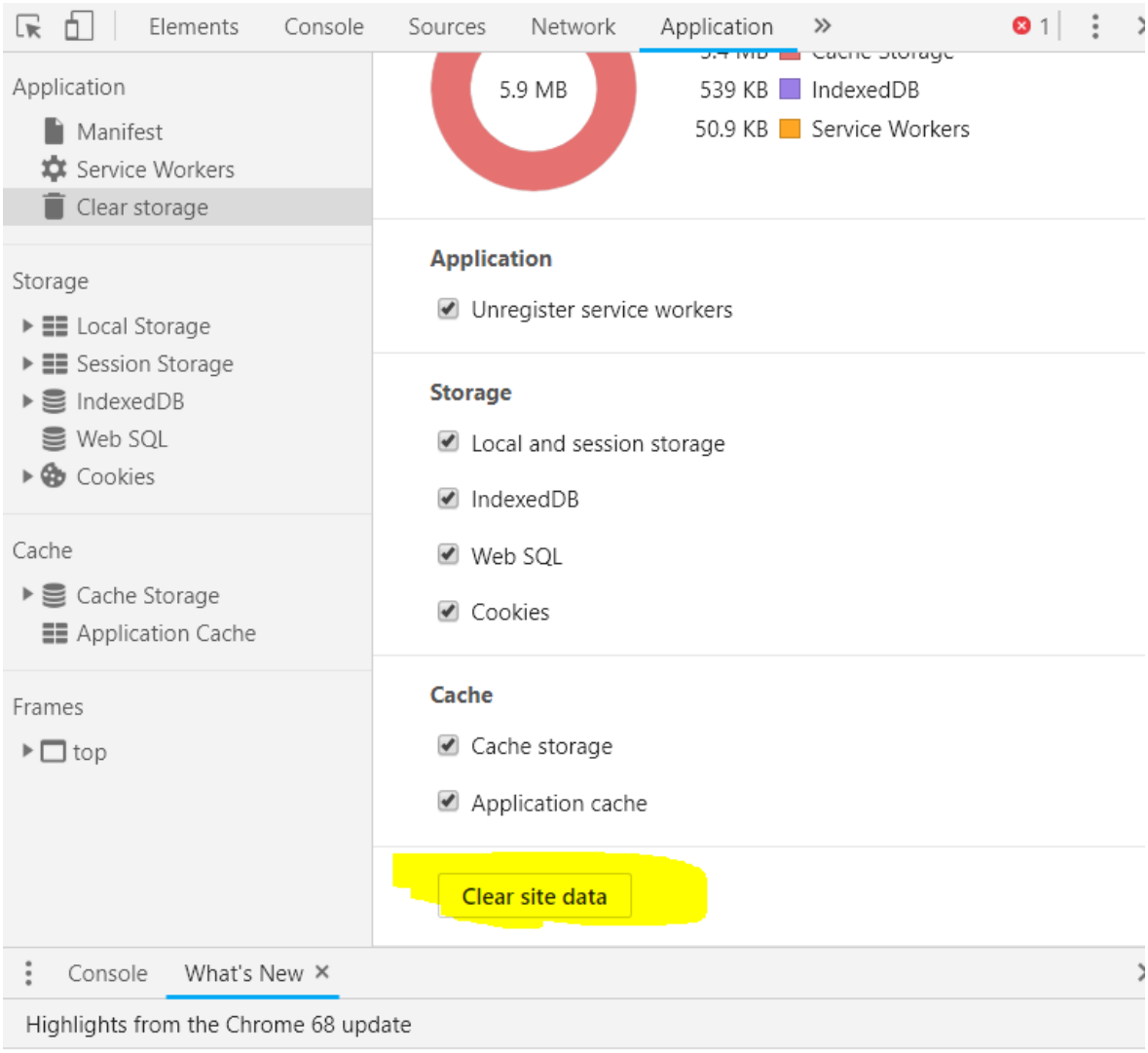


Fig. 117. Clearing the browser cache to clear the site data

This series of steps will clear the cache. If clearing the browser cache does not resolve your issue, you can use the MicrobeTrace GitHub site to report and track any software issues. To report any MicrobeTrace issues, use the link to go to the MicrobeTrace "new issues" website https://github.com/CDCgov/MicrobeTRACE/issues/new. Access to this site requires a GitHub account. If you do not have a GitHub account, then you will be prompted to create one. GitHub accounts are free and access requires only an email address and a self-generated password.

On the MicrobeTrace issues website you will be prompted with a form requesting a title and description of the issue (Fig 118).

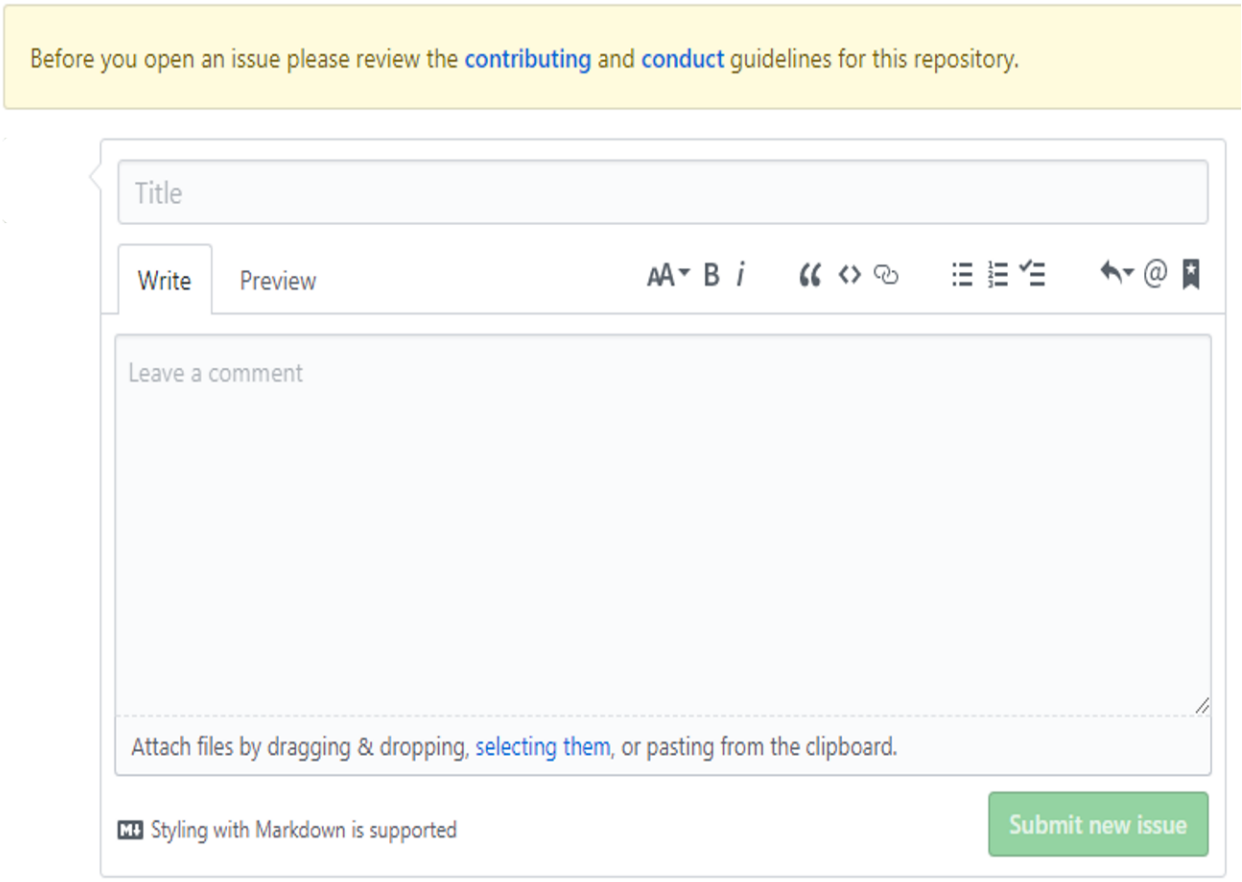


Fig. 118. Reporting issues to the MicrobeTrace team

In the title, please include what View in the software program you were using. In the comment section, please provide as much detail about the issue as possible, including the operating system used. You can also use this form to let us know how we can improve the software.

For example, Issue #27 is titled “The Group key table won't scroll past the bottom of the page”. While being clear about how the bug violates user expectations, this statement does not describe what the “Group Key Table” is or where it can be found when encountered using MicrobeTrace. In contrast, issue #125 “Option to Switch Distance Matrix from TN93s to SNPs” describes where the problem was encountered (in the Distance Matrix) and what is desired (the option to switch between the TN93 nucleotide substitution model for the analysis, and SNPs, which is available but unused).

You can also send an e-mail to request help from a CDC MicrobeTrace support representative at microbetrace@cdc.gov. When drafting your email, please be as thorough as

possible, listing out every action you took in the program prior to encountering the problem. The more detail you can provide, the greater the likelihood that we will be able to resolve the issue.

References

1. Ellsworth M.Campbell, Anthony Boyles, Anupama Shankar, Jay Kim, Sergey Knyazev, William M. Switzer. MicrobeTrace: Re-tooling Molecular Epidemiology for Rapid Public Health Response, PLOS Computational Genomics. In press. Pre-print is here: https://www.biorxiv.org/content/10.1101/2020.07.22.216275v1
2. Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Mol Biol Evol.* 2018 Jul 1;35(7).
3. Guy Yachdav, Sebastian Wilzbach, Benedikt Rauscher, Robert Sheridan, Ian Sillitoe, James Procter, Suzanna E. Lewis, Burkhard Rost, Tatyana Goldberg; MSAViewer: interactive JavaScript visualization of multiple sequence alignments, *Bioinformatics*, Volume 32, Issue 22
4. Hightower GK, May SJ, Perez-Santiago J, Pacold ME, Wagner GA, Little SJ, et al. HIV-1 clade B pol evolution following primary infection. *PLoS ONE*. 2013; 8(6):e68188. doi: 10.1371/journal.pone.0068188.
5. Oster AM, France AM, Mermin J. Molecular Epidemiology and the Transformation of HIV Prevention. *JAMA.* 2018;319(16):1657–1658. doi:10.1001/jama.2018.
6. Barré-Sinoussi F, Abdool Karim SS, Albert J, Bekker LG, Beyrer C, Cahn P, Calmy A, Grinsztejn B, Grulich A, Kamarulzaman A, Kumarasamy N, Loutfy MR, El Filali KM, Mboup S, Montaner JS, Munderi P, Pokrovsky V, Vandamme AM, Young B, Godfrey-Faussett P. Expert consensus statement on the science of HIV in the context of criminal law. *J Int AIDS Soc*. 2018;21(7):e25161. doi: 10.1002/jia2.25161.

Acknowledgments

The MicrobeTrace team consists of Tony Boyles, Ellsworth Campbell, Anupama Shankar, Jay Kim, Evan Moscoso, Francis Ambrosio, Roxana Cintron, Sergey Kniazev, and Bill Switzer. We thank the developers of HIV-TRACE for their pioneering work in HIV transmission research and

software development, including Joel Wertheim, Sergei Kosakovsky Pond, and Steven Weaver. We also thank the developers of all the open source projects on which MicrobeTrace depends. Finally, we thank the numerous beta-testers for their input. Special thanks to the CDC's Division of Tuberculosis Elimination epidemiology team (Sarah Talarico, Kathryn Winglee, Ben Silk) who gave us extremely valuable input that led to development of many new and improved features. We are very thankful for the CDC's Advanced Molecular Detection Initiative and the Division of HIV Prevention for funding development of MicrobeTrace.