
PHYLOViZ Documentation

Release 2.0

PHYLOViZ Team

September 07, 2016

1	Download and install	3
1.1	Binaries	3
1.2	Source	4
2	Loading data	5
2.1	File formats	5
2.2	Loading a Dataset	6
2.3	Loading a remote Dataset	10
3	Data analysis	15
3.1	goeBURST algorithm	15
3.2	goeBURST Full MST algorithm	19
3.3	Hierarchical Clustering	22
3.4	Neighbor Joining	26
4	Display and visualization	31
4.1	Interface features	31
4.2	Color conventions	39
5	Querying and visualizing the data	41
5.1	The isolate data tab	41
5.2	The typing data tab	42
5.3	Regular expression primer	42
5.4	Queries using the table view	43
5.5	Queries using the tree view	46
5.6	Exporting the results to an image file	49
6	Project management	51
6.1	Saving	51
6.2	Loading	52

PHYLOViZ is a platform independent JAVA software that allows the analysis of sequence-based typing methods that generate allelic profiles and their associated epidemiological data.

Download and install

PHYLOViZ core and several plugins are free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.

Certain source files distributed by the PHYLOViZ Team are under the terms of the GNU General Public License Version 3 with the following clarification and special exception, but only where PHYLOViZ Team has expressly included it in the particular source file's header.

```
Linking this library statically or dynamically with other modules is
making a combined work based on this library. Thus, the terms and
conditions of the GNU General Public License cover the whole combination.
```

```
As a special exception, the copyright holders of this library give you
permission to link this library with independent modules to produce an
executable, regardless of the license terms of these independent modules,
and to copy and distribute the resulting executable under terms of your
choice, provided that you also meet, for each linked independent module,
the terms and conditions of the license of that module. An independent
module is a module which is not derived from or based on this library.
If you modify this library, you may extend this exception to your version
of the library, but you are not obligated to do so. If you do not wish
to do so, delete this exception statement from your version.
```

Code licensed under this license may be reused in commercial products provided that changes made directly in the sources - bug fixes or enhancements - must be contributed back to PHYLOViZ, but new source files (as in new plugins) which you write that link to PHYLOViZ code do not need to be.

Choose the appropriate version for your operating system or the .jar file. The OS specific versions already contain some memory specific parameters to enhance the software performance when using large datasets.

See details about available plugins and the licenses under which they are covered.

1.1 Binaries

A cross-platform zip distribution package is [available](#).

Just unzip the package, enter the created directory and in the sub-directory `bin/` run `phyloviz.exe` or `phyloviz64.exe` (Windows) or `'phyloviz'` (Linux/MacOS) accordingly to your operating system.

NOTE: You may need to adjust some parameters in `etc/phyloviz.conf` with respect to memory usage. These settings have a strong impact on visualization features. For instance, in Windows, you may achieve better results with:

```
default_options="--branding phyloviz -J-Xss8M -J-Xms32m -J-Xmx1024M --laf javax.swing.plaf.metal.Met
```

IMPORTANT NOTICE: After installing always go to the “Help” menu and “Check for updates” to install any novel plugins or latest updates to PHYLOViZ software. The SNP analysis plugin is installed in this way to demonstrate the plugin capability.

1.2 Source

All the Source code is available in the new code repository for in [bitbucket.org](https://bitbucket.org/phyloviz/phyloviz-main). Check it out at <https://bitbucket.org/phyloviz/phyloviz-main>.

PHYLOViZ is built on top of the NetBeans Platform, thus we recommend NetBeans for the development of new plugins.

Loading data

2.1 File formats

To be able to analyze and visualize your data, PHYLOViZ needs two separate files: One file contains the allelic profile data of the method you are using (Typing Data), while the other will contain accessory data (Isolate Data). In the example image below they are *sampleAPfile.txt* and *sampleADfile.txt* respectively.

ST	gki	qtr	murI	mutS	recP	ygt	ygiZ
1	10	6	6	6	12	13	8
2	5	4	4	1	2	15	2
3	5	3	4	1	6	2	1
4	2	2	4	1	8	7	2
5	2	2	4	1	12	12	7
6	1	3	1	1	1	1	4
7	1	1	1	1	1	1	3
8	1	1	1	1	1	1	4
9	1	1	1	1	1	1	2
10	10	4	7	7	12	13	8
11	11	3	4	1	2	7	5
12	4	4	5	2	17	6	2
13	10	5	6	6	12	13	9
14	10	4	7	6	12	13	8
15	3	3	2	2	9	8	2
16	4	4	1	2	17	1	2
17	4	4	1	2	17	6	2
18	4	2	4	1	8	7	2
19	3	8	4	1	8	7	2
20	3	3	2	8	9	6	6
21	3	8	2	2	9	8	2
22	3	3	2	8	1	11	6
23	3	3	4	1	3	1	1
24	3	2	1	5	15	4	3
25	3	2	1	5	7	4	3
26	3	2	1	1	7	10	2
27	2	2	4	1	13	12	7
28	3	3	4	2	16	14	2
29	3	2	4	2	7	1	3
30	3	2	4	1	7	10	2
31	3	2	4	1	8	7	2
32	3	2	4	1	4	10	5
33	3	8	2	8	9	6	6
34	3	7	4	1	14	15	10
35	3	7	4	1	14	15	2
36	4	4	1	2	17	6	3
37	3	2	1	1	4	1	3
38	1	1	1	1	1	1	1
39	1	1	1	1	1	21	4
40	1	1	1	4	1	1	4
41	1	1	1	9	1	1	1
42	1	1	4	1	1	1	4
43	2	2	4	1	1	1	2
44	2	2	4	2	3	7	1
45	2	2	4	10	8	19	2
46	2	4	4	1	19	17	6
47	3	2	1	1	20	1	3
48	3	2	1	2	10	4	2
49	3	2	3	1	5	5	2
50	3	2	3	1	5	18	2
51	3	2	4	1	11	1	2
52	3	2	4	1	11	3	5

Strain	emm type	Group	carbohydrate	ST	Location	Collection
168554	stG485	G	47	Portugal	UL	
171712	stG488	G	38	Portugal	UL	
220269	stG2078	G	15	Portugal	UL	
223754	stC839	C	3	Portugal	UL	
230631	stG488	G	8	Portugal	UL	
231995	stC74a	G	29	Portugal	UL	
241940	stC36	C	50	Portugal	UL	
273600	stG166b	G	65	Portugal	UL	
299298	stG643	G	8	Portugal	UL	
313247	stG6	G	25	Portugal	UL	
363962	stG2078	G	17	Portugal	UL	
378119	stC839	G	15	Portugal	UL	
380870	stG488	G	41	Portugal	UL	
386041	stC839	C	3	Portugal	UL	
394314	stG2078	G	72	Portugal	UL	
423738	stG62647	C	20	Portugal	UL	
450784	stG10	G	15	Portugal	UL	
460880	stG10	G	15	Portugal	UL	
493188	stG485	C	69	Portugal	UL	
542567	stG6	G	62	Portugal	UL	
613230	emm57	G	57	Portugal	UL	
SH0094	stG6792	G	4	Portugal	UL	
SH0015	stG6	G	25	Portugal	UL	
SH0032	stG166b	G	15	Portugal	UL	
SH0102	stG2078	G	17	Portugal	UL	
SH0107	stG643	G	52	Portugal	UL	
SH0110	stG6	G	25	Portugal	UL	
SH0113	stG6792	G	4	Portugal	UL	
SH0124	stG6792	G	4	Portugal	UL	
SH0218	stG245	G	15	Portugal	UL	
SH0254	stG485	C	69	Portugal	UL	
SH0257	stC6979	C	80	Portugal	UL	
SH0259	stG652	G	71	Portugal	UL	
SH0275	stG485	G	55	Portugal	UL	
SH0330	stC36	C	49	Portugal	UL	
SH0336	stG5420	G	25	Portugal	UL	
G121	stC74a	G	29	Australia	QIMR	
G122	stC74a	G	29	Australia	QIMR	
GCS10128	stC1400	C	46	Australia	QIMR	
GC52816	stG62647	C	20	Australia	QIMR	
GC36894	stG62647	C	20	Australia	QIMR	
GC36929	stG62647	C	20	Australia	QIMR	
GG3075	stG166b	G	56	Australia	QIMR	
GG3101	stG643	G	12	Australia	QIMR	
GG3106	stG6	G	44	Australia	QIMR	
GG31172	stC74a	G	29	Australia	QIMR	
GG31543	stG643	G	12	Australia	QIMR	
GG3120	stG4831	G	74	Australia	QIMR	
GG319	stC1400	G	64	Australia	QIMR	
GG32	stG10	G	15	Australia	QIMR	
GG324	stG6	G	44	Australia	QIMR	
GG3430	stG643	G	12	Australia	QIMR	

The Typing data should be a tab separated file containing the allelic profiles, formatted as follows: the first line should contain the column headers (usually locus identifiers be it either SNP, MLST or cg/wgMLST locus). The first column should be the allelic profile identifier (for MLST this would be the Sequence Type number, for any other method could be an unique strain ID. however if two strains have the same profile they should be given the same ID). The following columns are the loci used in the analysis.

If the Isolate data file is not used, the Typing data file should also represent the number of repeated profiles in a

collection, that is to say that if a given profile appears in a collection n times it should be repeated in the Typing data file n times.

In case of an Isolate data file is used the frequency of each type will be represented by the number of entries with a given Sequence type, in the Isolate file only and the frequency represented by repeated profiles in the Typing data file will not be used.

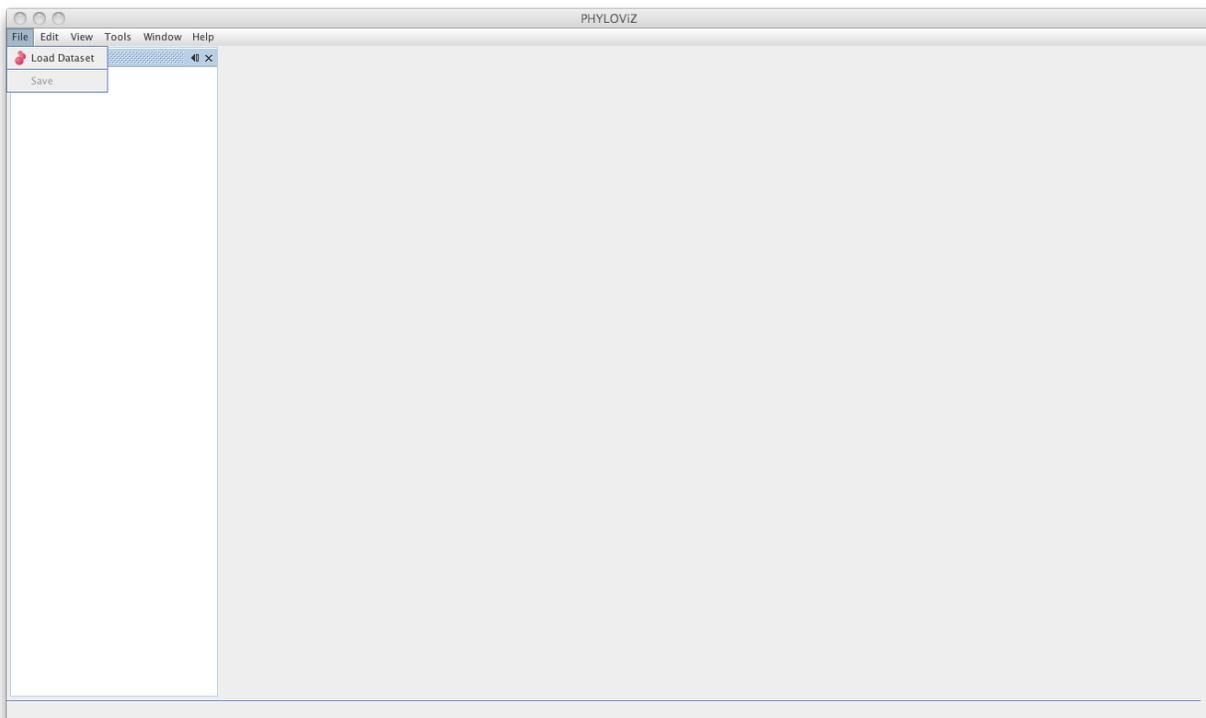
You can find an example of MLST data correctly formatted here. Note that in this file several STs are represented by more than one isolate (e.g. ST3 was found in 6 isolates).

The Isolate data file can contain epidemiological and/or demographic data or any other data you want to visualize overlaid onto the results of the analysis algorithms. The link between the data in the two files is made by the Sequence Type identifier. You can find an example file correctly formatted here.

2.2 Loading a Dataset

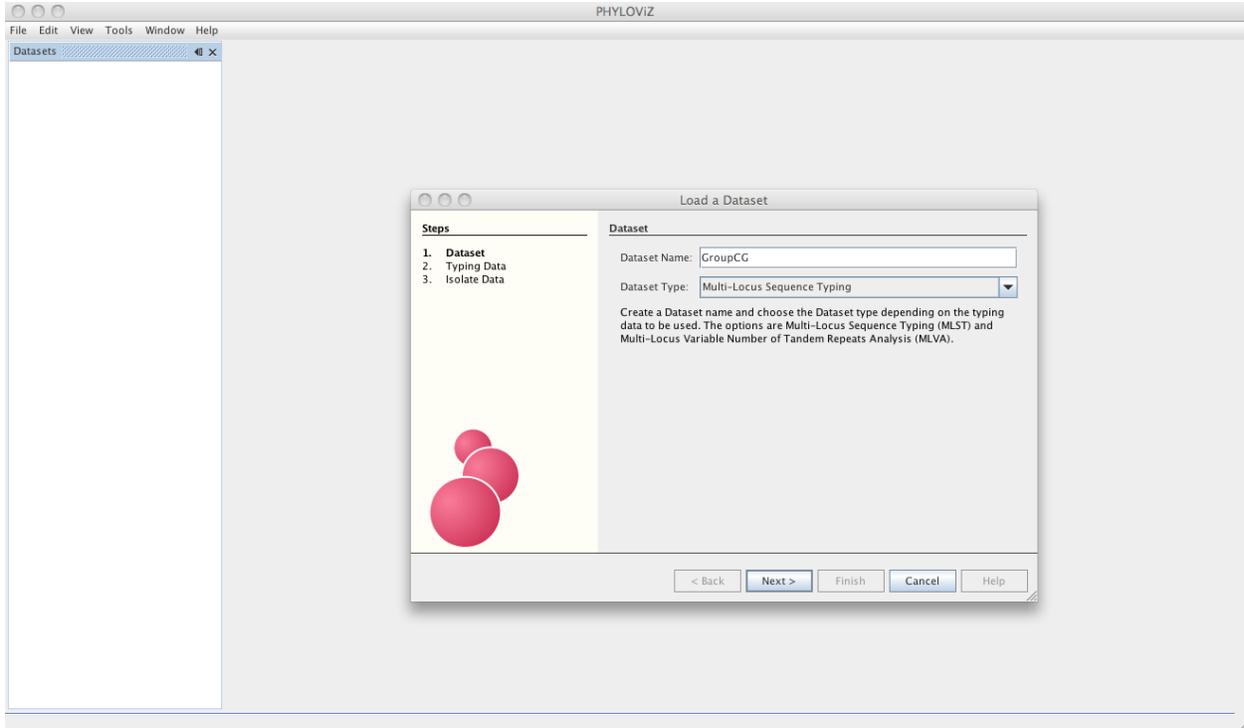
Go to *File* menu and choose *Load Dataset*.

If any errors in the data loading process are found they will be displayed in the session *Tab*. In the following screenshot you can see an example where allelic profiles were repeated with different identifiers. In the example data, we created ST81 as copy of ST1 profile and PHYLOViZ detects it and eliminates it from the analysis.

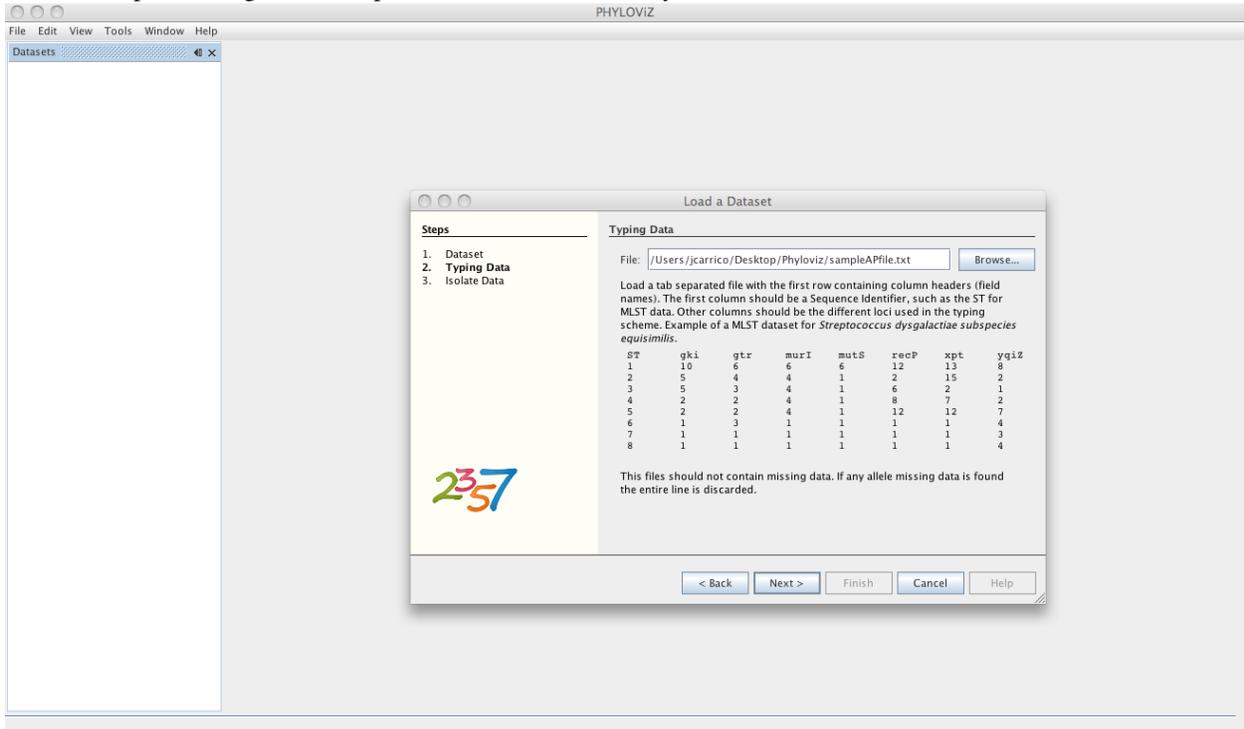


The dialog will now guide the user in the loading of the data. The first step is choosing a name for your Dataset since now PHYLOViZ supports multiple datasets open simultaneously. You must also choose the Dataset Type from the dropbox menu.

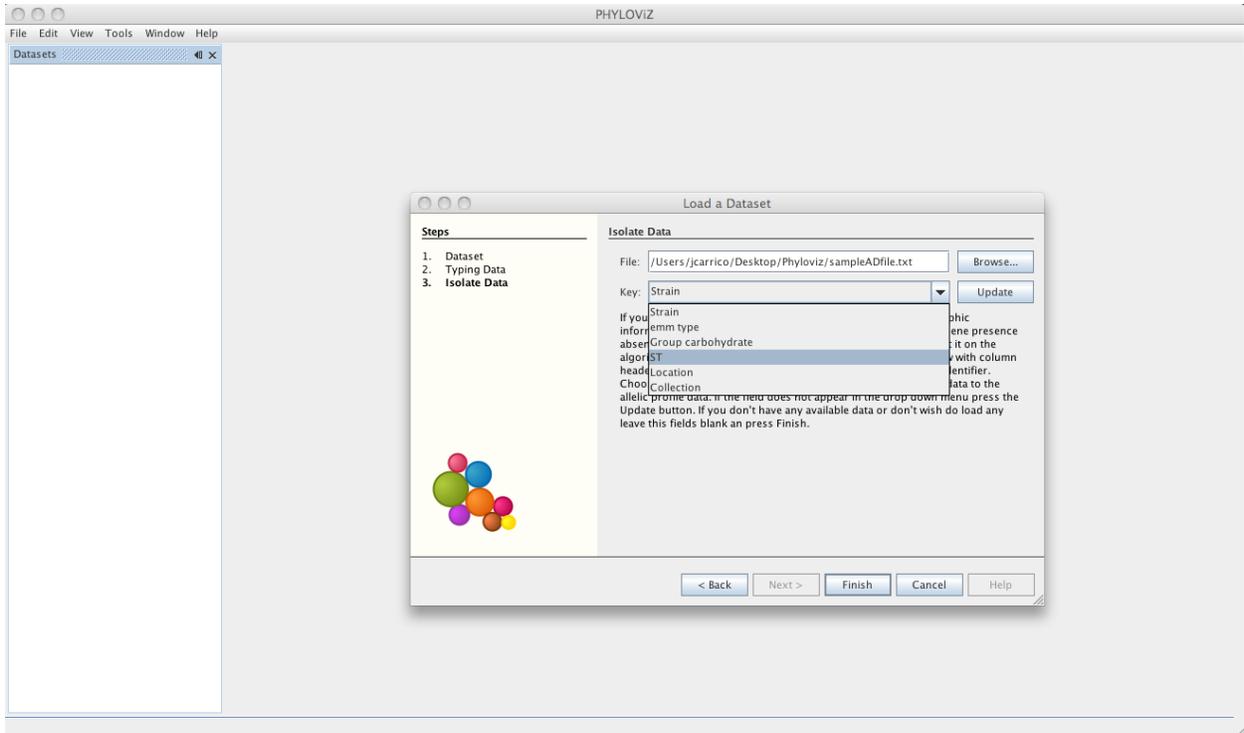
The Dataset type can be MLST or MLVA datasets with any number of loci, without any missing data. Lines with missing data will be excluded on load. If you have installed the Single Nucleotide Polymorphism (SNP) plugin, you can also access it on the Dataset type. See the [Sample Datasets](#) page to access some test data for the sequence-based typing methods available.



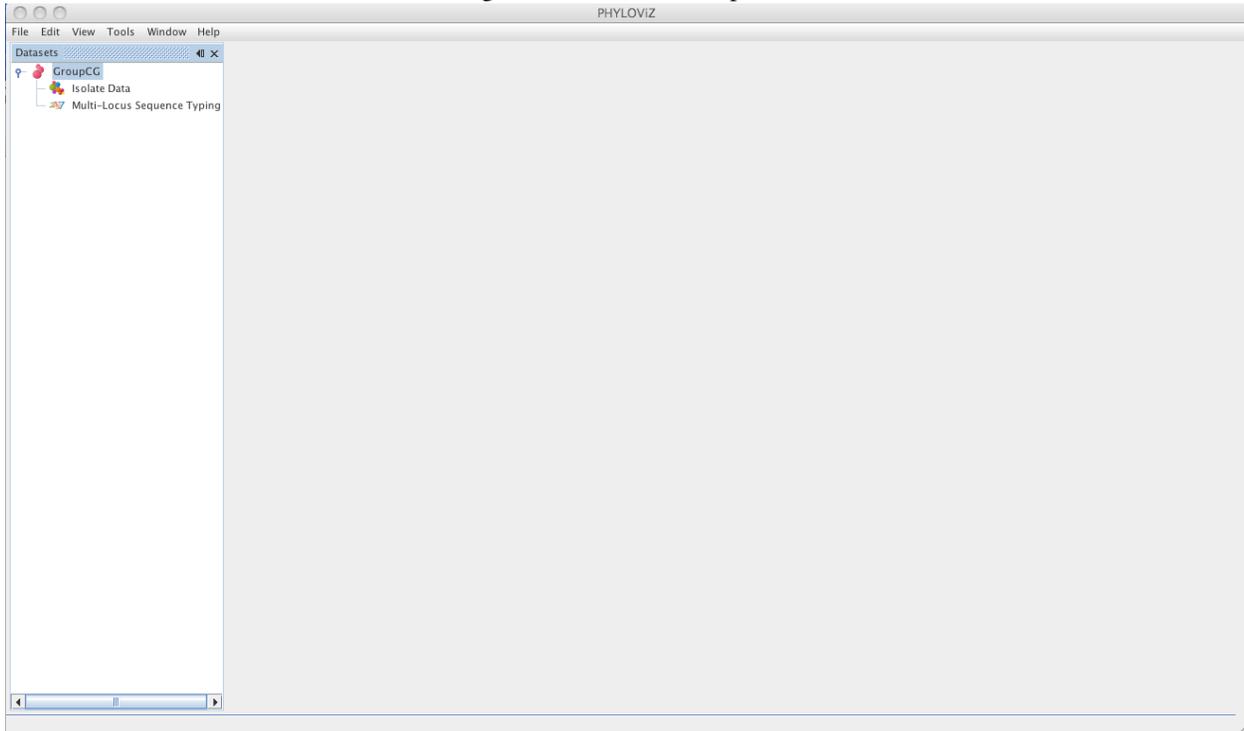
The next step is loading the allelic profile data for the method you selected.



After loading the allelic profile data, you can choose a file with information on your isolates for which the allelic profile was loaded. The linking field, as explained before, should be the Sequence Identifier and should be selected in the Key dropdown menu.



Then the dataset is loaded and double clicking on the dataset name opens the available data.



Double clicking on *Isolate Data* and *Typing Data* in the tree menu under the dataset name opens the respective tabs.

PHYLOViZ

File Edit View Tools Window Help

Datasets: GroupCG: Isolate Data x

GroupCG: Isolate Data

Multi-Locus Sequence Typing

View: table tree Regex filter:

Strain	emm type	Group carb...	ST	Location	Collection
168554	stG485	G	47	Portugal	UL
171712	stG480	G	38	Portugal	UL
220269	stG2078	G	15	Portugal	UL
223754	stC839	C	3	Portugal	UL
230631	stG480	G	8	Portugal	UL
231995	stC74a	G	29	Portugal	UL
241940	stC36	C	50	Portugal	UL
273600	stG166b	G	65	Portugal	UL
299298	stG643	G	8	Portugal	UL
313247	stG6	G	25	Portugal	UL
363962	stG2078	G	17	Portugal	UL
378119	stC839	G	15	Portugal	UL
380870	stG480	G	41	Portugal	UL
386041	stC839	C	3	Portugal	UL
394314	stG2078	G	72	Portugal	UL
423738	stG2647	C	20	Portugal	UL
450784	stG10	G	15	Portugal	UL
460880	stG10	G	15	Portugal	UL
493188	stG485	C	69	Portugal	UL
542567	stG6	G	62	Portugal	UL
618280	emm57	G	57	Portugal	UL
SH0004	stG6792	G	4	Portugal	UL
SH0015	stG6	G	25	Portugal	UL
SH0032	stG166b	G	15	Portugal	UL
SH0102	stG2078	G	17	Portugal	UL
SH0107	stG643	G	52	Portugal	UL
SH0110	stG6	G	25	Portugal	UL
SH0113	stG6792	G	4	Portugal	UL
SH0124	stG6792	G	4	Portugal	UL
SH0218	stG245	G	15	Portugal	UL
SH0254	stG485	C	69	Portugal	UL

GroupCG: Isolate Data (Selection view)

Legend

Load Save

PHYLOViZ

File Edit View Tools Window Help

Datasets: GroupCG: Isolate Data x GroupCG: Multi-Locus Sequence Typing x

GroupCG: Multi-Locus Sequence Typing

View: table tree Regex filter:

ST	gki	gtr	murI	mutS	recP	xpt	yqiZ
1	10	6	6	6	12	13	8
2	5	4	4	1	2	15	2
3	5	3	4	1	6	2	1
4	2	2	4	1	8	7	2
5	2	2	4	1	12	12	7
6	1	3	1	1	1	1	4
7	1	1	1	1	1	1	3
8	1	1	1	1	1	1	4
9	1	1	1	1	1	1	2
10	10	4	7	7	12	13	8
11	11	3	4	1	2	7	5
12	4	4	5	2	17	6	2
13	10	5	6	6	12	13	9
14	10	4	7	6	12	13	8
15	3	3	2	2	9	8	2
16	4	4	1	2	17	1	2
17	4	4	1	2	17	6	2
18	4	2	4	1	8	7	2
19	3	8	4	1	8	7	2
20	3	3	2	8	9	6	6
21	3	8	2	2	9	8	2
22	3	3	2	8	1	11	6
23	3	3	4	1	3	1	1
24	3	2	1	5	15	4	3
25	3	2	1	5	7	4	3
26	3	2	1	1	7	10	2
27	2	2	4	1	13	12	7
28	3	3	4	2	16	14	2
29	3	2	4	2	7	1	3
30	3	2	4	1	7	10	2
31	3	2	4	1	8	7	2

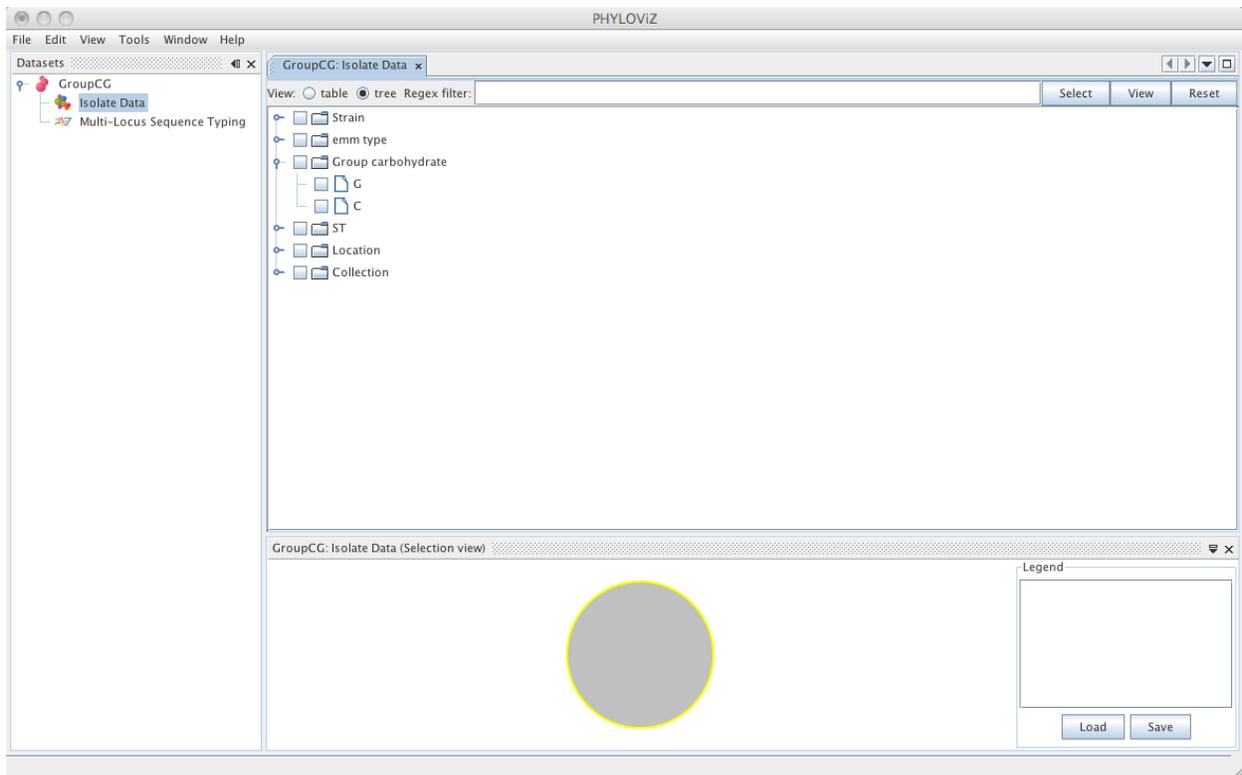
GroupCG: Isolate Data (Selection view)

GroupCG: Multi-Locus Sequence Typing (Selection view)

Legend

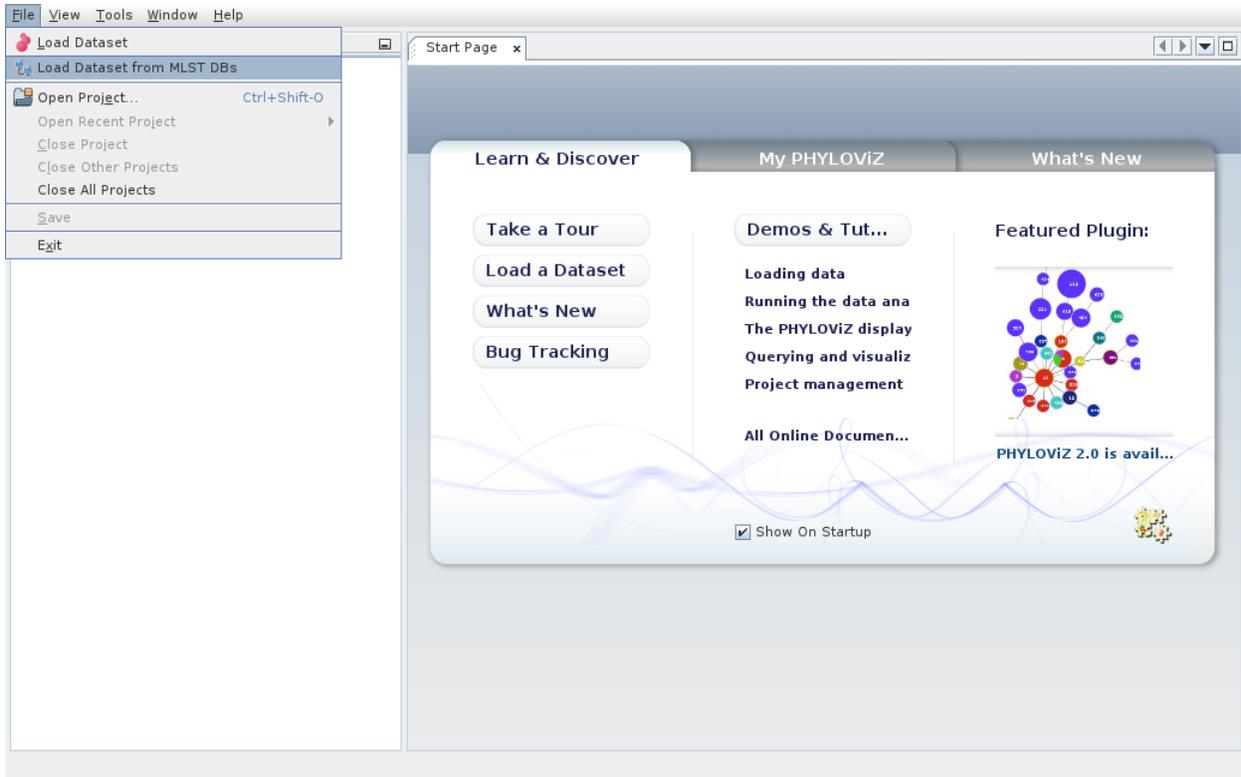
Load Save

The default view is the *table* view. Also available is the *tree* view, where it is easier to visualize what information is available in the different fields and to select combinations of fields with specific values.



2.3 Loading a remote Dataset

We can also load datasets from remote databases and services. PHYLOViZ contains already a list of available databases. We can choose *Load Dataset from MLST DBs*.



There are several datasets available from several providers. In the following example we select the *Streptococcus pneumoniae* dataset from PubMLST.org.

Steps

1. Database
2. Typing Data
3. Isolate Data
4. Sequence Data

Database

Dataset Name:

Public DB Name:

This plugin create isolate data from
 . MLST.net
 . Pubmlst.org
 . mlst.UCC.ie
 . www.pasteur.fr/
 . www.shigatox.net

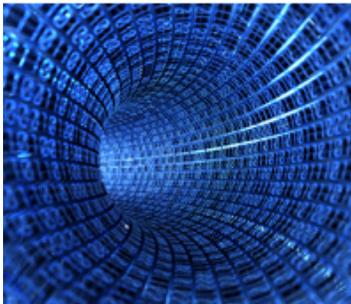
pubmlst.org - Stenotrophomonas maltophilia
 pubmlst.org - Streptococcus agalactiae
 pubmlst.org - Streptococcus canis
 pubmlst.org - Streptococcus dysgalactiae equis
 pubmlst.org - Streptococcus gallolyticus
 pubmlst.org - Streptococcus oralis
 pubmlst.org - Streptococcus pneumoniae
 mlst.net - Streptococcus pyogenes

and

Choose a dataset name and an available online database for a given microorganism to continue.

An internet connection is necessary to communicate with the available webservices.

This is a BETA version of the plugin. Please contact us if some bug is detected.



The next step is to download the dataset.

Steps

1. Database
2. **Typing Data**
3. Isolate Data
4. Sequence Data

Typing Data

Dataset Name: Streptococcus pneumoniae

Profile: aroE gdh gki recP spi xpt ddl

Dataset Size: 10061 STs

Loading from public databases the typing data of a specific dataset.

Each entry in a dataset is composed by a Sequence Identifier, followed by the different loci in the typing scheme.

Press the *Start/Stop* button to initiate the download.

If you press the *Back* button, you will have to restart the download.



In the next window we can load ancillary data on isolates. In this example we choose to not load any data.

Steps

1. Database
2. Typing Data
3. **Isolate Data**
4. Sequence Data

Isolate Data

File:

Key:

If you have any type of ancillary data about the isolates (Demographic information, Epidemiological information, Antibiotic Resistance, Gene presence absence, etc) you can load a tab separated file to further represent it on the algorithm. This file should be a tab separated file with the first row with column headers (field names). One of the fields should be the Sequence identifier. Choose this field in the Key drop down menu to link the ancillary data to the allelic profile data. If the field does not appear in the drop down menu press the Update button. If you don't have any available data or don't wish to load any leave these fields blank and press Finish.



We can also load sequence data for each allele. They are downloaded individually and loaded as typing data.

Steps

1. Database
2. Typing Data
3. Isolate Data
4. **Sequence Data**

Sequence Data

- No sequence data
- Load sequence data from public databases

aroE	Done!	or	<input type="text"/>	Browse...
gdh	Download	or	<input type="text"/>	Browse...
gki	Download	or	<input type="text"/>	Browse...
recP	Download	or	<input type="text"/>	Browse...
spi	Download	or	<input type="text"/>	Browse...
xpt	Download	or	<input type="text"/>	Browse...
ddl	Download	or	<input type="text"/>	Browse...

Choose this option to download all sequence data available on the *loci* from the public database for the selected microorganism.

To import each *locus* independently, you can either:

- . Press **Download**, to transfer the sequences from a public database;
- . Press **Browse**, to load a Fasta file containing the sequences.

Finally press **Finish** to create the dataset and proceed with the analysis.

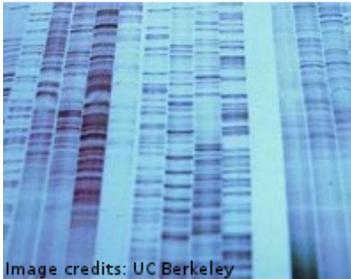


Image credits: UC Berkeley

At the end we have seven typing data items to explore and analyze.

File View Tools Window Help

Datasets x

- spneumo
 - Multi-Locus Sequence Typing (MLST)
 - aroE locus sequence
 - gdh locus sequence
 - gki locus sequence
 - recP locus sequence
 - spi locus sequence
 - xpt locus sequence
 - ddl locus sequence

Start Page x spneumo: aroE locus sequence x

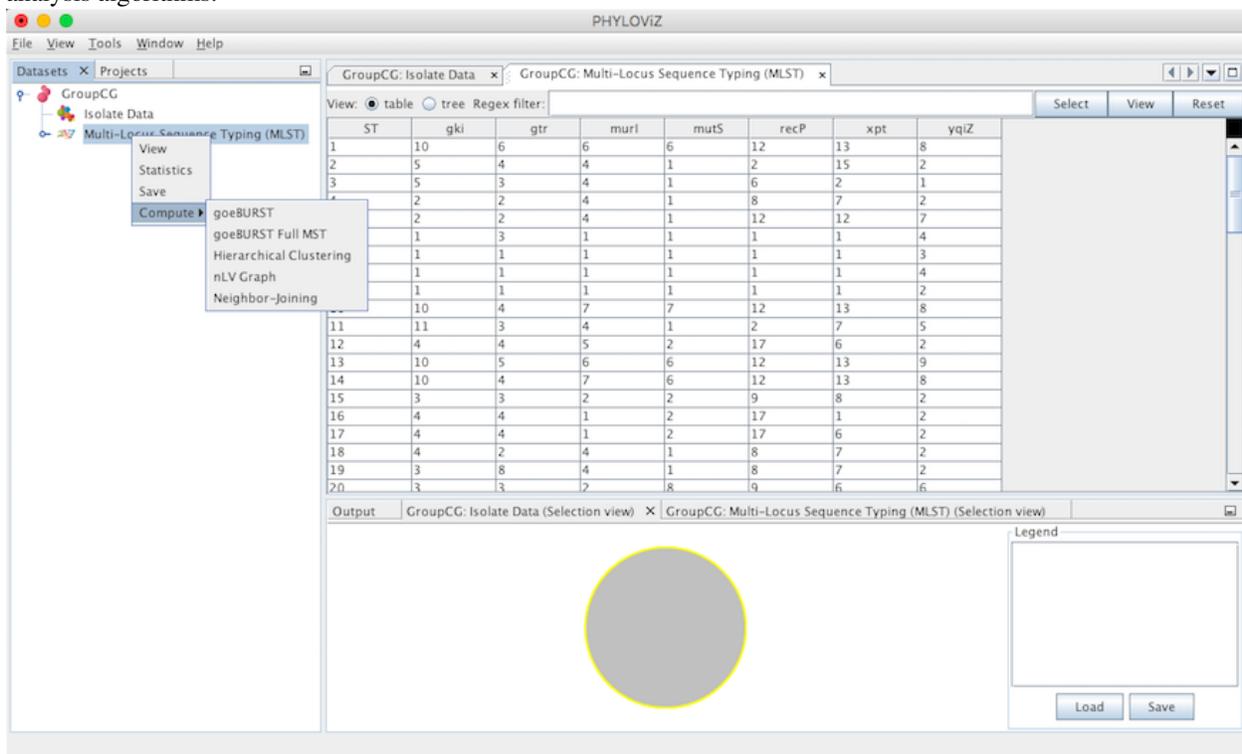
View: table tree Regex filter:

ID	s[0]	s[1]	s[2]	s[3]	s[4]	s[5]	s[6]	s[7]
aroE_1	G	A	A	G	C	G	A	G
aroE_2	G	A	A	G	C	G	A	G
aroE_3	G	A	A	G	C	G	A	G
aroE_4	G	A	A	G	C	G	A	G
aroE_5	G	A	A	G	C	G	A	G
aroE_6	G	A	A	G	C	G	A	G
aroE_7	G	A	A	G	C	G	A	G
aroE_8	G	A	A	G	C	G	A	G
aroE_9	G	A	A	G	C	G	A	G
aroE_10	G	A	A	G	C	G	A	G
aroE_11	G	A	A	G	C	G	A	G
aroE_12	G	A	A	G	C	G	A	G
aroE_13	G	A	A	G	C	G	A	G
aroE_14	G	A	A	G	C	G	A	G
aroE_15	G	A	A	G	C	G	A	G
aroE_16	G	A	A	G	C	G	A	G
aroE_17	G	A	A	G	C	G	A	G
aroE_18	G	A	A	G	C	G	A	G
aroE_19	G	A	A	G	C	G	A	G
aroE_20	G	A	A	G	C	G	A	G
aroE_21	G	A	A	G	C	G	A	G
aroE_22	G	A	A	G	C	G	A	G
aroE_23	G	A	A	G	C	G	A	G
aroE_24	G	A	A	G	C	G	A	G
aroE_25	G	A	A	G	C	G	A	G
aroE_26	G	A	A	G	C	G	A	G
aroE_27	G	A	A	G	C	G	A	G
aroE_28	G	A	A	G	C	G	A	G
aroE_29	G	A	A	G	C	G	A	G
aroE_30	G	A	A	G	C	G	A	G
aroE_31	G	A	A	G	C	G	A	G
aroE_32	G	A	A	G	C	G	A	G

spneumo: aroE locus sequence (Selection view)

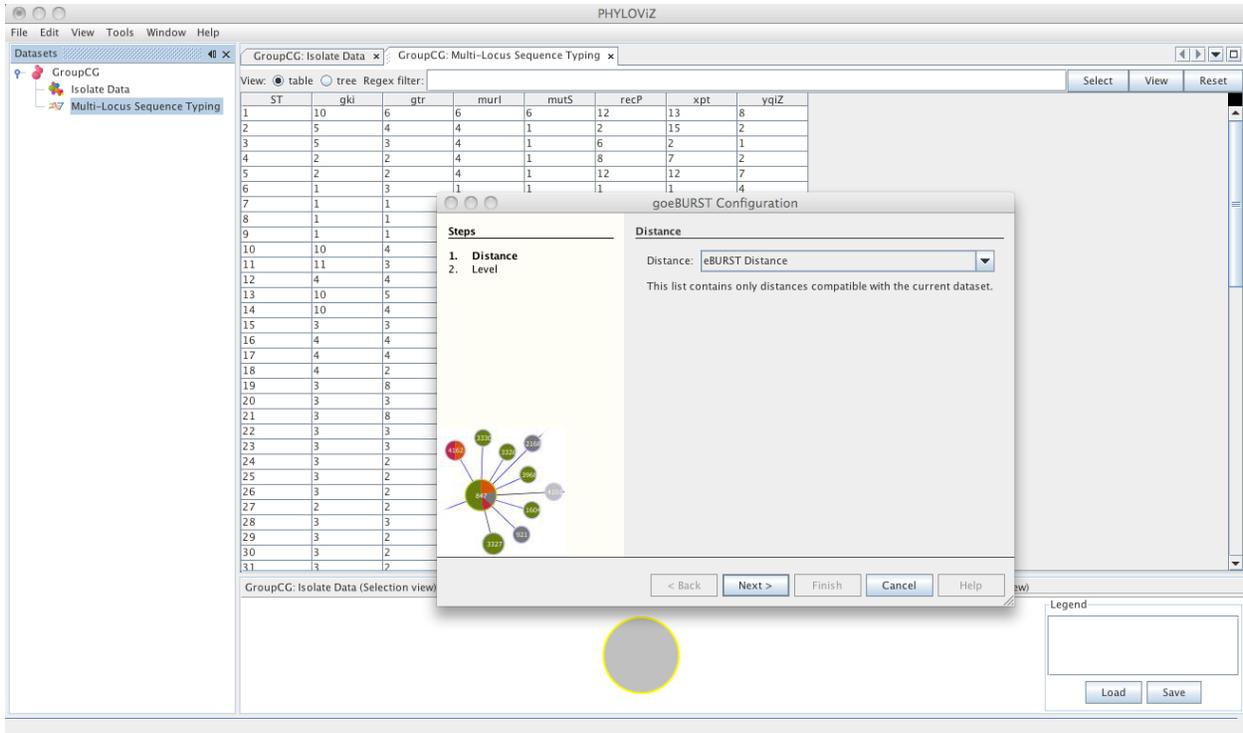
Data analysis

In the current version of PHYLOViZ, you can analyze your data using the several algorithms described below. Press the *Right Mouse Button* on the *Typing Data* (now named with the method) and choose compute to access the available analysis algorithms.

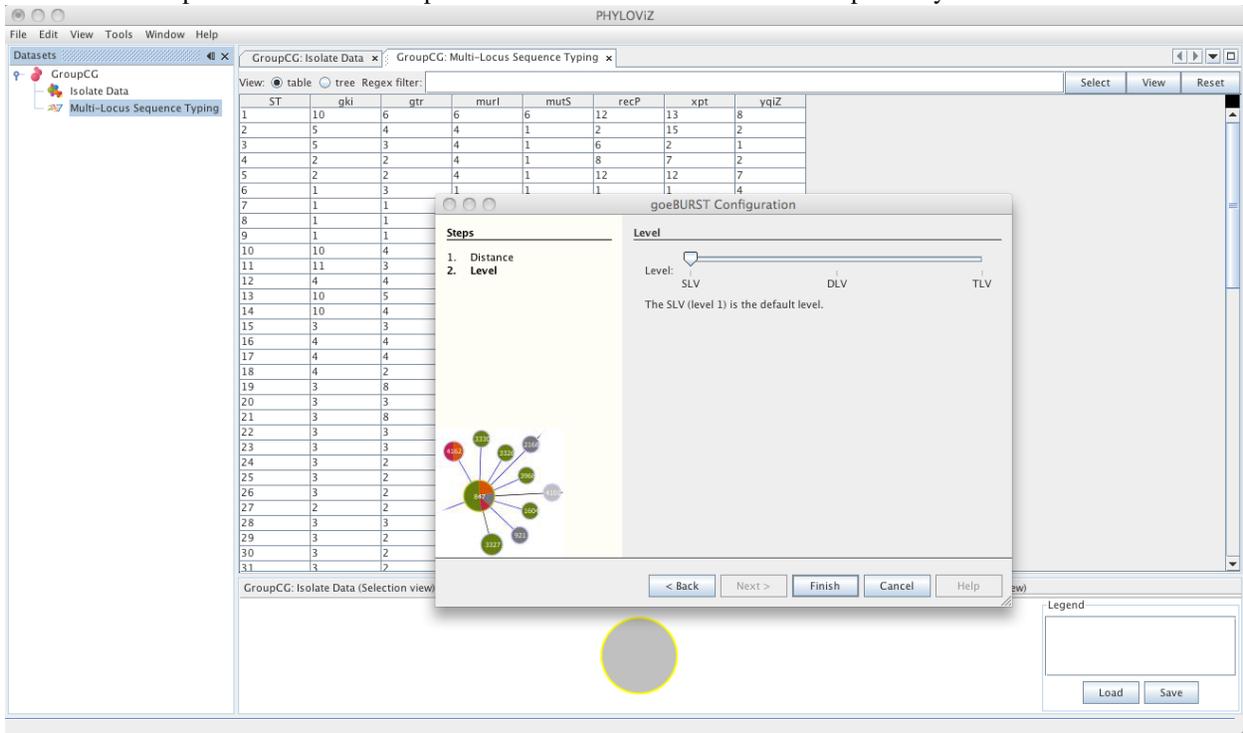


3.1 goeBURST algorithm

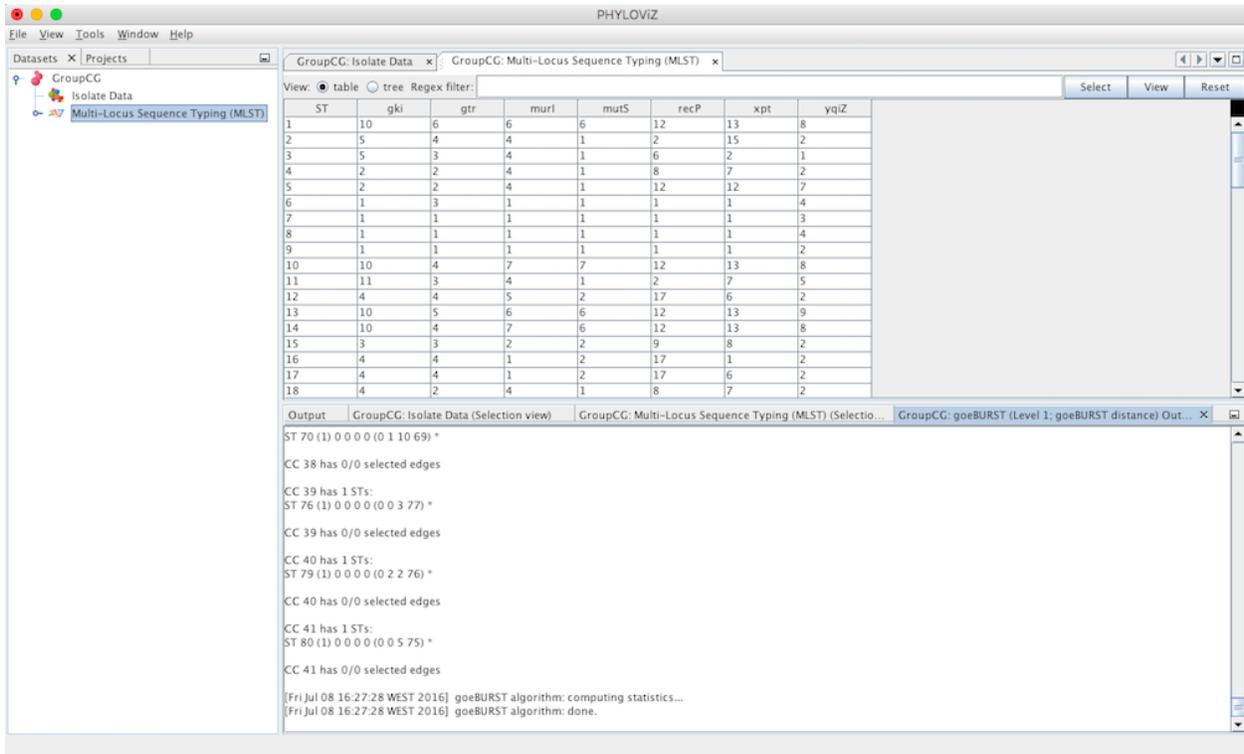
Selecting the goeBURST algorithms opens the dialog for the goeBURST algorithm. This algorithm was typically used for MLST data analysis and was originally described in the article [Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach](#). The first step is choosing the *Distance* to be used. Currently eBURST Distance is the only one available, but others could be implemented. The eBURST distances follows the tiebreak rules discussed in the article.



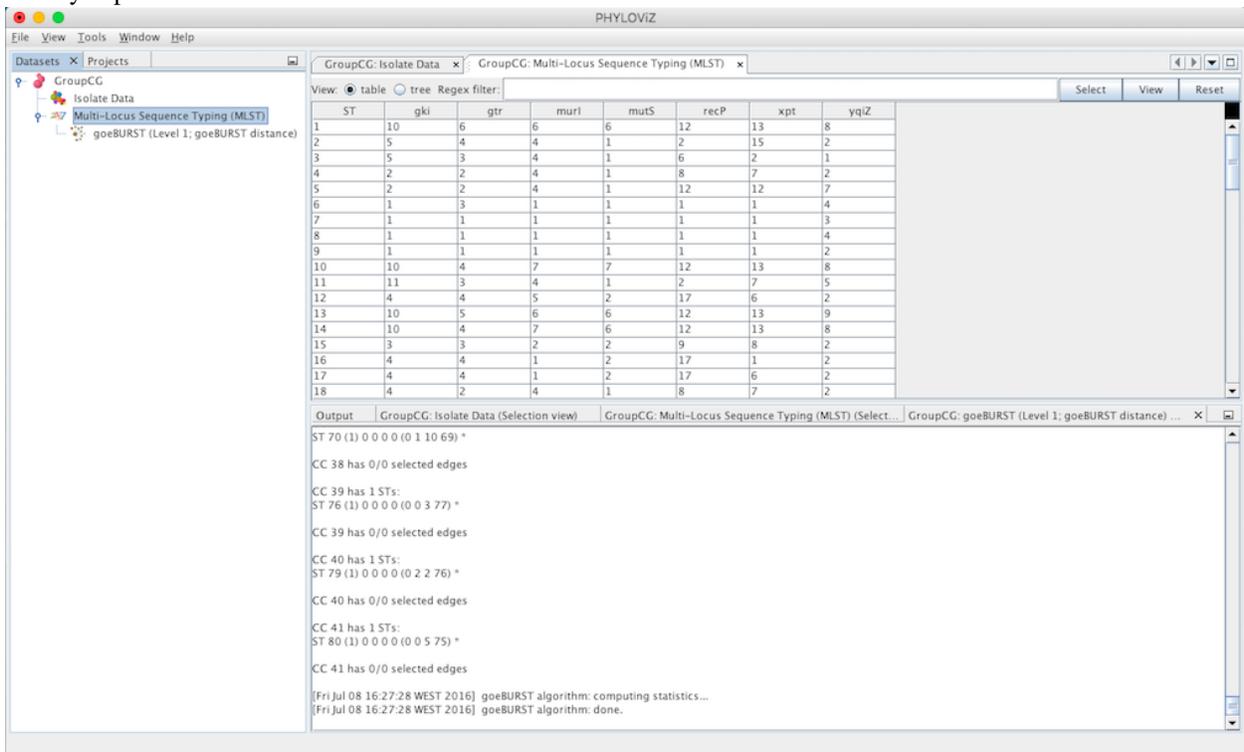
The second step is the choice of the level to which clonal complexes will be formed. The usual default for MLST analysis is SLV Level. Choosing DLV or TLV level will take longer calculation times, but could provide some insight to the relationships between clonal complexes formed at SLV and DLV level respectively.



A goeBURST *Output* tab will appear and display the goeBURST algorithm results. It will contain information about the Clonal Complexes (CCs), namely the Sequence Types that compose them and what edges (the links between STs) were drawn in each CC.

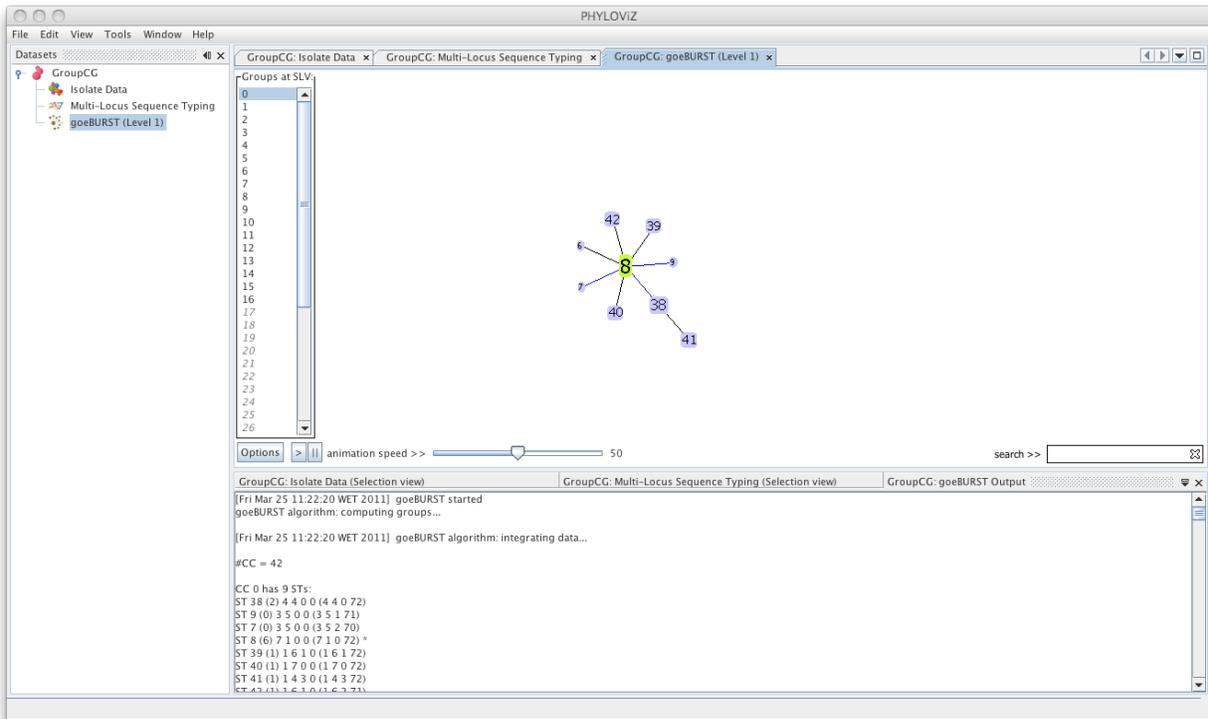


In order to display the goEburst tree view, it is necessary to expand the typing data on the DataSets' tab, if it is not already expanded.

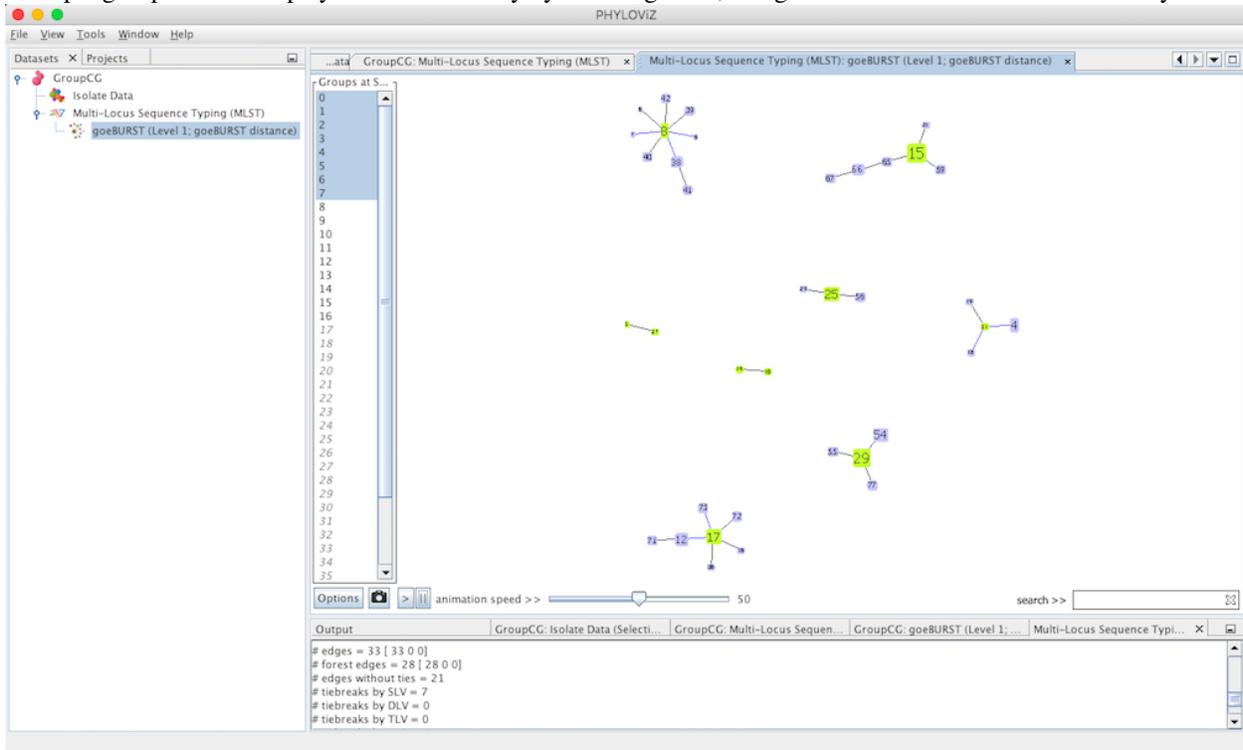


Double clicking on the goeBURST item that is now on the Dataset tree menu will show the display. The clonal complexes will be arbitrarily numbered starting from 0 (for the CC with most STs) and contains all the data relevant to the goeBURST analysis (STs in each group and the drawn SLVs edges). The following screenshot summarizes the

output for a single clonal complex with the test dataset used.



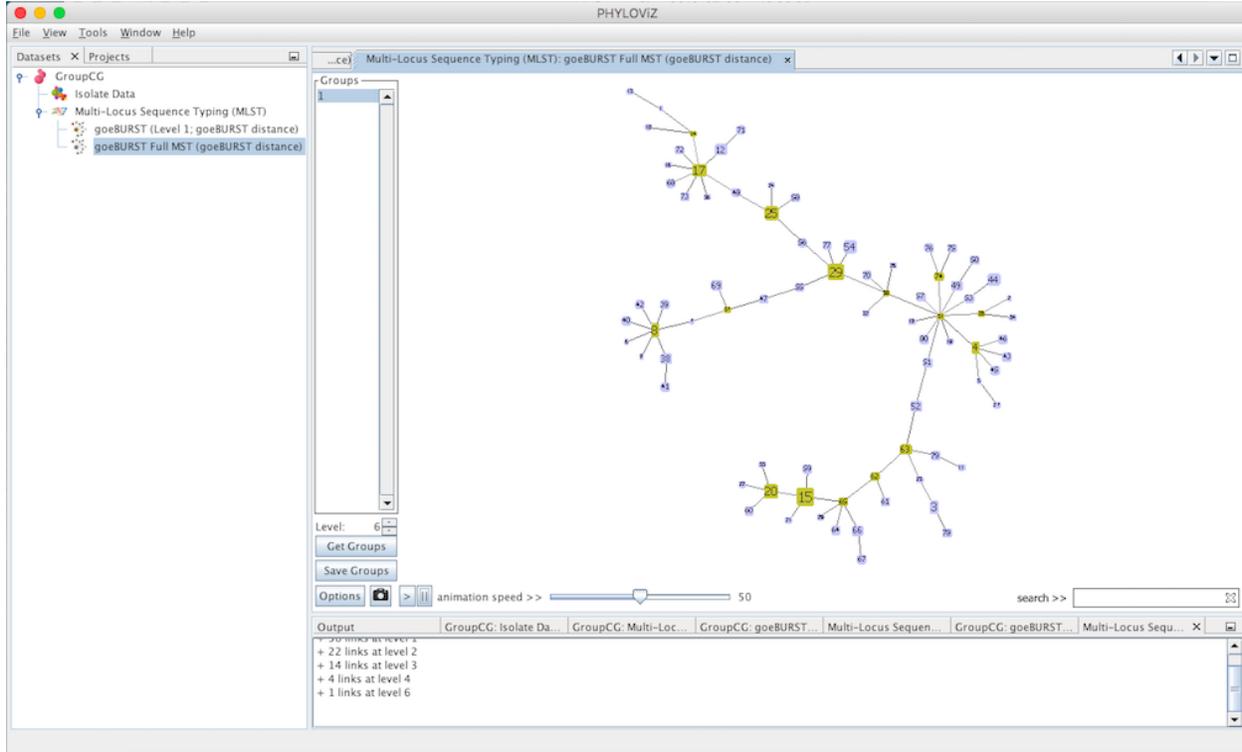
Multiple groups can be displayed simultaneously by selecting them, using the CTRL /CMD and/or SHIFT keys.



3.2 goeBURST Full MST algorithm

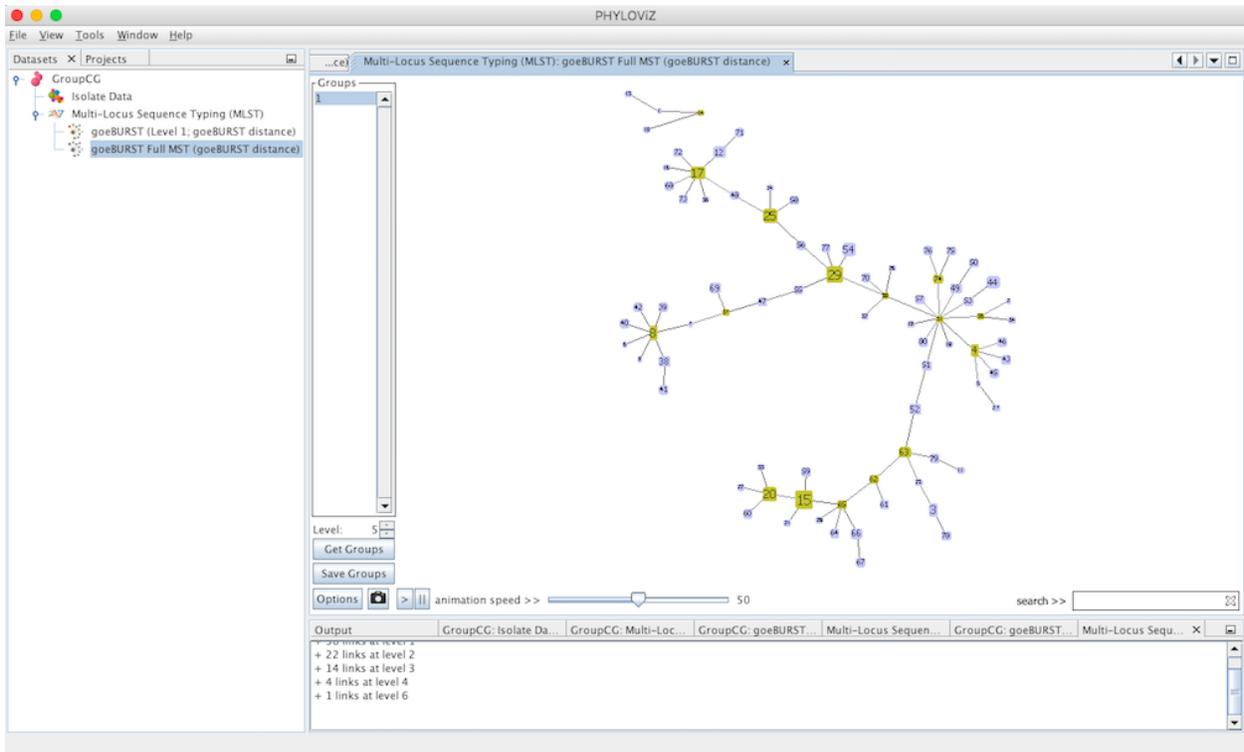
Using an extension of the goeBURST rules up to $\setminus(n)$ LV level (where $\setminus(n)$ equals to the number of loci your dataset uses), a Minimum Spanning Tree-like structure can be drawn. This is typically used for SNP or cg/wgMLST datasets with dozens to thousand of loci.

Select *goeBURST Full MST* in the *Compute* options to draw it. Contrary to the standard goeBURST, the link statistics are not presented. After computation, double click on the *goeBURST Full MST* that appears under the dataset heading to visualize the result.

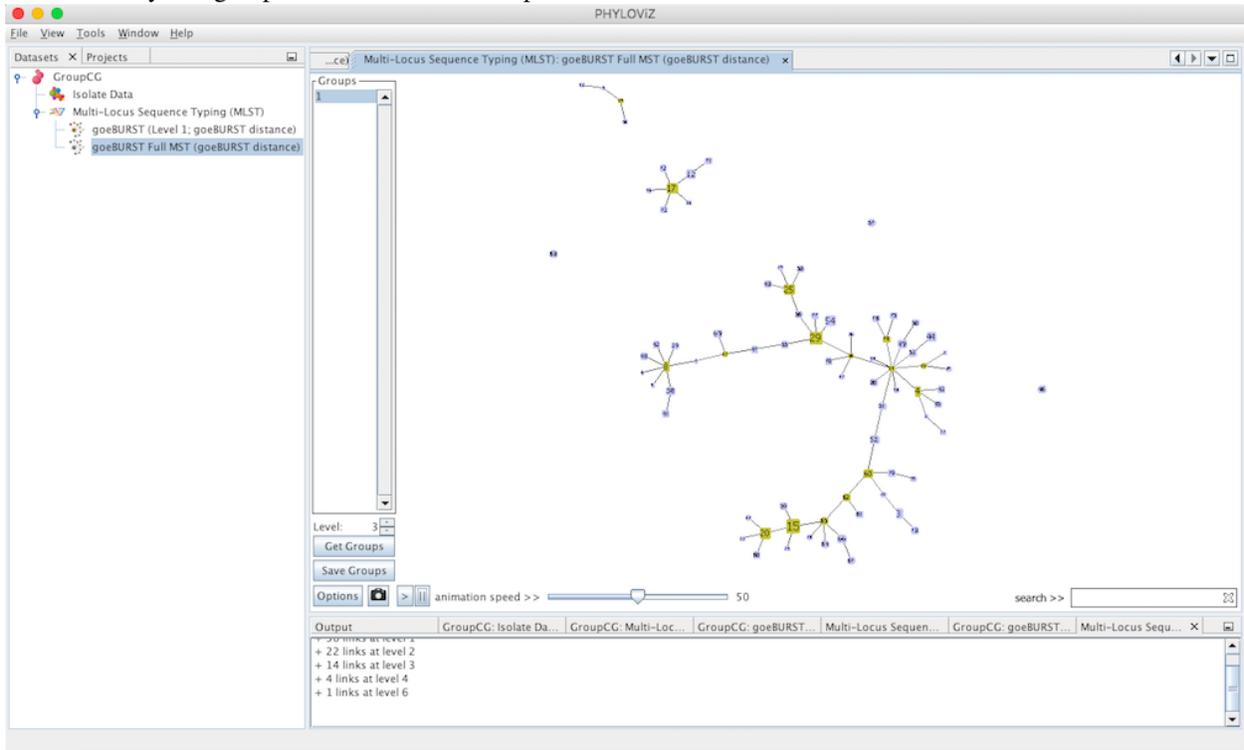


New options appear on the display: The *Level* selector and two new buttons *Get Groups* and *Save Groups*. The *Level* represents the *Locus Variant* level and allows the removal of all the links greater than the number represented. The user can use the up and down arrows or directly edit the number by clicking on it. The *Get Groups* button allows separate the display of groups that are not connected at the level chosen in order to simplify the analysis of larger datasets. This will generate a display very similar to that of goeBURST, but at a higher link level. The *Save Groups* creates an extra column in the isolate data with the title label *goeBURST MST[\setminus(x)]* with $\setminus(x)$ being equal to the level used to create the groups.

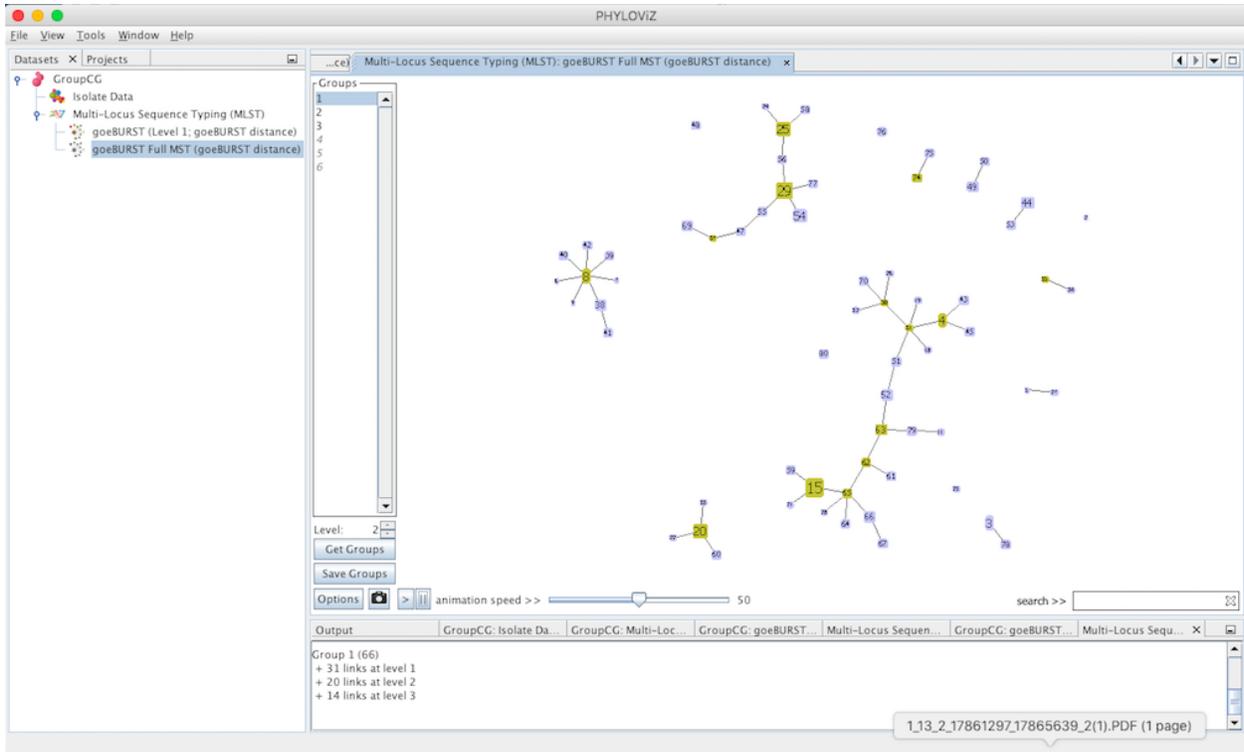
Decreasing the *Level* selector, allows the user to see how clonal complexes would relate to each other at a certain level. Level 1, 2 and 3 are equivalent to calculating goeBURST at those levels (SLV,DLV and TLV respectively). The following images shows what happens to the dataset when you decrease the level. Level 4 is not displayed since no new groups are formed at that level.



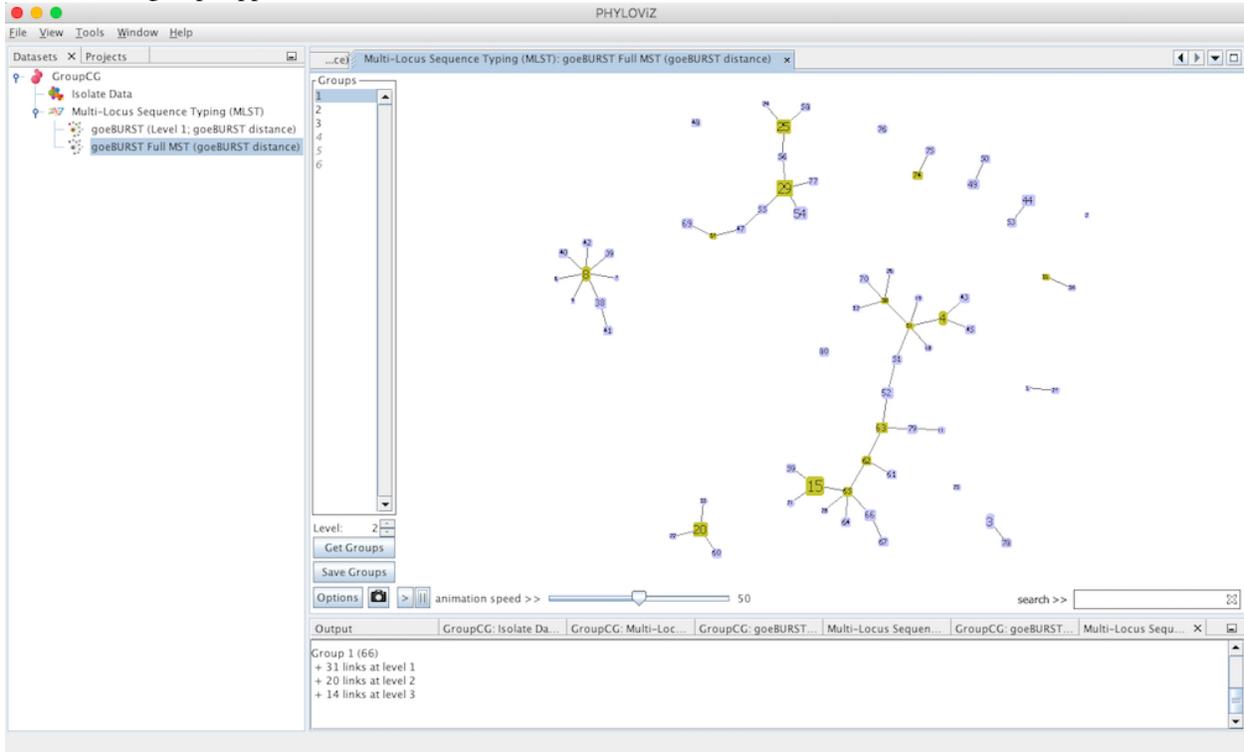
At level 5 only two groups are formed in the sample dataset.



At level 3 (TLV level) some singletons appear. Level 4 is not shown since no changes were observed in the graph. This means that there are no two STs in the dataset that differ in 4 of the loci of their profiles.



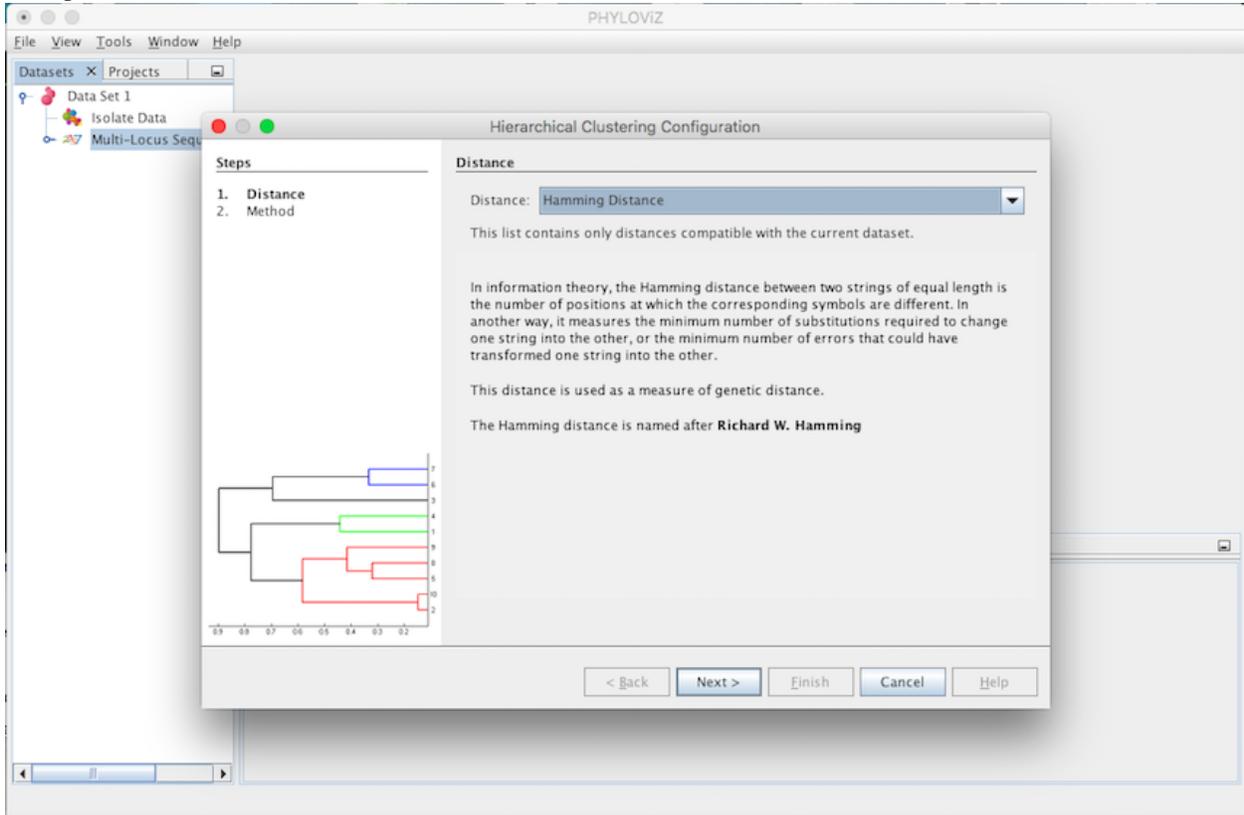
At level 2 , 6 groups appear with 4 or more STs each.



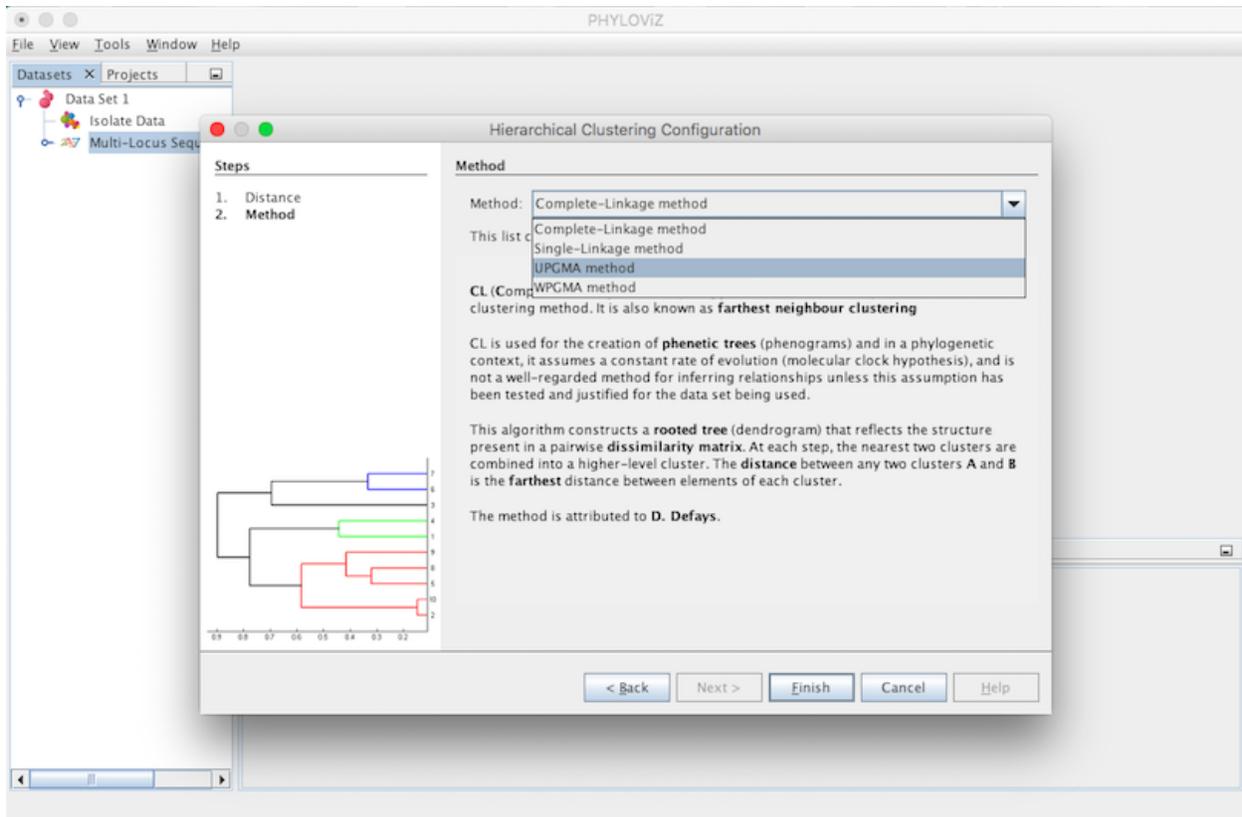
And finally at level 1, the equivalent of the most commonly used Clonal Complex definition by goeBURST, 17 groups with 2 or more STs are formed and there are 25 singletons on the dataset.

3.3 Hierarchical Clustering

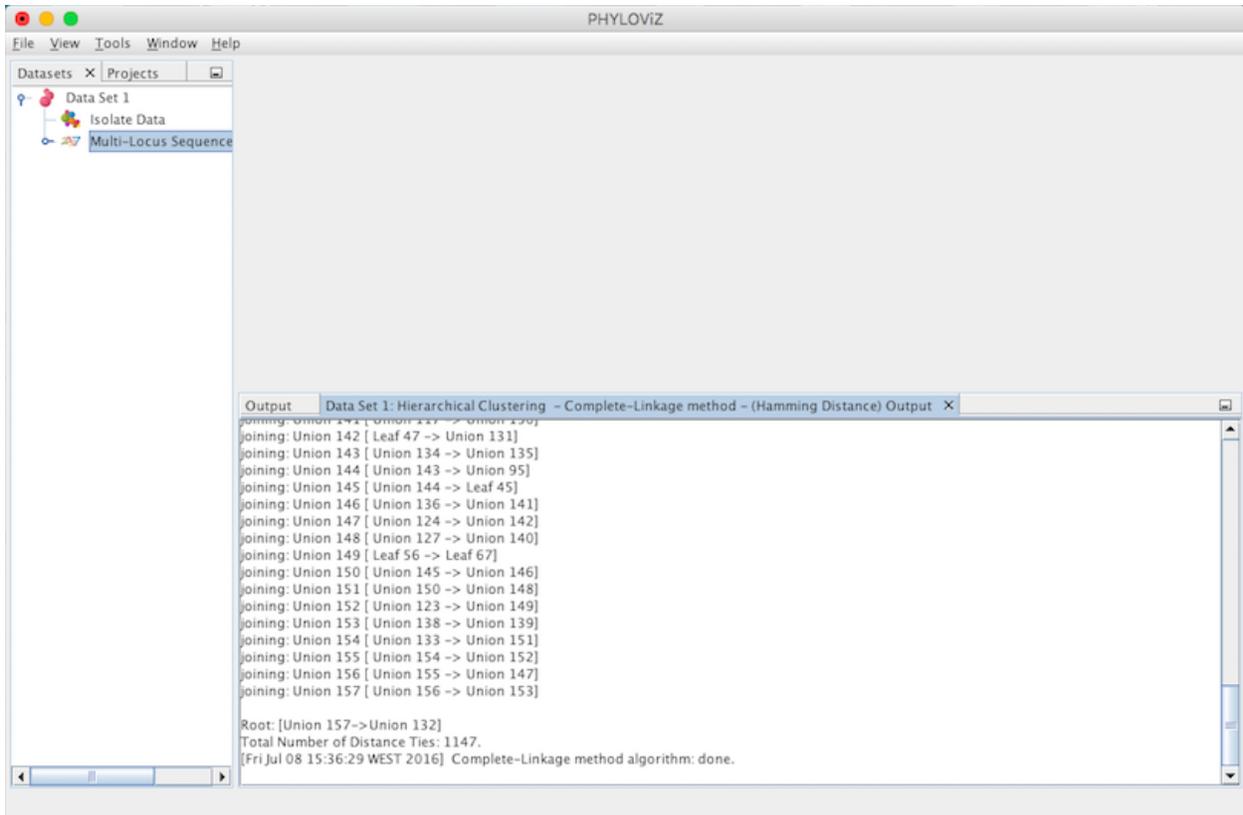
Selecting the Hierarchical Clustering opens the dialog where you can select what method you want to apply. The first step is choosing the *Distance* to be used. Currently the hamming distance is the only one available, but others could be implemented.



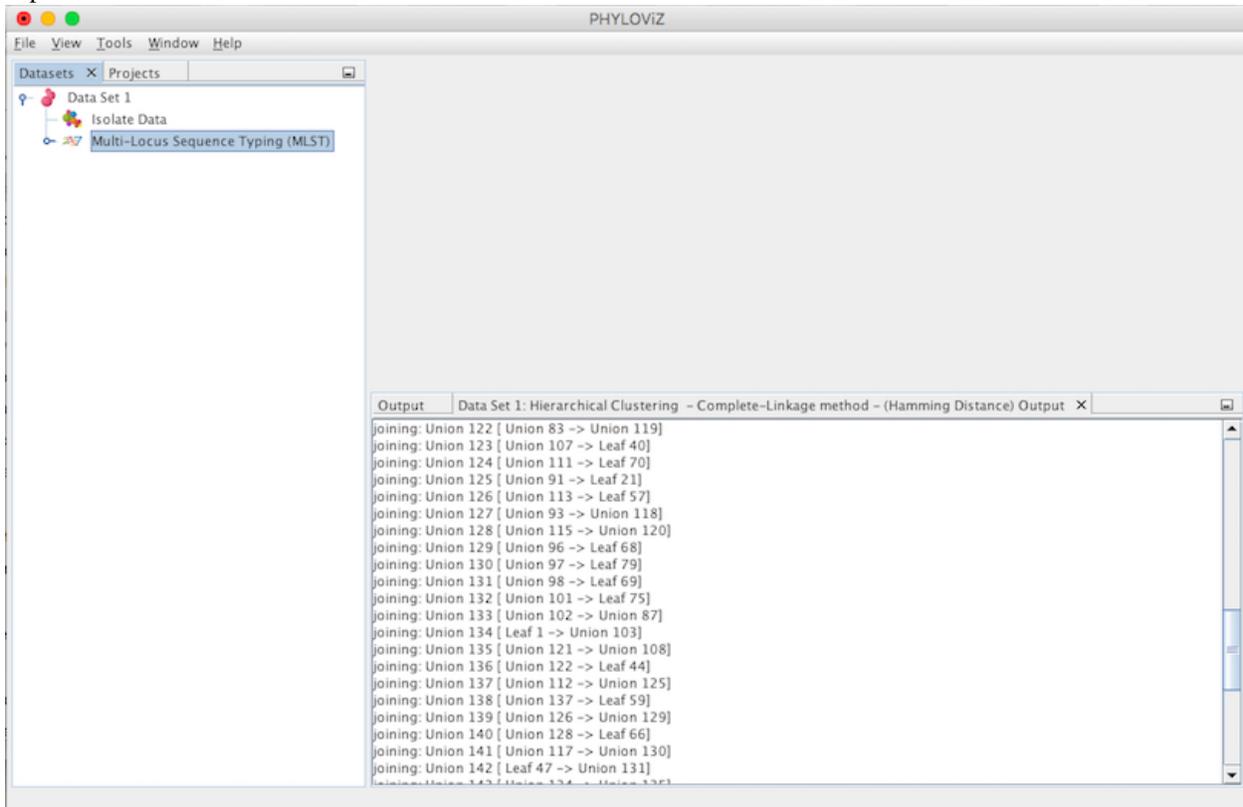
The second step is to select the *Method*. You can choose between complete-linkage, single-linkage, UPGMA (Unweighted Pair Group Method with Arithmetic mean) and WPGMA (Weighted Pair Group Method with Arithmetic mean). Selecting the method corresponds to selecting the criterion of minimal dissimilarity.



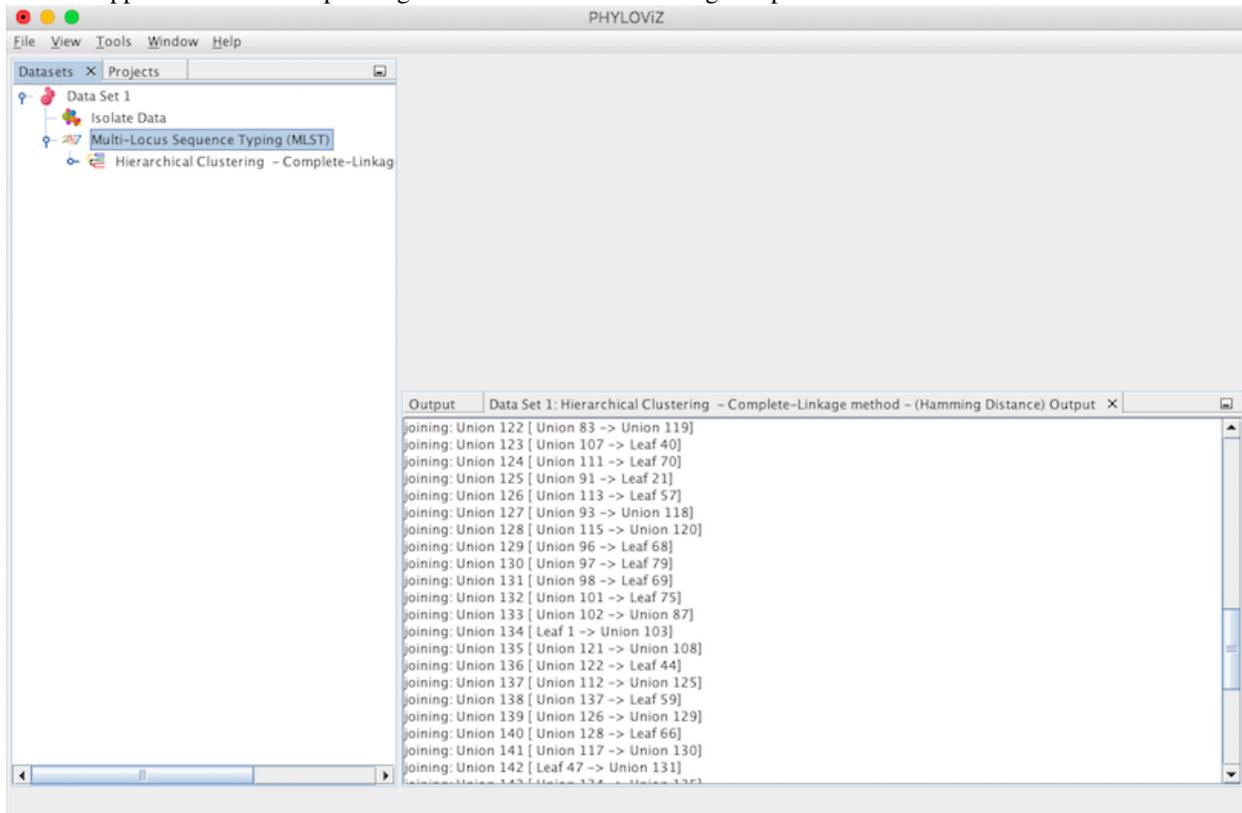
A Hierarchical Clustering *Output Tab* will appear and display the results of the application of the chosen method. A *Leaf* represents a Sequence Type and a *Union* represents a group that results of joining Leaves or Unions with Leafs. This process of joining is displayed step by step by the algorithm in the *Output's Tab*. Finally we have the number of ties occurred. The tie break applied is to always choose the first one found.



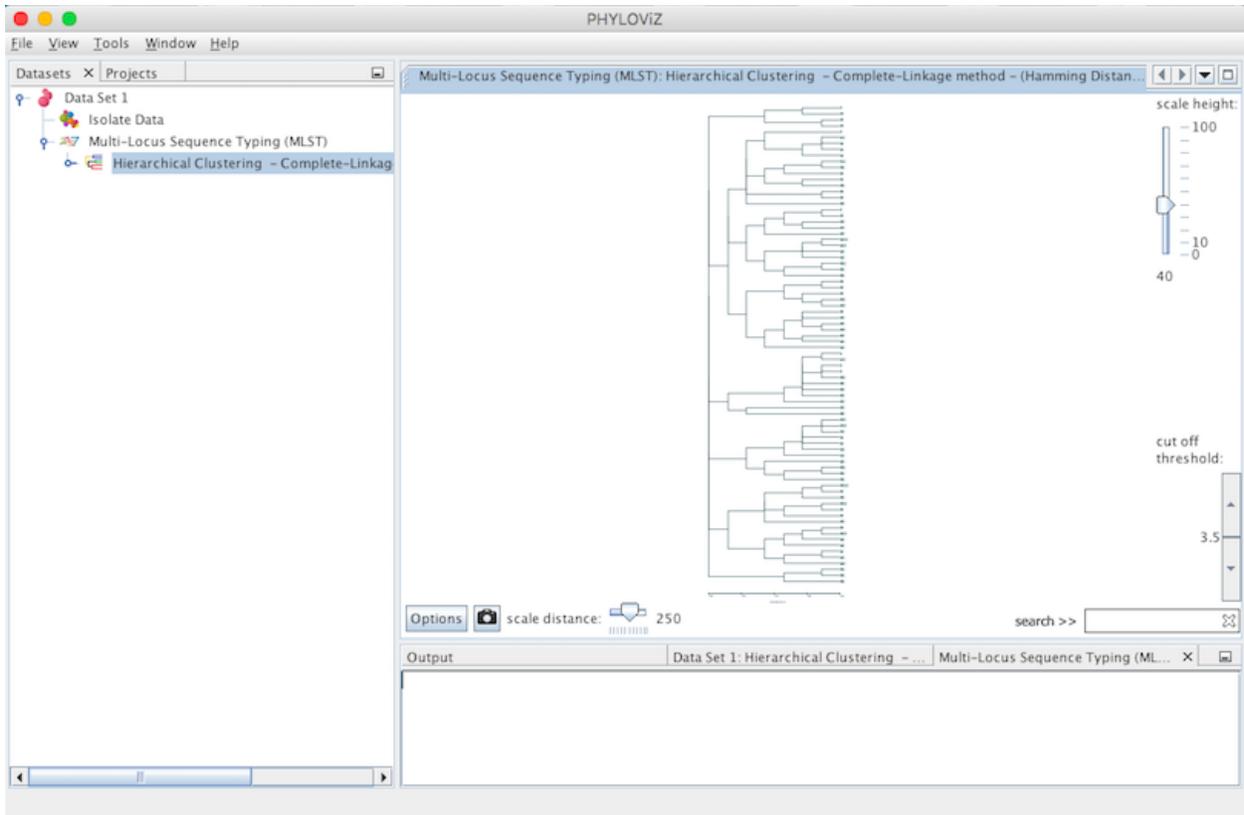
In order to display the dendrogram view, it is necessary to expand the typing data on the Datasets' tab, if it is not already expanded.



It should appear an icon corresponding to the hierarchical clustering computation



Double clicking on the Hierarchical Clustering item will show the display. This type of clustering is represented in the format of a dendrogram. The following screenshot summarizes the output for the previous dataset. Sometimes it is necessary to fit the image to see all the display at once. To do this, please right click on the mouse over the display.



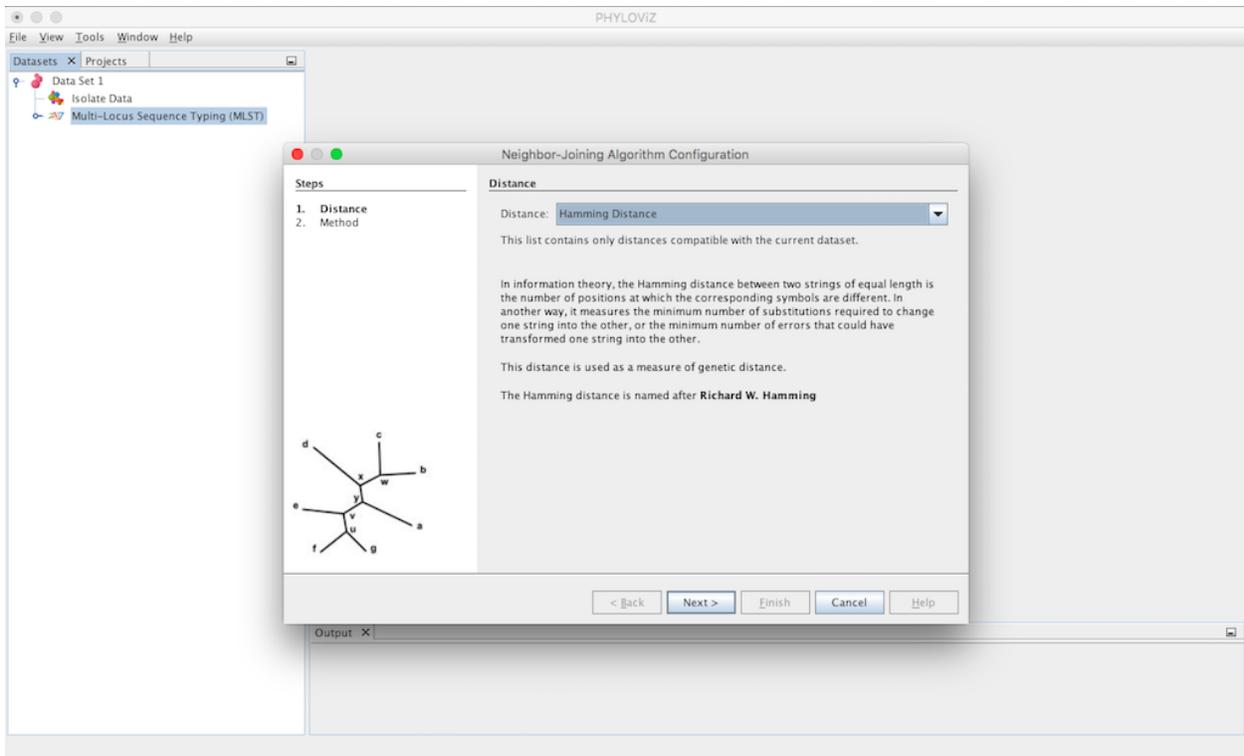
Some features were added to the visualization to improve and facilitate the analysis. These features are the following:

1. Height scale
2. Width scale
3. Options Panel
4. Search ST
5. Filter by distance (cut off threshold)
6. Export image

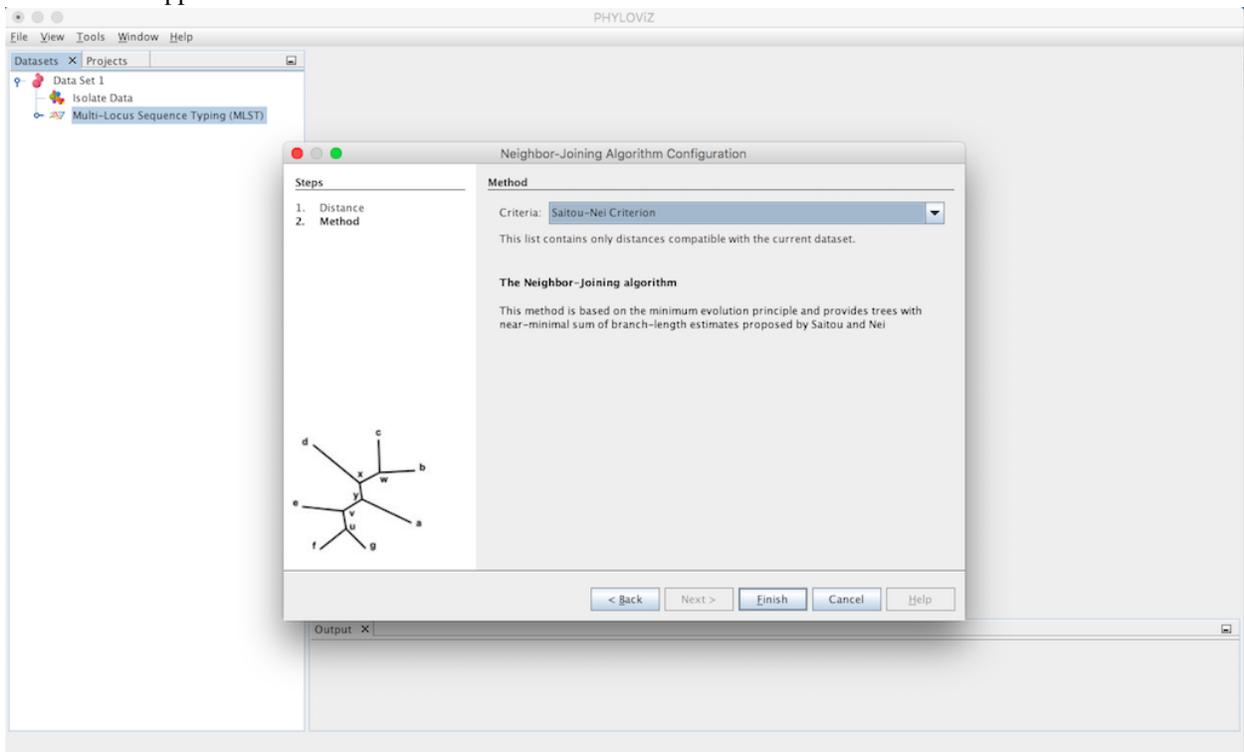
See section display and visualization for more information on these features.

3.4 Neighbor Joining

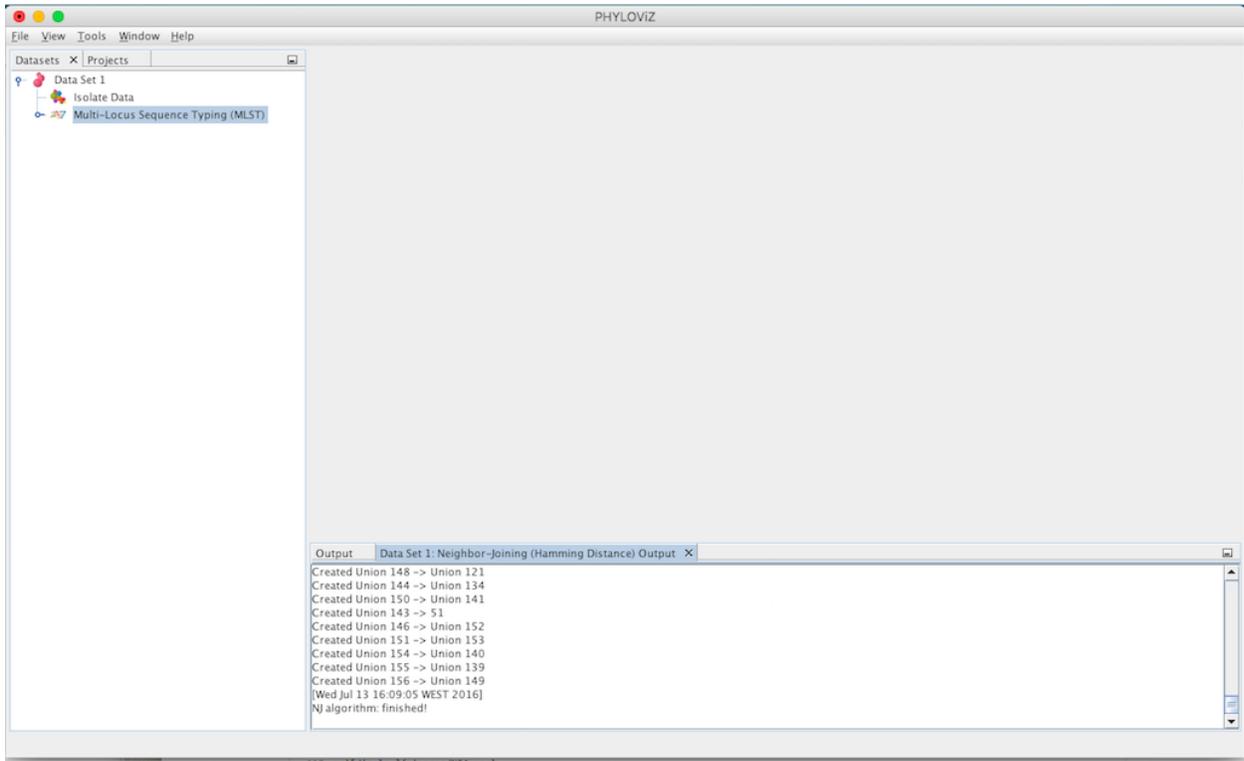
Selecting the Neighbor Joining algorithm opens the dialog where you can select what method you want to apply. The first step is choosing the *Distance* to be used.



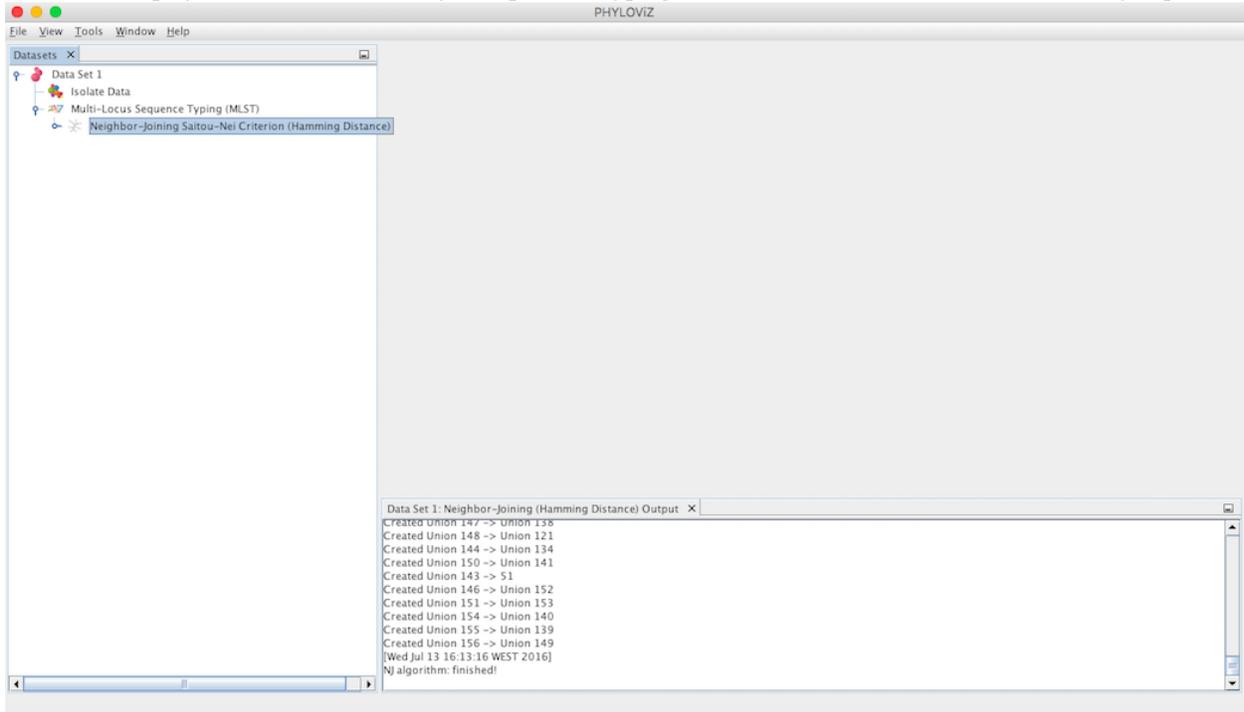
The second step is to select the *Criteria* of the tree branch-length minimization. You can choose between Saitou-Nei and Studier-Keppler criterion.



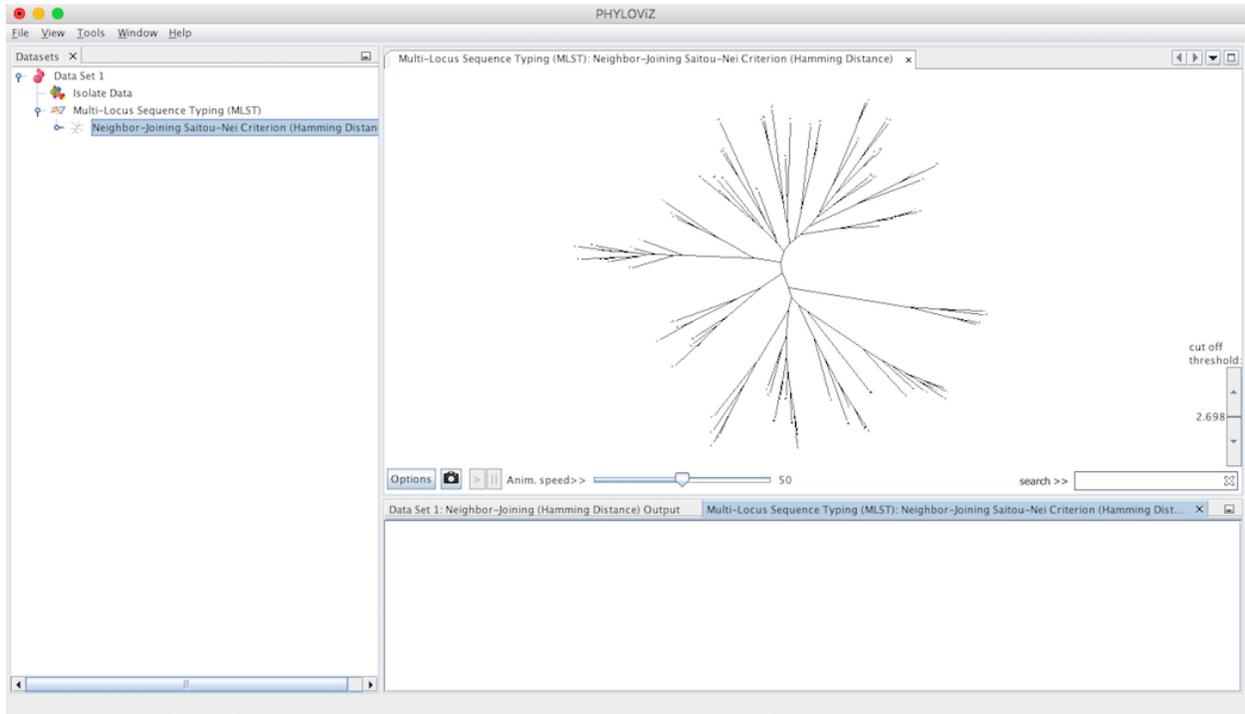
A Neighbor Joining *Output Tab* will appear and display the results of the application of the chosen method. The information displayed represents the same as the Hierarchical Clustering *Output Tab*.



In order to display the view, it is necessary to expand the typing data on the Dataset's tab, if it is not already expanded.



Double clicking on the Neighbor Joining item will show the display. By default it is represented in the format of a radial tree. The following screenshot summarizes the output for the previous dataset.



Some features were added to the visualization to improve and facilitate the analysis. These features are the following:

1. Options Panel that includes changing the tree layout
2. Search ST
3. Filter by distance
4. Export image

Display and visualization

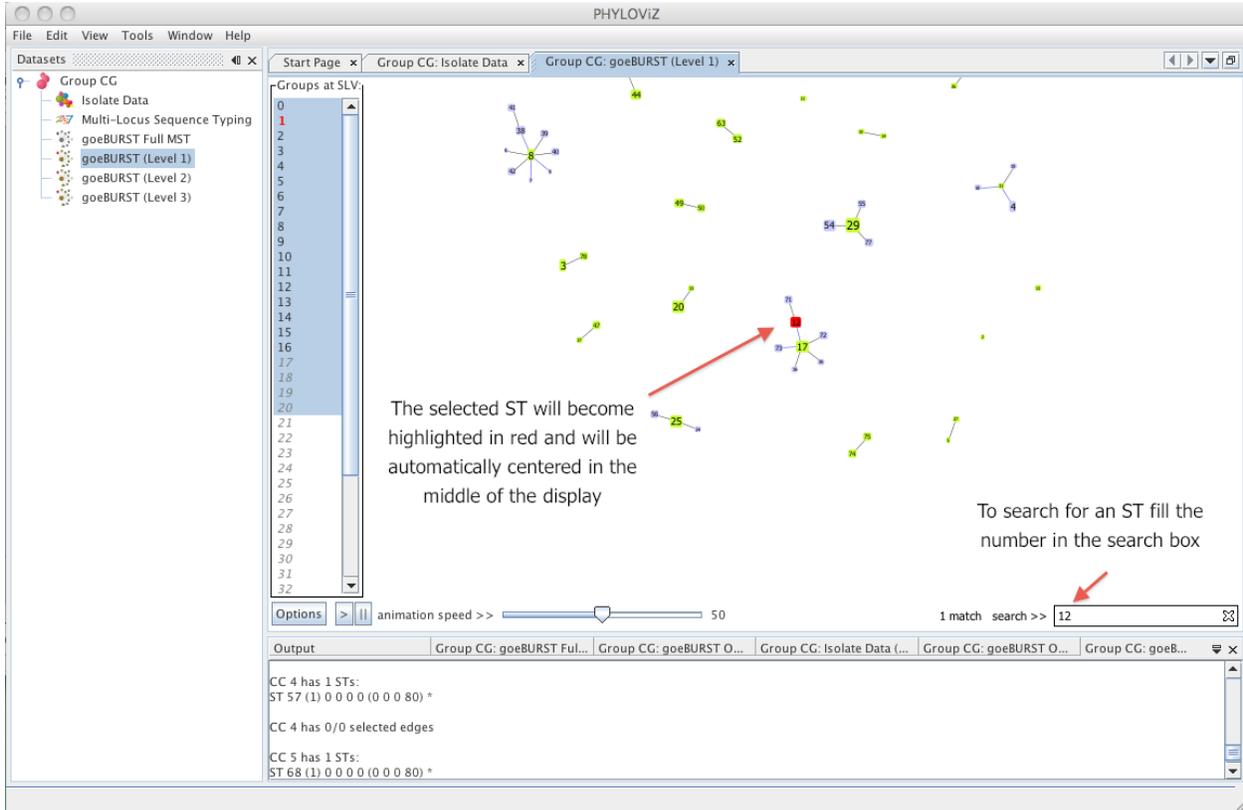
4.1 Interface features

After running the selected algorithm, you will notice that the program then tries to optimize the display of the group with the largest number of elements in the data set. You can change the speed at which this occurs by moving the animation speed slider.

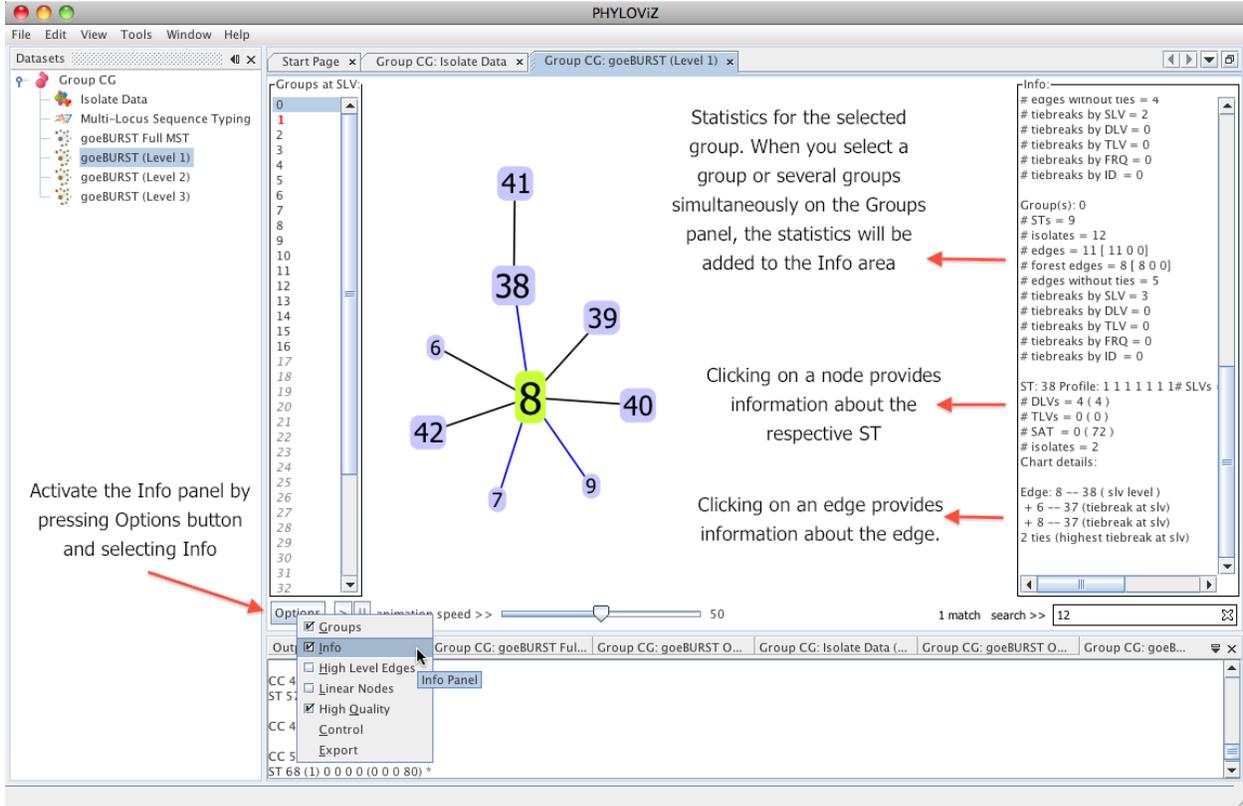
The Display tab offers the user the ability to search for an isolate, Highlight the SLVs and DLVs, control the animation speed, select different different or multiple groups. You can fit any displayed graphs to the window by right-clicking any open space (i.e. with no link or ST node) on the window.

4.1.1 Common features

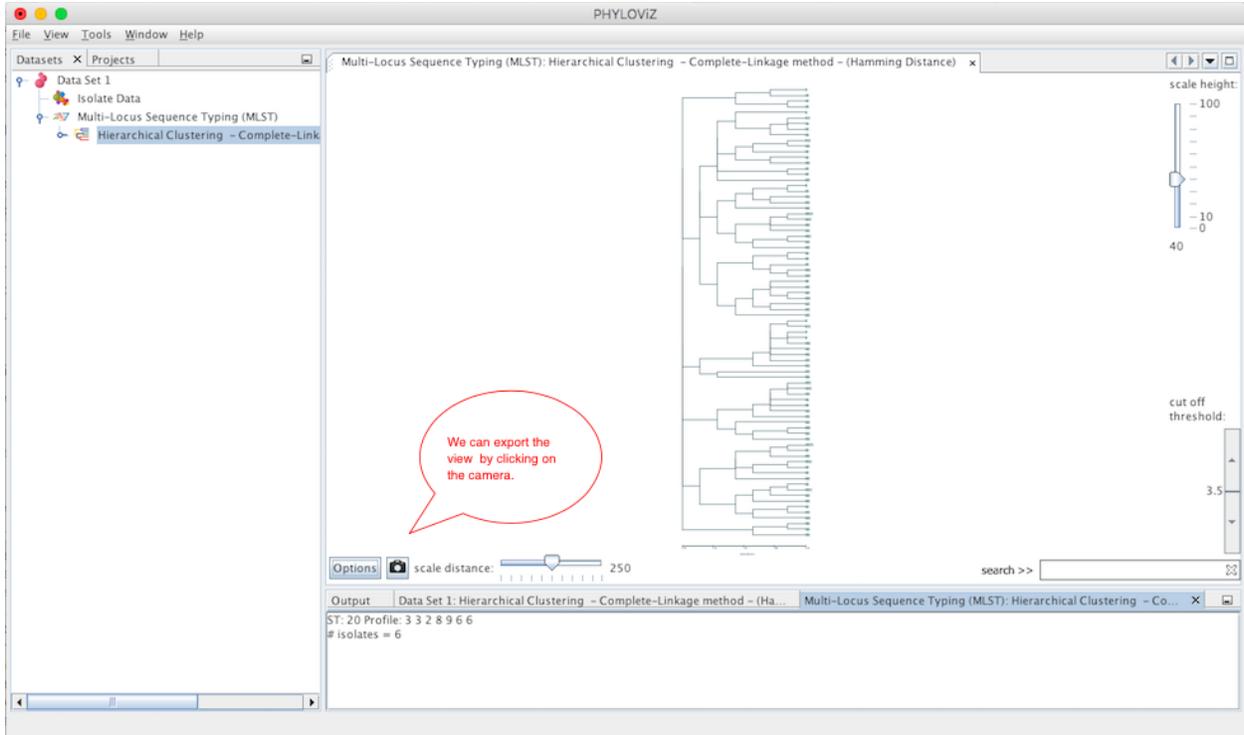
- ST search



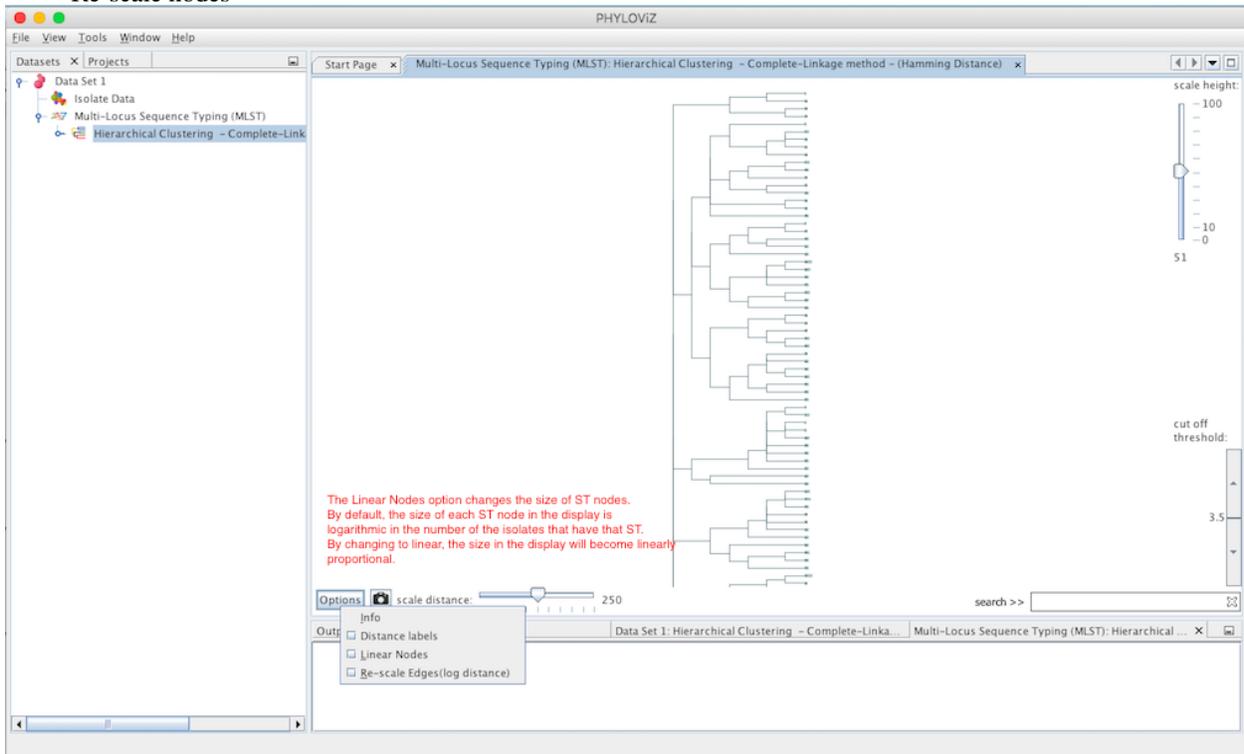
• Information Area



• Exporting Image



- Re-scale nodes



4.1.2 GoeBURST and GoeBURST Full MST features

- Basic Interface

File Edit View Tools Window Help

Datasets

- Group CG
 - Isolate Data
 - Multi-Locus Sequence Typing
 - goeBURST Full MST
 - goeBURST (Level 1)
 - goeBURST (Level 2)
 - goeBURST (Level 3)

Start Page x Group CG: Isolate Data x Group CG: goeBURST (Level 1) x

Groups at SLV

0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Options > || animation speed >> 50

Output

Group CG: goeBURST Ful... Group CG: goeBURST O... Group CG: Isolate Data (... Group CG: goeBURST O... Group CG: goeB...

CC 4 has 1 STs:
ST 57 (1) 0 0 0 0 (0 0 0 80) *

CC 4 has 0/0 selected edges

CC 5 has 1 STs:
ST 68 (1) 0 0 0 0 (0 0 0 80) *

Annotations:

- Indication of what level the Clonal Complexes were calculated
- Tabs with the data and algorithm results. To close the tab press the x symbol. To reopen double-click on the Data set tree.
- To manually change an ST position, click on the ST with the left mouse button and drag to the selected position.
- Search box: Insert here the ST ID you want to search
- Algorithm output tabs
- Animation Play
- Animation Pause
- Animation speed scrollbar
- Groups are ordered from the largest to the smallest; Singletons are in grey and italic

• SLV/DLV highlighting

File Edit View Tools Window Help

Datasets

- Group CG
 - Isolate Data
 - Multi-Locus Sequence Typing
 - goeBURST Full MST
 - goeBURST (Level 1)
 - goeBURST (Level 2)
 - goeBURST (Level 3)

Start Page x Group CG: Isolate Data x Group CG: goeBURST (Level 1) x

Groups at SLV

0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Options > || animation speed >> 50

Output

Group CG: goeBURST Ful... Group CG: goeBURST O... Group CG: Isolate Data (... Group CG: goeBURST O... Group CG: goeB...

CC 4 has 1 STs:
ST 57 (1) 0 0 0 0 (0 0 0 80) *

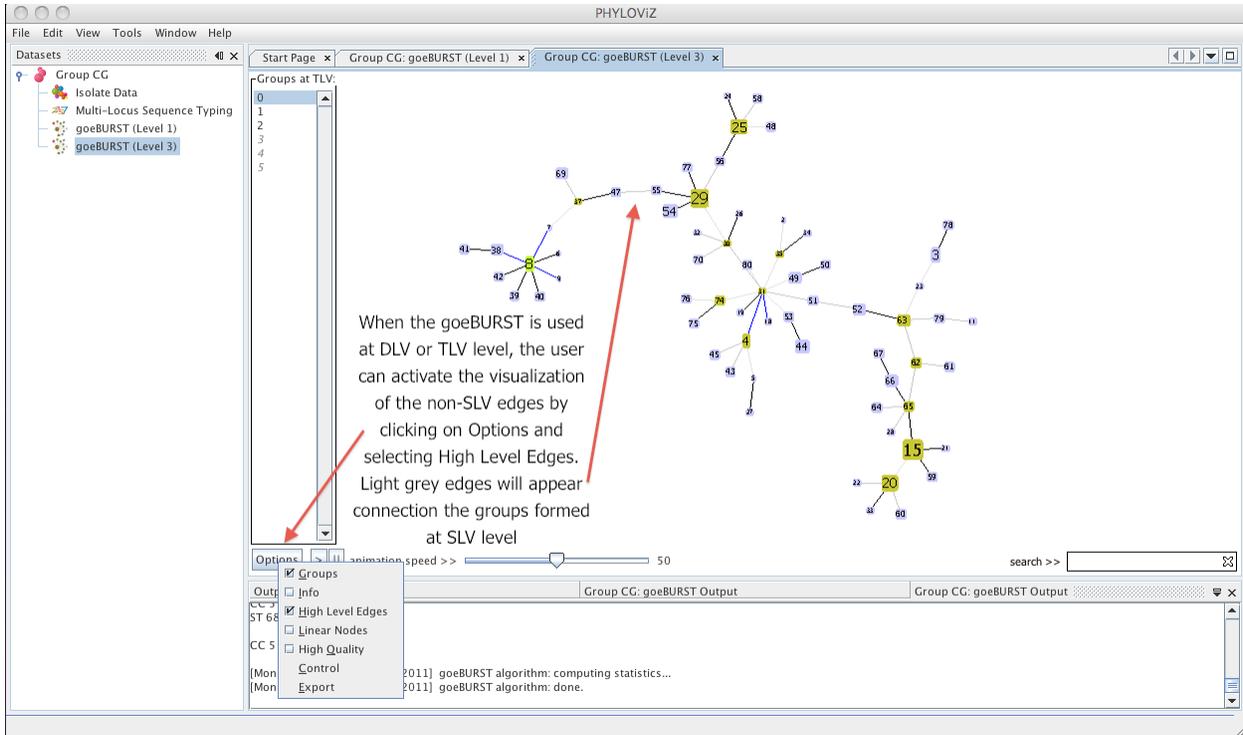
CC 4 has 0/0 selected edges

CC 5 has 1 STs:
ST 68 (1) 0 0 0 0 (0 0 0 80) *

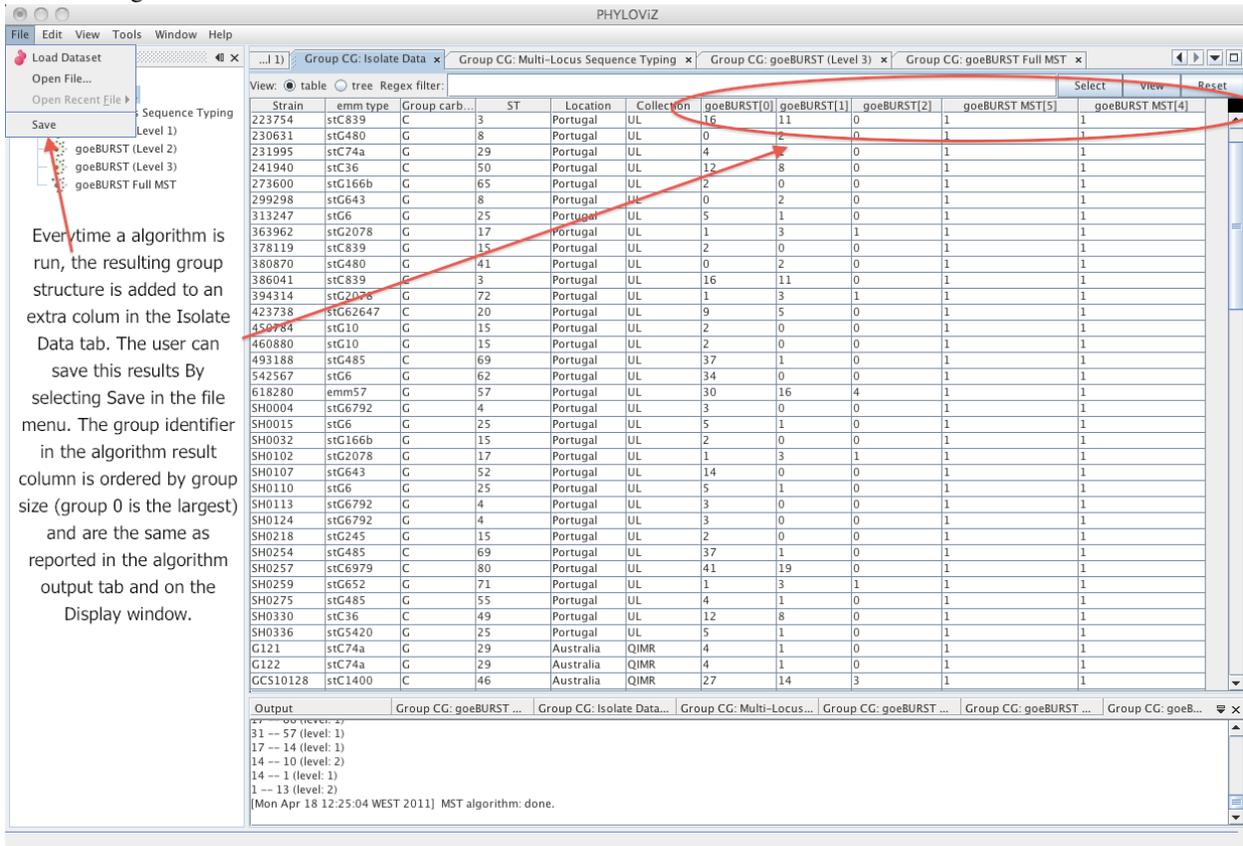
Annotations:

- To select a node just hover the mouse pointer to the node of interest. The selected node is represented in red; All the SLVs of the selected node are represented in orange and the DLVs are represented in yellow.

• High Level Edges

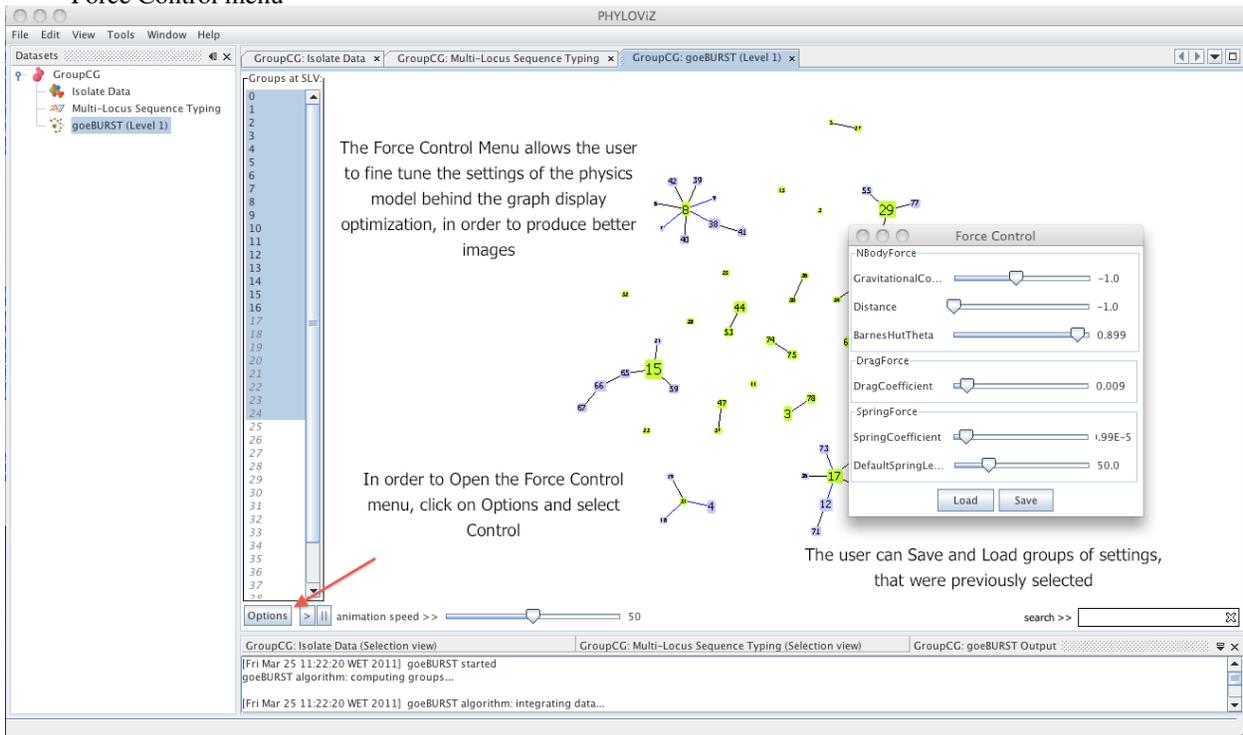


• Saving Results



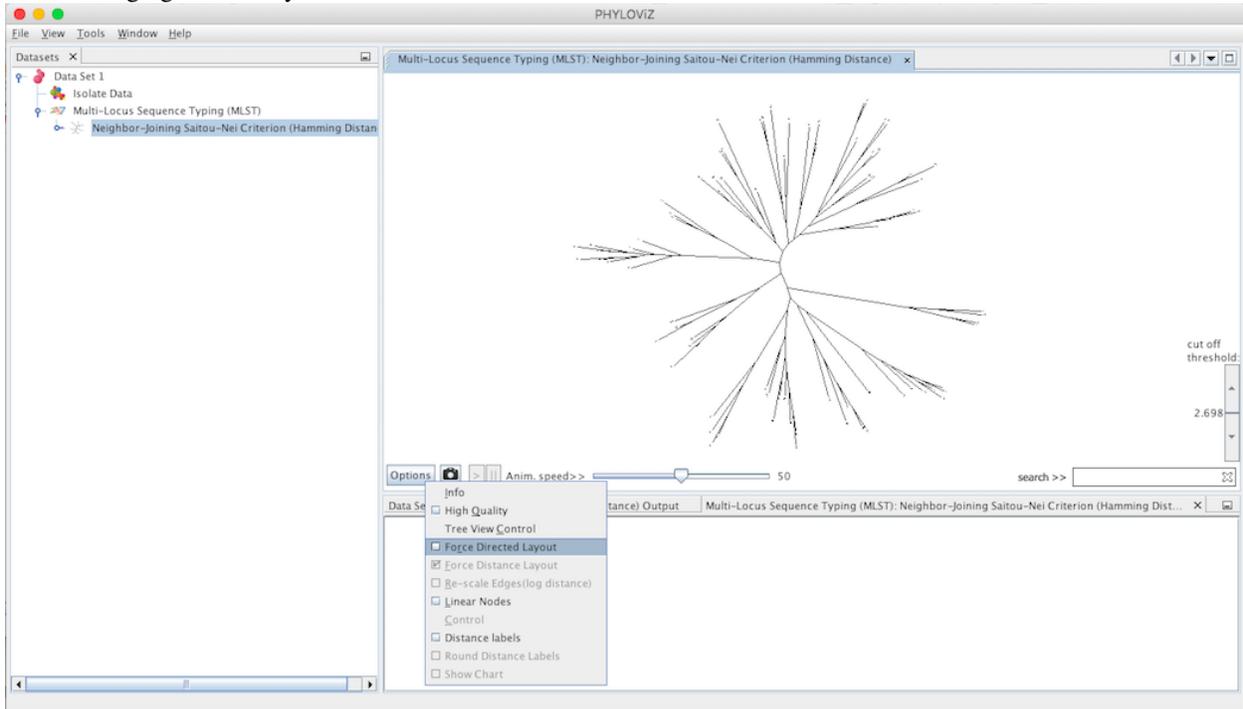
4.1.3 GoeBURS, GoeBURST Full MST and Neighbor Joining features

- Force Control menu



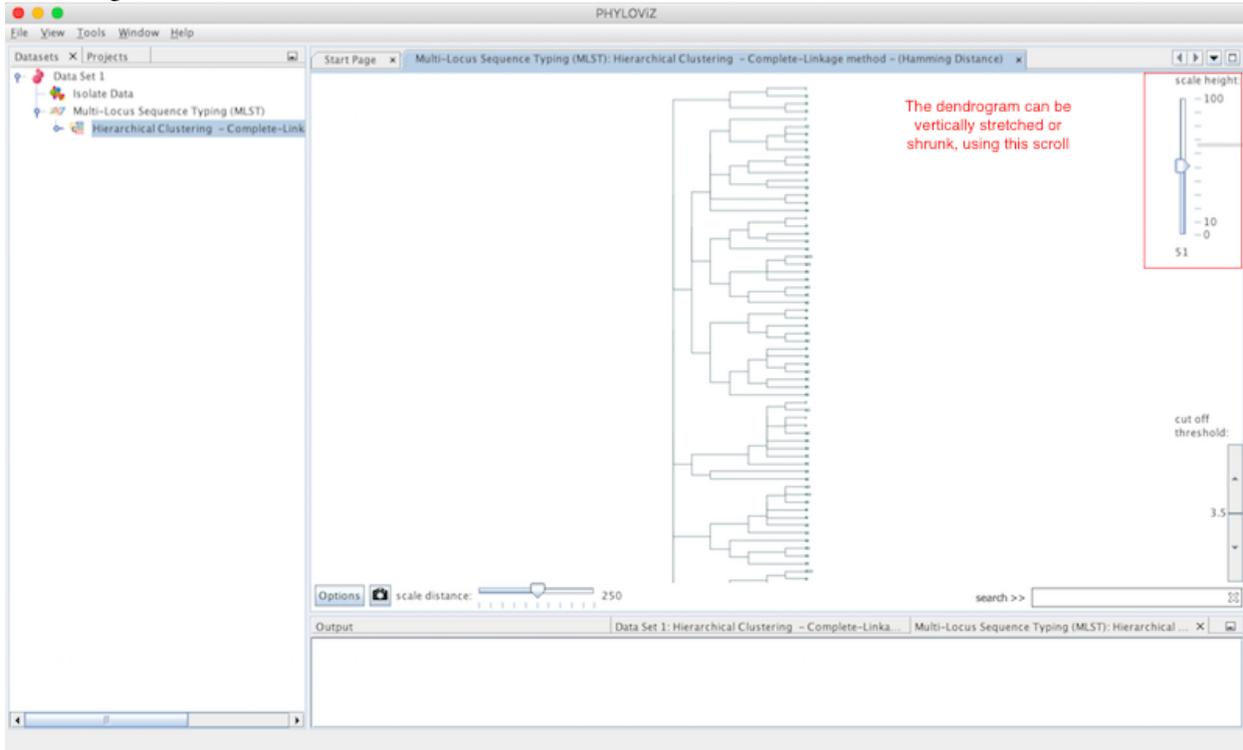
4.1.4 Neighbor Joining features

- Changing the tree layout

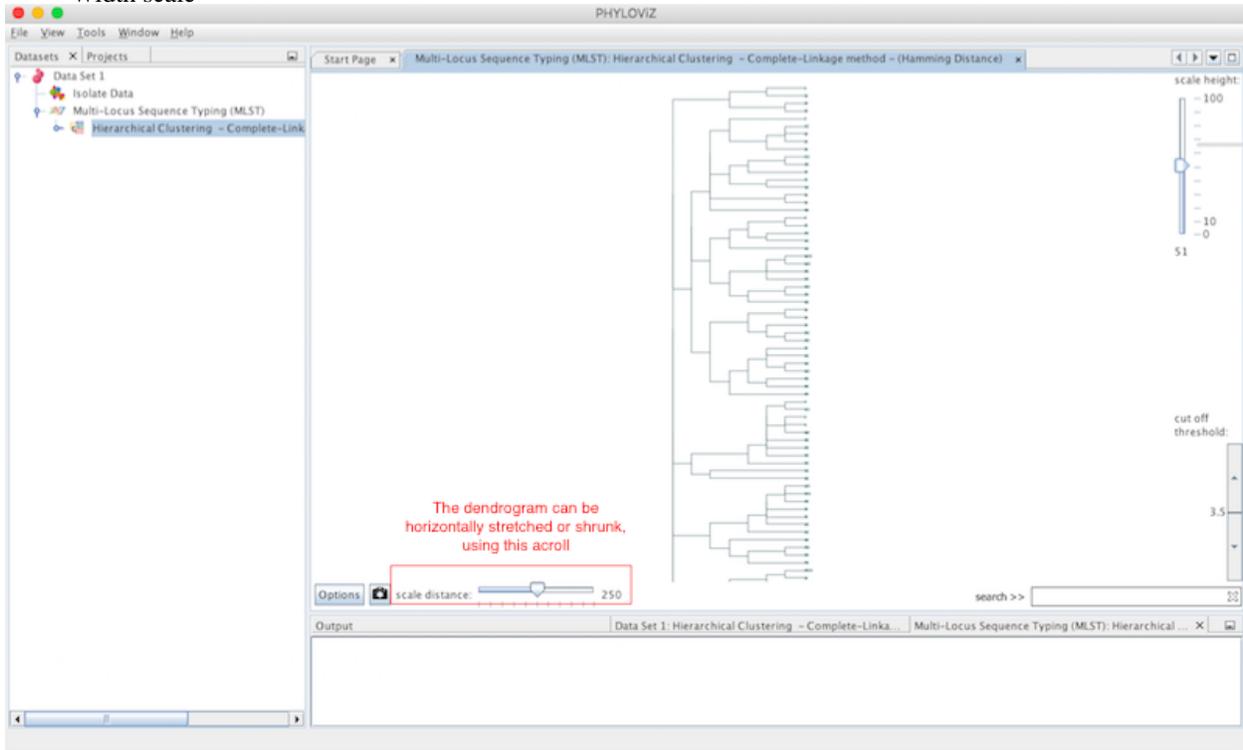


4.1.5 Hierarchical Clustering and Neighbor Joining features

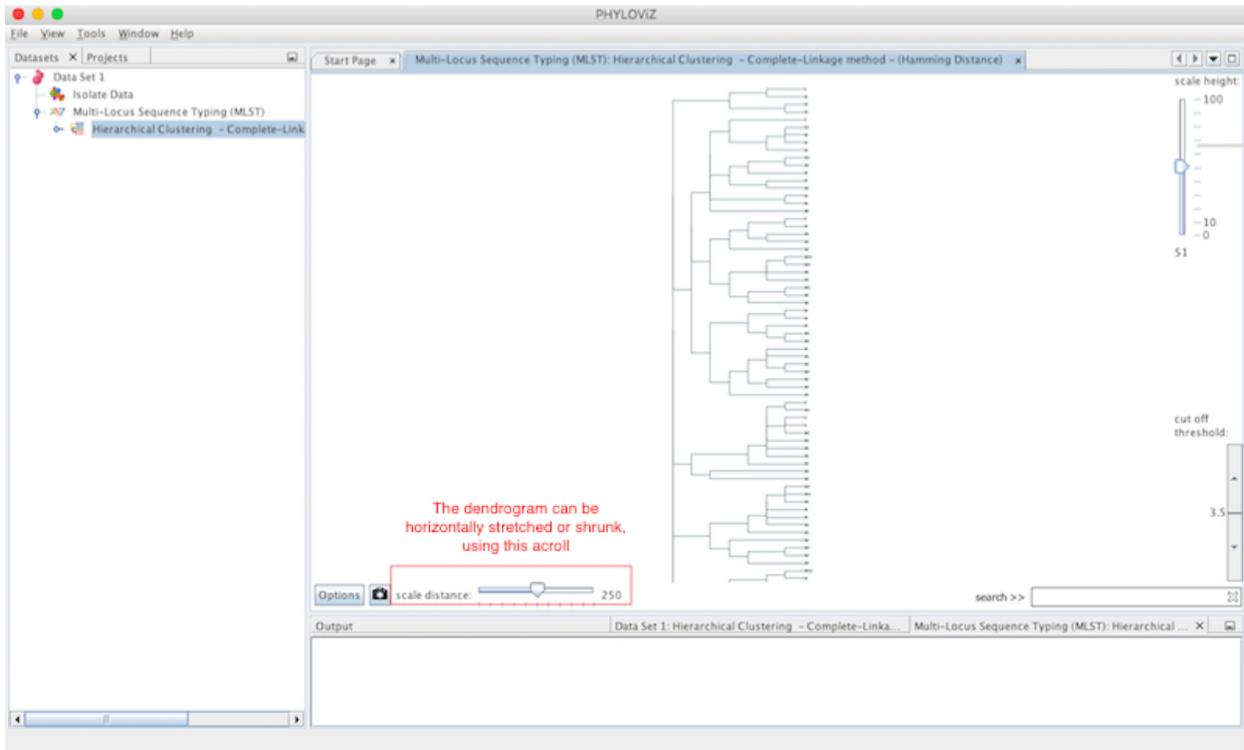
- Height scale



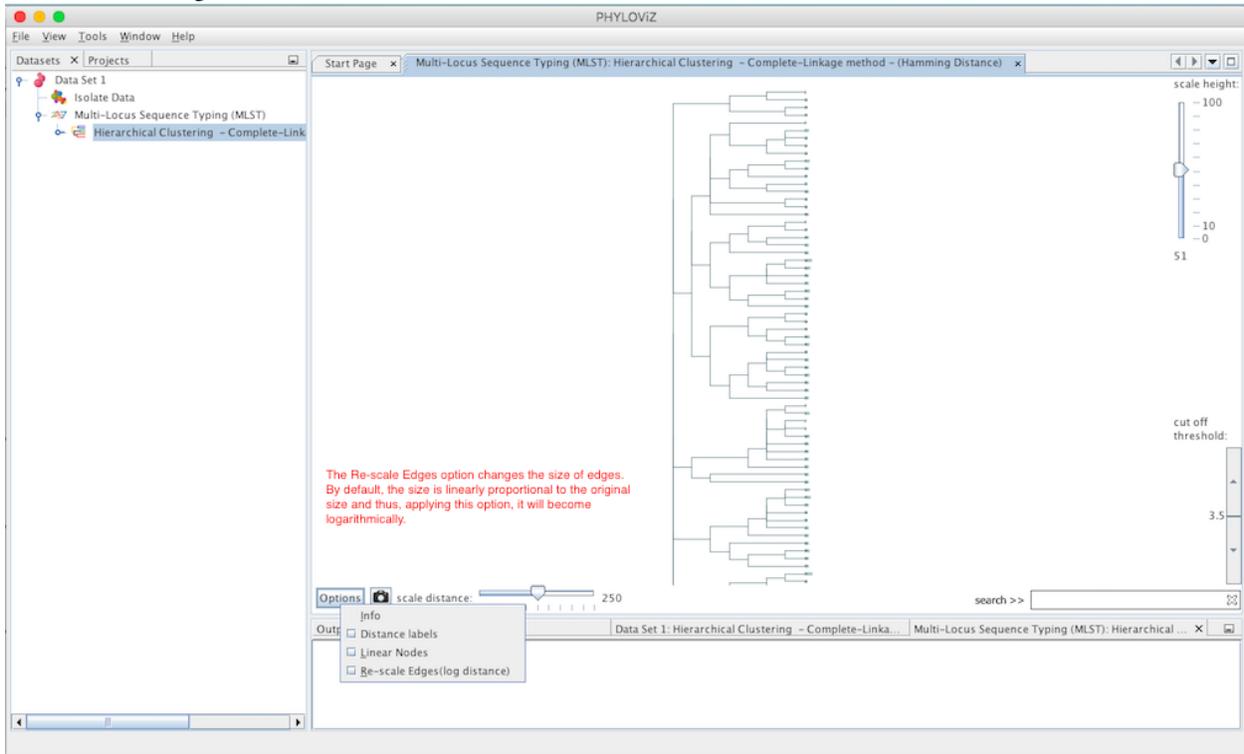
- Width scale



- Filter by distance (cut off threshold)



- Re-scale edges



4.2 Color conventions

Link colors for goeBURST results:

- Black - Link drawn without recourse to tiebreak rules,
- Blue - Link drawn using tiebreak rule 1 (number of SLVs),
- Green - Link drawn using tiebreak rule 2 (number of DLVs),
- Red - Link drawn using tiebreak rule 3 (number of TLVs),
- Yellow - Link drawn using tiebreak rule 4 or 5 (Frequency found on the data set and ST number , respectively),
- Gray - Links drawn at DLV (darker gray) or TLV (lighter gray) if the groups are constructed at DLV/TLV level.

Link colors for goeBURST Full MST results: The goeBURST Full MST algorithm links uses a grayscale with darker links having less differences between the profiles than the lighter gray links. To know the number of differences that the link represents click on the link in the Display window.

ST nodes colors:

- Light green - Group founder
- Dark green - Sub-group founder
- Light blue - Common node
- Red - Selected node

Querying and visualizing the data

The main goal of PHYLOViZ is to provide a data visualization tool for overlaying accessory data on the data analysis algorithms result. This allows to test the method's adequacy to the data, or the proposal of novel hypothesis. This section will explain the basics on how this can be achieved in our software. The user can query the data using regular expressions, or simply manually selecting the desired fields from the table or, even just use the checkboxes in the tree view. Using your dataset and this instructions you should be able to create visualizations similar to the ones found in the [PHYLOViZ website](#).

5.1 The isolate data tab

The *Isolate Data* tab is displayed by double clicking on the *Isolate Data* on the *Dataset* tree. The following screenshot resumes the basic functionality of the display on the table view.

The screenshot shows the PHYLOViZ software interface. The main window displays a table view of data. The table has the following columns: Strain, Imm type, Group carb..., ST, Location, Collection, goeBURST[0], goeBURST[1], goeBURST[2], goeBURST MST(5), and goeBURST MST(1). The 'Strain' column contains values like 223754, 230637, 231955, etc. The 'Imm type' column contains 'C' or 'G'. The 'Group carb...' column contains values like 3, 8, 29, etc. The 'ST' column contains values like 'Portugal', 'Portugal', 'Portugal', etc. The 'Location' column contains 'Portugal', 'Portugal', 'Portugal', etc. The 'Collection' column contains 'UL', 'UL', 'UL', etc. The 'goeBURST' columns contain numerical values. The 'goeBURST MST' columns contain binary values (0 or 1).

Annotations in the screenshot include:

- Double click on the *Isolate Data* to active the *Isolate Data* tab. The user can choose to visualize the dataset as a table or a tree.
- Column headers
- The Regex filter allows the creation of complex queries
- Left clicking on a column header will sort the column in ascending or descending alphabetical order.
- Right clicking on a table header will select that field for limiting the regex queries to it. By pressing the Select button will also select all entries in the selected column. After the selection is performed, the user can press View to plot the selection onto the resulting algorithm graphs. Reset will clear all the selections made.

5.2 The typing data tab

The *Typing Data* tab contains the allelic profiles loaded in the dataset. The name of displayed on the tab, and on the *Dataset* tree, is the name of the selected method during the *Load Dataset* procedure. The user can also query, select and visualize the data of the allelic profiles, similarly to operations describe in the *Isolate Data* tab.

The screenshot shows the PHYLOViZ software interface. The main window displays a table of allelic profiles with columns for ST, gki, gtr, murl, mutS, recP, xpt, and yqiZ. The table contains 41 rows of data. Below the table, there is an 'Output' section and a legend for the 'recP' field, showing three categories: recP:9 (n = 28, 23.93%), recP:7 (n = 25, 21.37%), and recP:1 (n = 14, 11.97%).

ST	gki	gtr	murl	mutS	recP	xpt	yqiZ
6	1	3	1	1	1	4	
7	1	1	1	1	1	3	
8	1	1	1	1	1	4	
9	1	1	1	1	1	2	
22	3	3	2	8	1	11	6
38	1	1	1	1	1	1	
39	1	1	1	1	1	21	4
40	1	1	1	4	1	1	4
41	1	1	1	9	1	1	1
42	1	1	4	1	1	1	4
43	2	2	4	1	1	1	2
57	3	2	8	9	1	16	2
48	3	2	1	2	10	4	2
51	3	2	4	1	11	1	2
52	3	2	4	1	11	3	5
63	3	3	4	1	11	3	5
1	10	6	6	6	12	13	8
5	2	2	4	1	12	12	7
10	10	4	7	7	12	13	8
13	10	5	6	6	12	13	9
14	10	4	7	6	12	13	8
27	2	2	4	1	13	12	7
34	3	7	4	1	14	15	10
35	3	7	4	1	14	15	2
24	3	2	1	5	15	4	3
28	3	3	4	2	16	14	2
12	4	4	5	2	17	6	2
16	4	4	1	2	17	1	2
17	4	4	1	2	17	6	2
36	4	4	1	2	17	6	3
21	4	3	5	2	17	6	2

5.3 Regular expression primer

Some basic regular expressions that can be used in PHYLOViZ. For more complex expressions there are extensive tutorials on regular expressions online. Just search Regular Expression or regex.

- . (period mark) - represents any character.
- [] (square brackets) - Match anything inside the square brackets for one character position once and only once. Examples: [40] will match any field with 4 or 0; [7-9] will match any field will 7, 8 or 9 (- is the range separator).
- ^ (caret) - Starts with. Ex: ^P will give you all the fields that start with a P. Inside the square brackets means negation. Example [^a-c] means anything not a, b or c.
- \$ (dollar sign) - Ends with. Ex. 7\$ will give you all fields that end in a 7.
- ? (question mark) - Matches the preceding character 0 or 1 times only. Example: colour?r will find color and colour.
- * (asterisk) - Matches the preceding character 0 or more times. Example: tre* would find tree, tread and trough.
- + (plus) - Matches the preceding character 1 or more times. Example: tre+ would find tree, tread but not trough.

- {n} (any integer between brackets) - Matches the preceding character exactly n times. Example: AT[GC]{2} would match ATGC, ATCG, ATGG or ATCC but not ATGA.

All these operators can be combined to create complex search expressions. For example: `^st[G|C].*6$` would find any field that starts with `st` followed by a `C` or a `G` then as 0 or more characters and ends with a `6`. The following screenshot shows the result on the test dataset:

The screenshot shows the PHYLOViZ interface with the following components:

- Table View:** A table with columns: Strain, emm type, Group carb..., ST, Location, Collection, goeBURST[0], goeBURST[1], goeBURST[2], goeBURST MST[5], goeBURST MST[4]. The 'emm type' column is selected. The filter `^st[G|C].*6$` is applied.
- Text Box:** A text box explaining the query: "Complex queries to the dataset can be made by using regular expressions (REGEX). In this query, we select all of the strains which have emm type (notice the column header that is selected) that start with 'st', are followed by G or C, and end in 6."
- Legend:** A legend showing three categories: recP:9 (n = 28, 23.93%), recP:7 (n = 25, 21.37%), and recP:1 (n = 14, 11.97%).

Strain	emm type	Group carb...	ST	Location	Collection	goeBURST[0]	goeBURST[1]	goeBURST[2]	goeBURST MST[5]	goeBURST MST[4]
241940	stC36	C	50	Portugal	UL	12	8	0	1	1
313247	stG6	G	25	Portugal	UL	5	1	0	1	1
542567	stG6	G	62	Portugal	UL	34	0	0	1	1
GC506ny	stC36	C	49	USA	NYMC	12	8	0	1	1
GC509ny	stC36	C	68	USA	NYMC	36	17	5	1	1
GG510b	stG6	G	44	Australia	QIMR	13	9	0	1	1
GG520ny	stC36	G	45	USA	NYMC	26	0	0	1	1
GG521ny	stG6	G	52	USA	NYMC	14	0	0	1	1
GG524	stG6	G	44	Australia	QIMR	13	9	0	1	1
MD04	stG6	G	25	Australia	QIMR	5	1	0	1	1
MD05	stG6	G	63	Australia	QIMR	14	0	0	1	1
MD07	stG6	G	58	Australia	QIMR	31	1	0	1	1
MD834	stC36	G	4	Australia	QIMR	3	0	0	1	1
NS752	stG6	G	44	Australia	QIMR	13	9	0	1	1
SH0015	stG6	G	25	Portugal	UL	5	1	0	1	1
SH0110	stG6	G	25	Portugal	UL	5	1	0	1	1
SH0330	stC36	C	49	Portugal	UL	12	8	0	1	1

5.4 Queries using the table view

In the *Table* view of the *Data* tab you can manually select any field you want to represent by left clicking on it. That will automatically display all the entries with the selected value and not only the selected ones. To select multiple fields you can press the CTRL key (or CMD on Mac) while clicking on the desired fields. If you keep the SHIFT key pressed you can select ranges of cells.

You can also automatically select multiple columns by clicking with the right mouse button on the table headers and pressing the *Select* button.

Finally to plot the data on the *Display* tab, press the *View* button, after all the desired selections are performed.

5.4.1 Query examples

- Table view with selections

PHYLOViZ

File Edit View Tools Window Help

Datasets

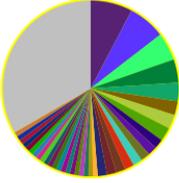
Start Page x Group CG: goeBURST (Level 1) x Group CG: Isolate Data x Group CG: Multi-Locus Sequence Typing x

View: table tree Regex filter: ^UL[2[0-9]{1}]

Strain	emm type	Group carb...	ST	Location	Collection	goeBURST[0]	goeBURST[1]	goeBURST[2]	goeBURST MST[5]	goeBURST MST[4]
CGS15ny	stC74a	G	29	Other	NYMC	4	1	0	1	1
CGS16ny	stG485	G	29	Other	NYMC	4	1	0	1	1
G121	stC74a	G	29	Australia	QIMR	4	1	0	1	1
G122	stC74a	G	29	Australia	QIMR	4	1	0	1	1
GCS2816	stG62647	C	20	Australia	QIMR	9	5	0	1	1
GCS6894	stG62647	C	20	Australia	QIMR	9	5	0	1	1
GCS6929	stG62647	C	20	Australia	QIMR	9	5	0	1	1
CGS11172	stC74a	G	29	Australia	QIMR	4	1	0	1	1
CGS539813	stC74a	G	29	Australia	QIMR	4	1	0	1	1
CGS540048	stG485	G	29	Australia	QIMR	4	1	0	1	1
CGS9225	stC74a	G	29	Australia	QIMR	4	1	0	1	1
MD04	stG6	G	25	Australia	QIMR	5	1	0	1	1
MD06	stC74a	G	29	Australia	QIMR	4	1	0	1	1
MD122	stC74a	G	29	Australia	QIMR	4	1	0	1	1
MD227	stC6979	C	20	Australia	QIMR	9	5	0	1	1
MD605	stG62647	C	20	Australia	QIMR	9	5	0	1	1
MD722	stC74a	G	29	Australia	QIMR	4	1	0	1	1
MD934	stG5420	G	25	Australia	QIMR	5	1	0	1	1
168554	stG485	G	47	Portugal	UL	11	1	0	1	1
171712	stG480	G	38	Portugal	UL	0	2	0	1	1
220269	stG2078	G	15	Portugal	UL	2	0	0	1	1
223754	stC839	C	3	Portugal	UL	16	11	0	1	1
230631	stG480	G	8	Portugal	UL	0	2	0	1	1
231995	stC74a	G	29	Portugal	UL	4	1	0	1	1
241940	stC36	C	50	Portugal	UL	12	8	0	1	1
273600	stG166b	G	65	Portugal	UL	2	0	0	1	1
299298	stG643	G	8	Portugal	UL	0	2	0	1	1

REGEX filter applied to the data set: select all entries from the dataset that start with UL (^UL) or (l) have any number with 2 digits starting with 2 (2[0-9]{1}). Since only the ST and Collection columns were selected (by right clicking on them) the search is only made on those columns. After writing the query the user presses the Select button to select all results (the blue background appears) and to view the results on the Display tabs, the user has to press the View button

Output



Legend

- ST:29 Collection:QIMR (n = 9, 7.6%)
- ST:15 Collection:QIMR (n = 8, 6.8%)
- ST:15 Collection:UL (n = 6, 5.13%)
- ST:20 Collection:QIMR (n = 5, 4.2%)
- ST:25 Collection:UL (n = 4, 3.42%)
- ST:17 Collection:QIMR (n = 3, 2.56%)
- ST:3 Collection:NYMC (n = 3, 2.56%)
- ST:4 Collection:UL (n = 3, 2.56%)
- ST:17 Collection:UL (n = 2, 1.71%)
- ST:25 Collection:QIMR (n = 2, 1.7%)

Load Save

- Query results *Chart*

PHYLOViZ

File Edit View Tools Window Help

Datasets

Start Page x Group CG: goeBURST (Level 1) x Group CG: Isolate Data x Group CG: Multi-Locus Sequence Typing x

View: table tree Regex filter: ^AUL2[0-9]{1}

Strain	emm type	Group carb...	ST	Location	Collection	goeBURST[0]	goeBURST[1]	goeBURST[2]	goeBURST MST[5]	goeBURST MST[4]
GGS15ny	stC74a	G	29	Other	NYMC	4	1	0	1	1
GGS16ny	stG485	G	29	Other	NYMC	4	1	0	1	1
G121	stC74a	G	29	Australia	QIMR	4	1	0	1	1
G122	stC74a	G	29	Australia	QIMR	4	1	0	1	1
GCS2816	stG62647	C	20	Australia	QIMR	9	5	0	1	1
GCS6894	stG62647	C	20	Australia	QIMR	9	5	0	1	1
GCS6929	stG62647	C	20	Australia	QIMR	9	5	0	1	1
GGS11172	stC74a	G	29	Australia	QIMR	4	1	0	1	1
GGS39813	stC74a	G	29	Australia	QIMR	4	1	0	1	1

Output

Group CG: goeBURST... Group CG: Isol... Group CG: goeBURST... Group CG: goeBURST... Group CG: goeBURST... Group CG: Multi-Locu...

Legend

- ST:29 Collection:QIMR (n = 9, 7.6%)
- ST:15 Collection:QIMR (n = 8, 6.8%)
- ST:15 Collection:UL (n = 6, 5.13%)
- ST:20 Collection:QIMR (n = 5, 4.27%)
- ST:25 Collection:UL (n = 4, 3.42%)
- ST:17 Collection:QIMR (n = 3, 2.56%)
- ST:3 Collection:NYMC (n = 3, 2.56%)
- ST:4 Collection:UL (n = 3, 2.56%)
- ST:17 Collection:UL (n = 2, 1.71%)
- ST:25 Collection:QIMR (n = 2, 1.71%)
- ST:29 Collection:NYMC (n = 2, 1.71%)
- ST:3 Collection:UL (n = 2, 1.71%)
- ST:69 Collection:UL (n = 2, 1.71%)
- ST:8 Collection:NYMC (n = 2, 1.71%)
- ST:8 Collection:QIMR (n = 2, 1.71%)
- ST:8 Collection:UL (n = 2, 1.71%)
- ST:15 Collection:NYMC (n = 1, 0.85%)
- ST:20 Collection:UL (n = 1, 0.85%)
- ST:29 Collection:UL (n = 1, 0.85%)
- ST:3 Collection:QIMR (n = 1, 0.85%)
- ST:38 Collection:NYMC (n = 1, 0.85%)
- ST:38 Collection:UL (n = 1, 0.85%)
- ST:4 Collection:QIMR (n = 1, 0.85%)
- ST:41 Collection:UL (n = 1, 0.85%)
- ST:47 Collection:UL (n = 1, 0.85%)
- ST:49 Collection:NYMC (n = 1, 0.85%)
- ST:49 Collection:UL (n = 1, 0.85%)
- ST:50 Collection:UL (n = 1, 0.85%)
- ST:52 Collection:NYMC (n = 1, 0.85%)
- ST:53 Collection:UL (n = 1, 0.85%)

All the combinations of the values are displayed in the chart with different colors automatically assigned.

The grey color ('others' in the chart legend) is reserved for all the combinations not found by the active selection.

In the figure legend it is also displayed for each combination, the absolute numbers of the isolates found and percentage of the dataset they represent.

- Results on *Display Tab*

When the user double clicks in an algorithm Display in the Dataset tree view, it opens the corresponding Display tab with the results of the query displayed on it.

Category colors can be changed by clicking on the color square in the legend. The defined colors can be saved and retrieved for use in future data analysis sessions.

The screenshot shows the PHYLOViZ application window with a menu bar (File, Edit, View, Tools, Window, Help) and several tabs. The 'Datasets' panel on the left shows a tree view with 'goeBURST (Level 1)' selected. The main area displays a network graph with nodes of various colors and sizes. Below the graph is an 'Options' panel with an 'animation speed' slider set to 50. At the bottom, there is an 'Output' panel with a 'Legend' section containing a list of categories and their counts, and 'Load' and 'Save' buttons.

5.5 Queries using the tree view

The *Tree* view offers a faster way to create simple queries. The user can also use the regex filter to search the dataset but all the possibilities for each dataset column are automatically indexed in a tree like manner. By pressing the *Select* button and switching to *Table* view the user can see the resulting selection. The users can alternate both views (Table and Tree) at will for creating the selection.

5.5.1 Query examples

- Tree view with selections

The screenshot shows the PHYLOViZ interface with the following components:

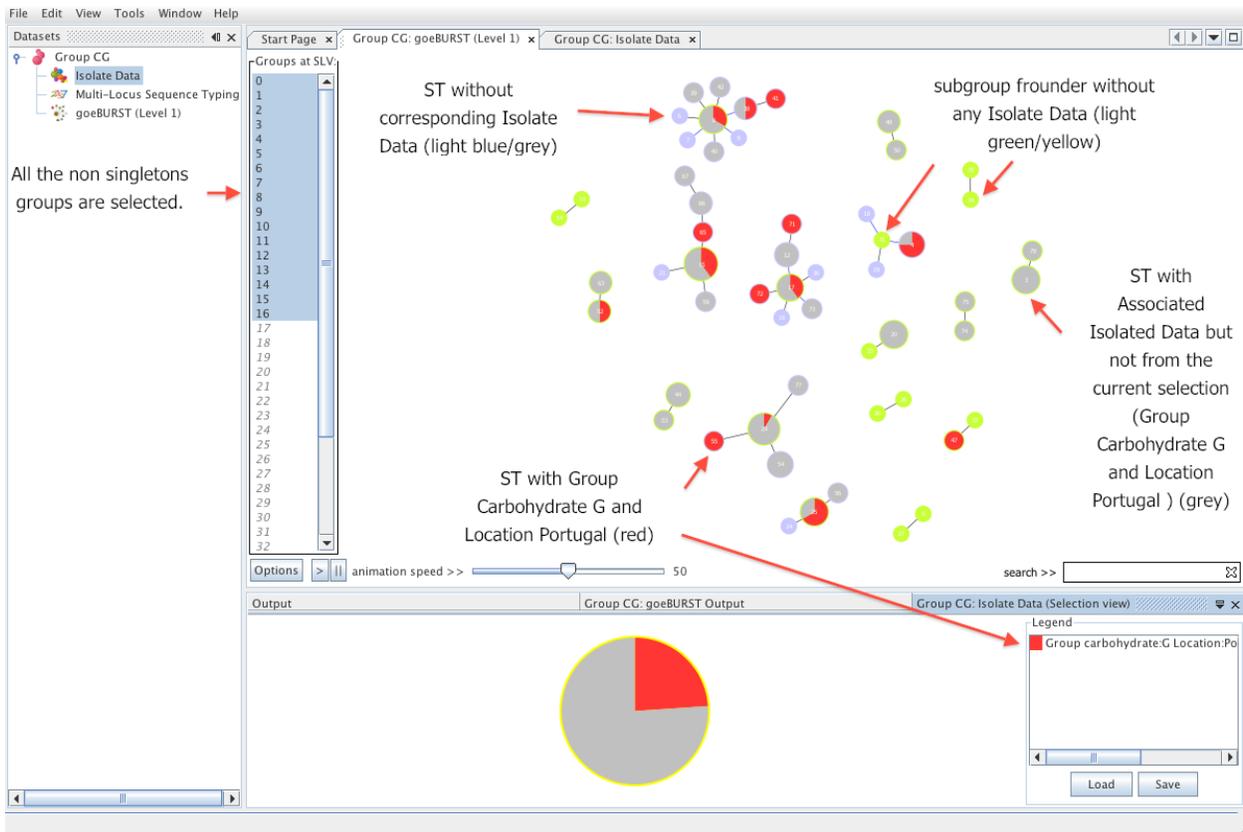
- Tree View:** A hierarchical tree on the left with the following structure:
 - Group CG
 - Isolate Data
 - Multi-Locus Sequence Typing
 - goeBURST (Level 1)
 - Strain
 - emm type
 - Group carbohydrate
 - G (checked)
 - C
- Location
 - Portugal (checked)
 - USA
 - Other
 - Australia
- Collection
- goeBURST[0]

Annotations:

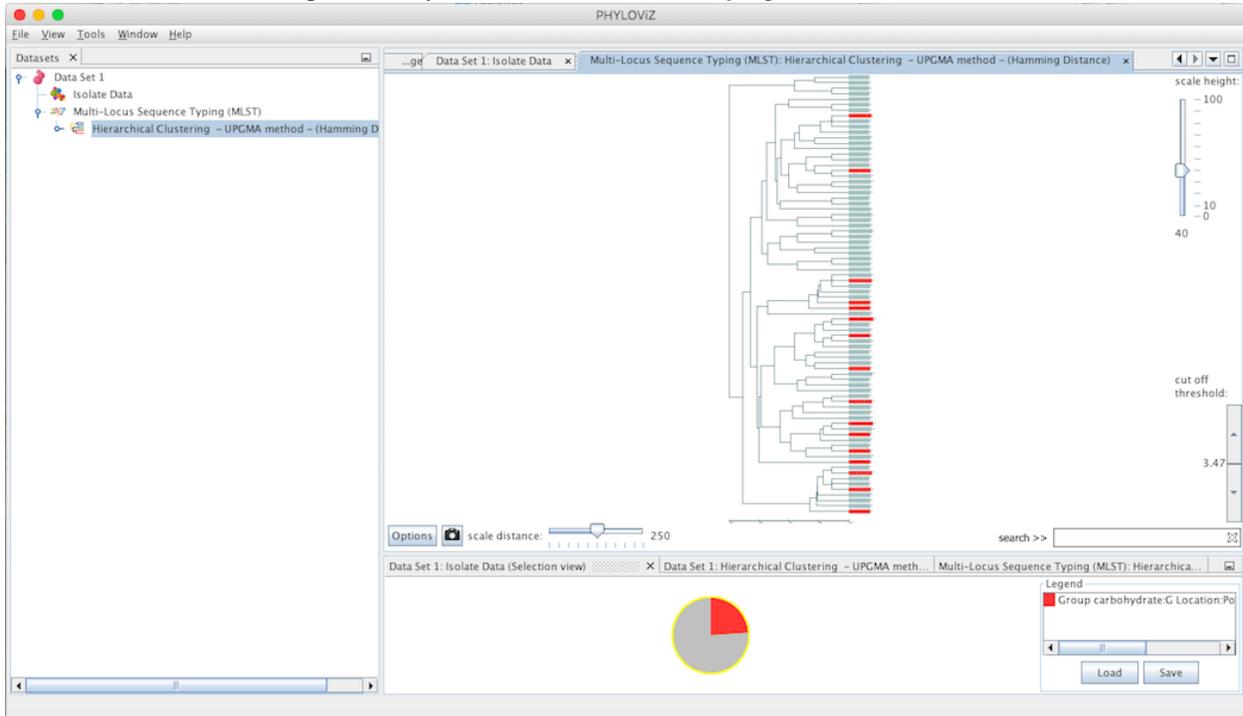
- Text: "This selection will choose all isolates that have Group Carbohydrate G and Location Portugal." with arrows pointing to the checked 'G' and 'Portugal' nodes.
- Text: "All the isolates with this resulted combination will be displayed in red. All other combinations will be displayed in grey. Isolates for which no data is available will use the default colors (light blue/grey except for subgroup founder that are light green/yellow)." with an arrow pointing to the red slice in the pie chart.

Output Window: Shows a pie chart with a small red slice and a legend titled "Legend" with the entry "Group carbohydrate:G Location:Po".

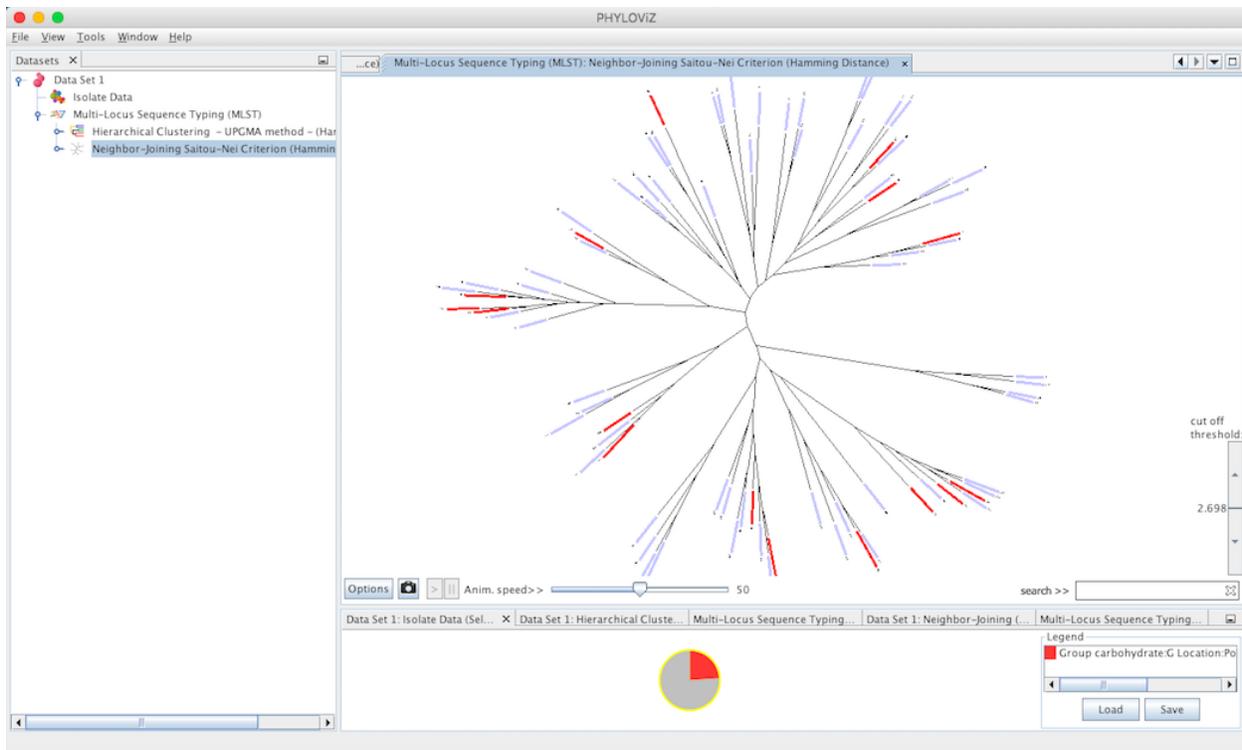
- Queries on the results produced by the *goeBURST* and *goeBURST Full MST algorithm*



- Queries on the results produced by the *Hierarchical Clustering* algorithm



- Queries on the results produced by the *Neighbor Joining* algorithm



5.6 Exporting the results to an image file

To export the resulting graphs to an image file. Click on the *Options* button and choose *Export*. Select the adequate file format for the intended purpose. We recommend the use of png images for presentation quality and eps for publication quality.

The screenshot shows the PHYLOViZ application window. At the top, there is a menu bar (File, Edit, View, Tools, Window, Help) and a tabbed interface with three tabs: 'Start Panel', 'Group CG: goeBURST (Level 1)', and 'Group CG: Isolate Data'. The 'Export view as...' dialog box is open, showing a list of file formats. A red arrow points to the 'Options' button in the dialog. The main window displays a phylogenetic tree with nodes colored in red, blue, green, and grey. Below the tree, there is an 'animation speed' slider set to 50. At the bottom, the 'Output' tab shows a pie chart with a red slice and a grey slice. A legend on the right side of the 'Output' tab shows a red square next to the text 'Group carbohydrate:G Location:Po'. There are 'Load' and 'Save' buttons at the bottom right of the legend.

A large selection of file formats are available. We recommend .png for preview quality and .eps for publication quality. Only the displayed area is exported so make sure you have all the groups you want to be visible in the image file in the Display tab

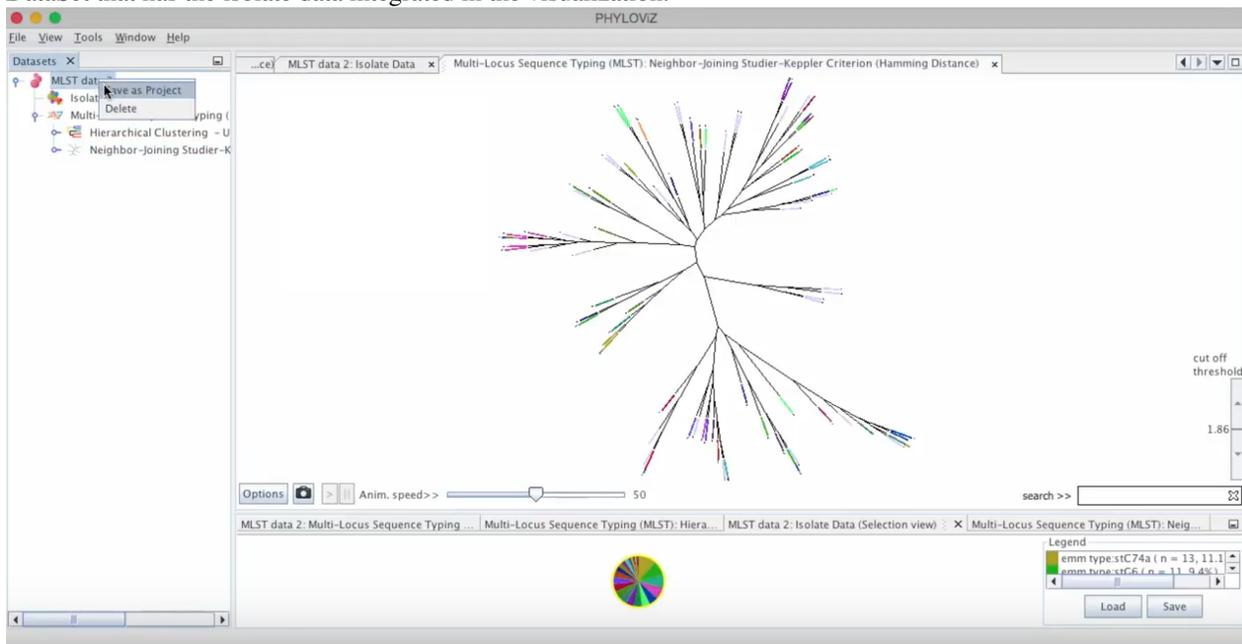
Click on Options and choose Export

Project management

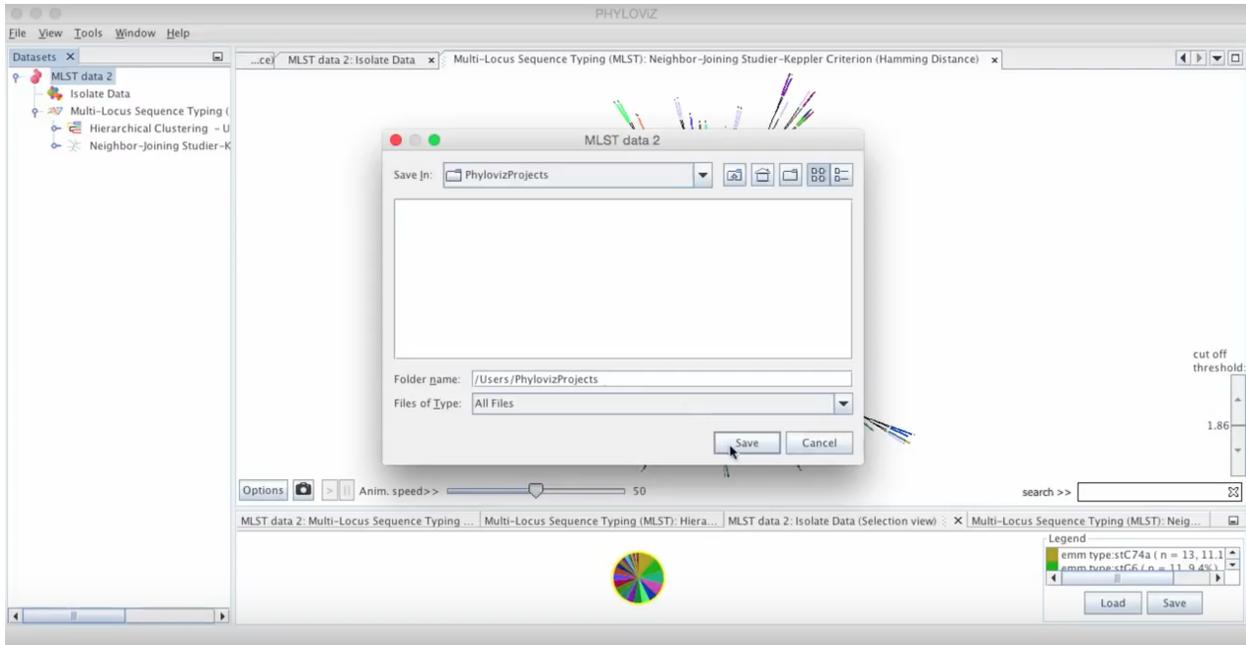
A PHYLOViZ Project allows users to save their ongoing studies and update them as needed. It is a time-saving feature when working with large data sets and essential for efficiently sharing results, since the saved projects can then be shared. Each project includes the data under analysis, results of inference algorithms, visualization serializations and related customizations.

6.1 Saving

Right click on the dataset you would like to save and choose the option *Save as Project*. As you can see we'll save a DataSet that has the isolate data integrated in the visualization.

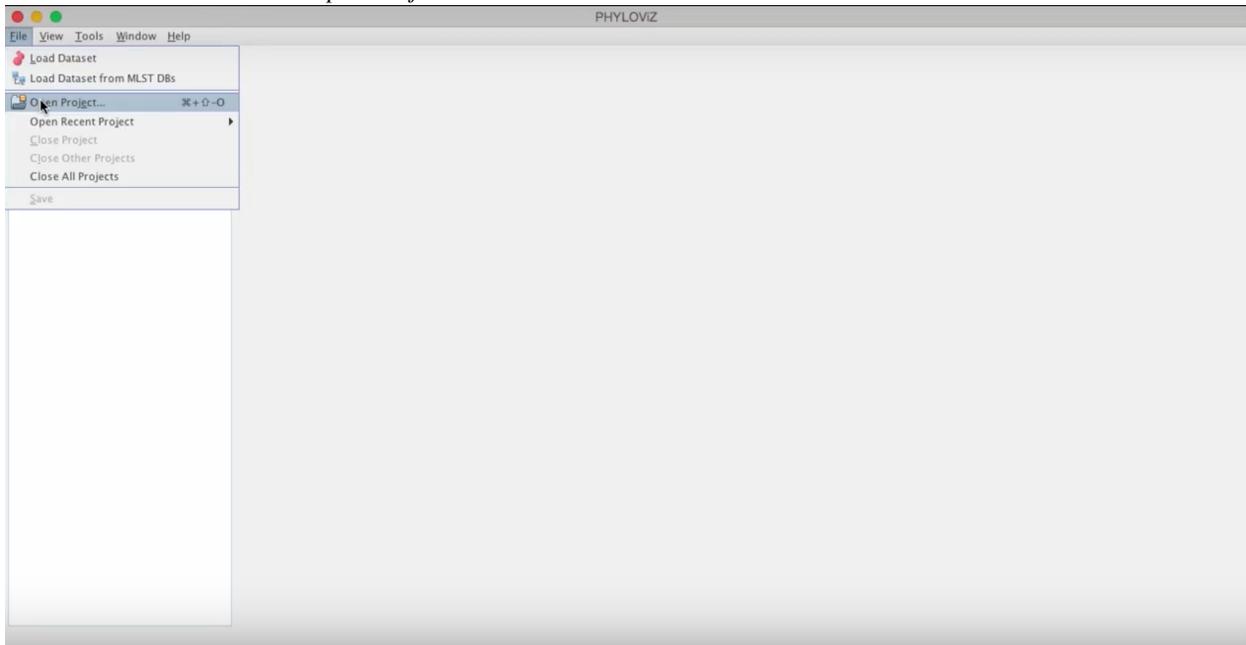


Finally you can choose where to save your project. A dialog appears if you are overriding an existing project or creating a new one with a name that was already taken in the chosen directory.

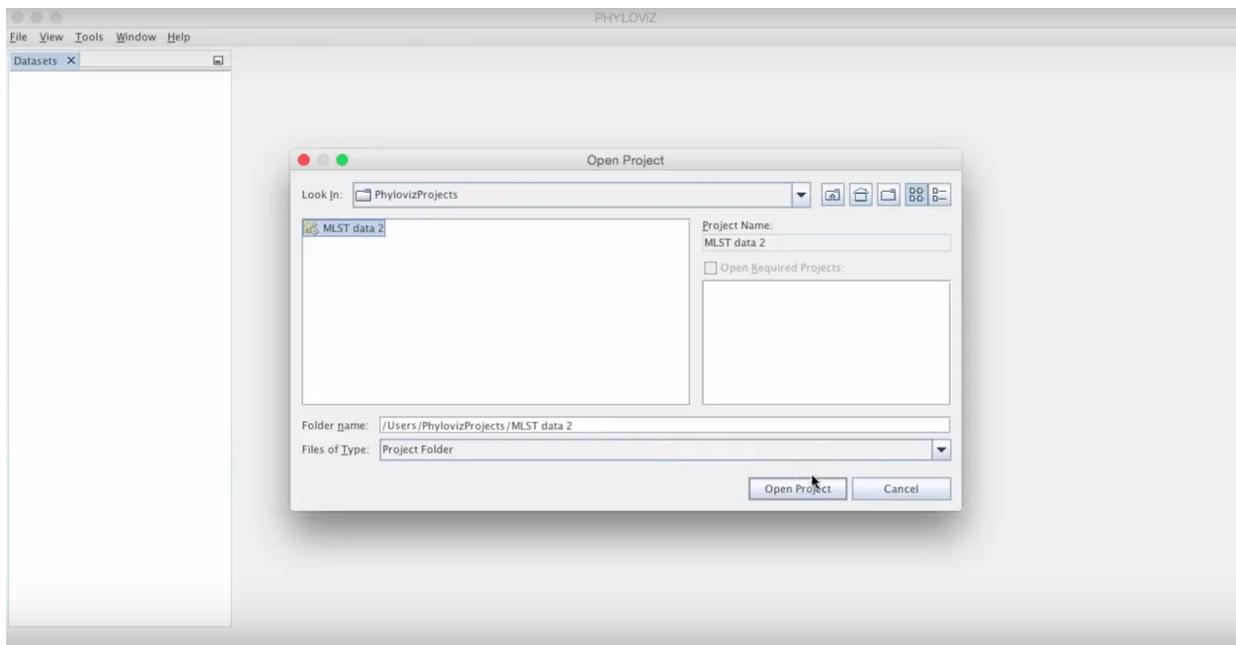


6.2 Loading

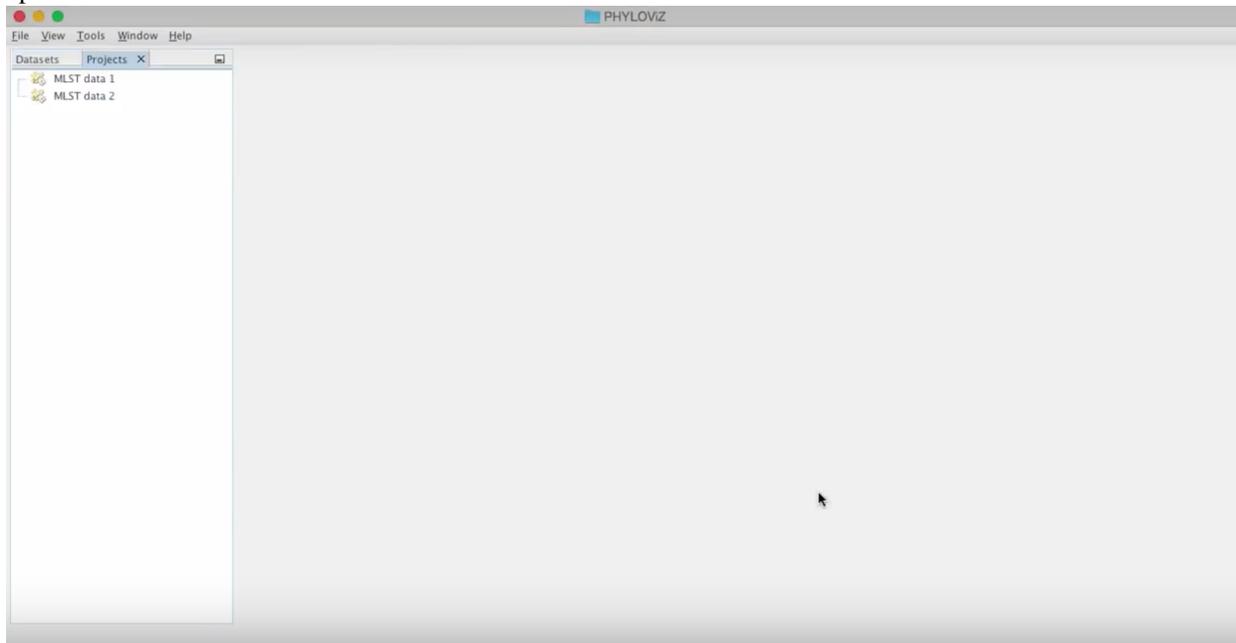
Go to *File* menu and choose *Open Project*.



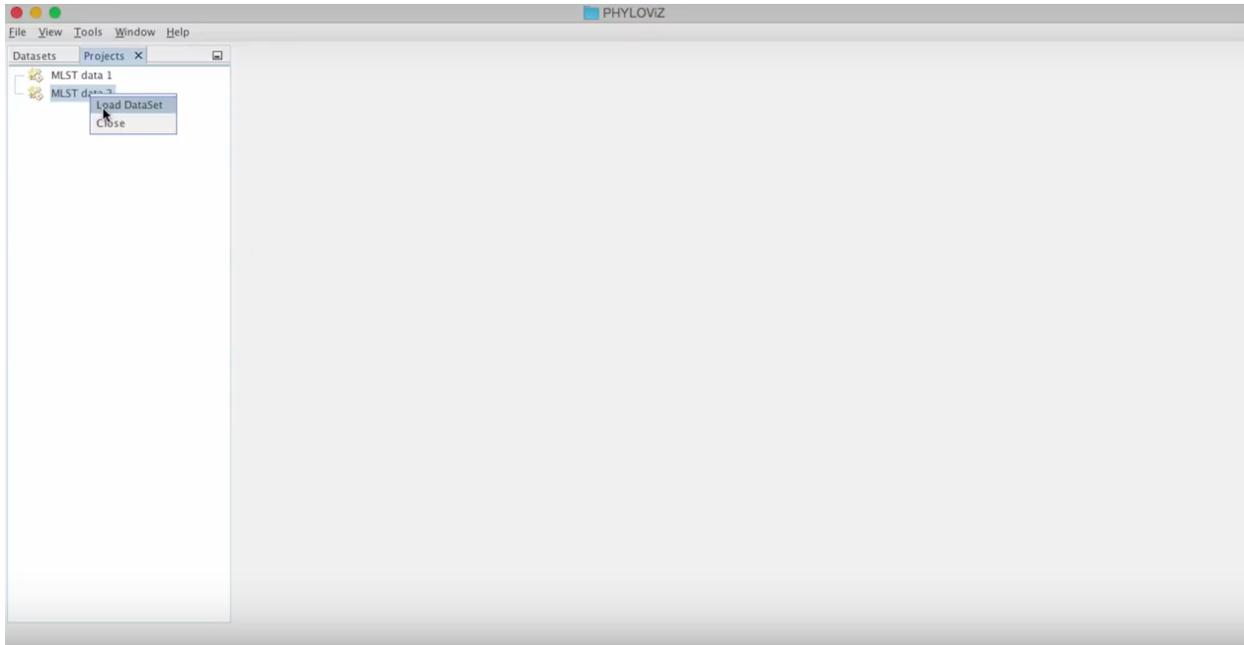
The next step is to find the project that you would like to load. After finding it click on *Open Project*.



This action will open the *Projects Tab* where you can see your project listed and many others that were previously opened.



Now for restoring the study just right click on the project and select *Load DataSet*. This will open the *Dataset's Tab* with your saved study.



The project is now loaded with all the study that was done before as we can see in the following screenshot. You can check that the isolated data integrated on saving was restored completely.

