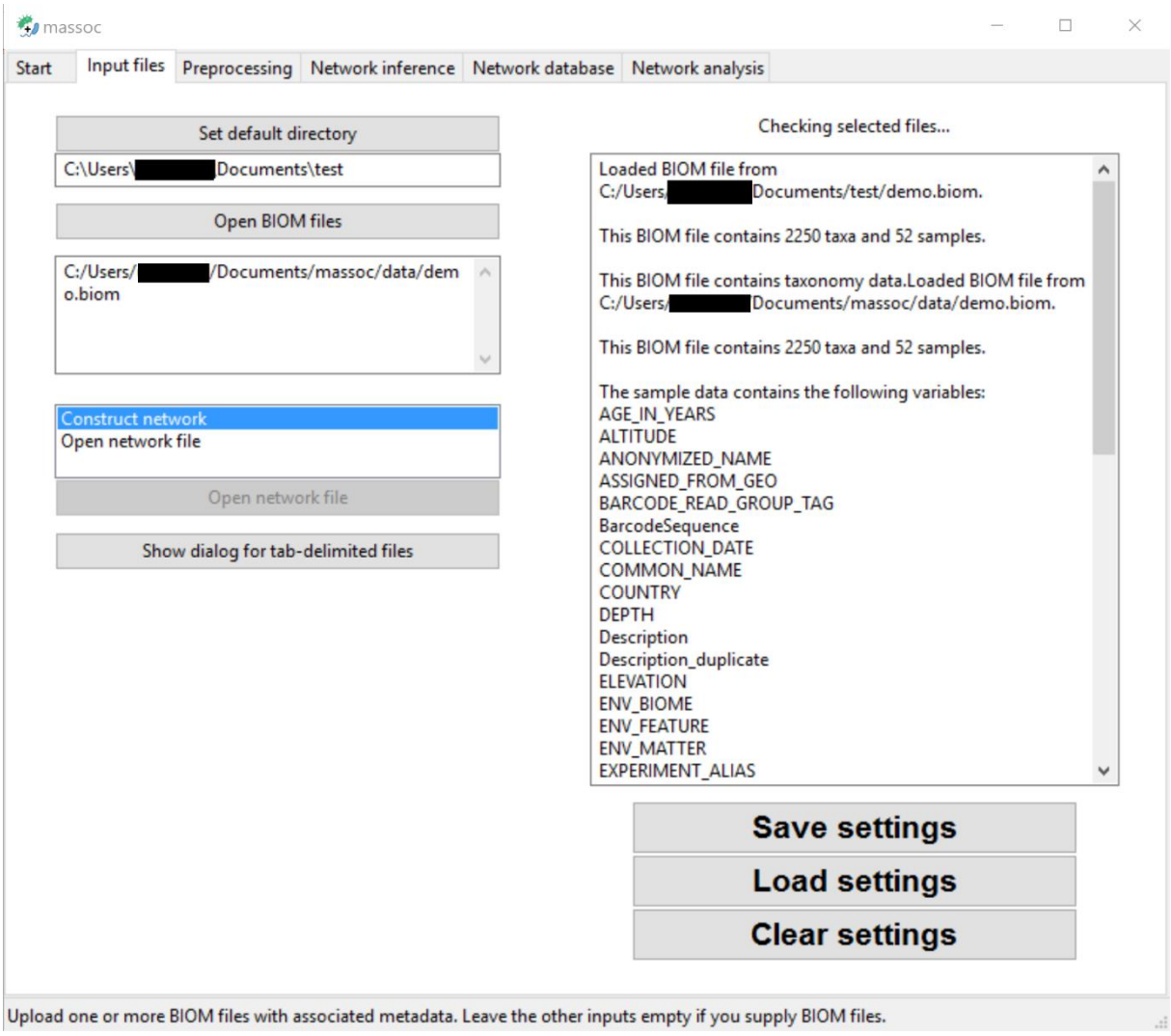
Case study: soil data from the Arctic

In this case study, we are going to perform an analysis of data collected from arctic soils. This data can be found here: https://qiita.ucsd.edu/study/description/104

Chu H, Fierer N, Lauber CL, *et al.*: Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ Microbiol.* 2010; **12**(11): 2998–3006.

In the original BIOM file, the mapping file is not included. The BIOM file in the Github repository has already had the mapping file added:
https://github.com/ramellose/massoc/raw/master/data/demo.biom

First, set the default directory and load the data. The BIOM file will be checked automatically, and if it can be imported, the pane on the right will list properties of the BIOM file.

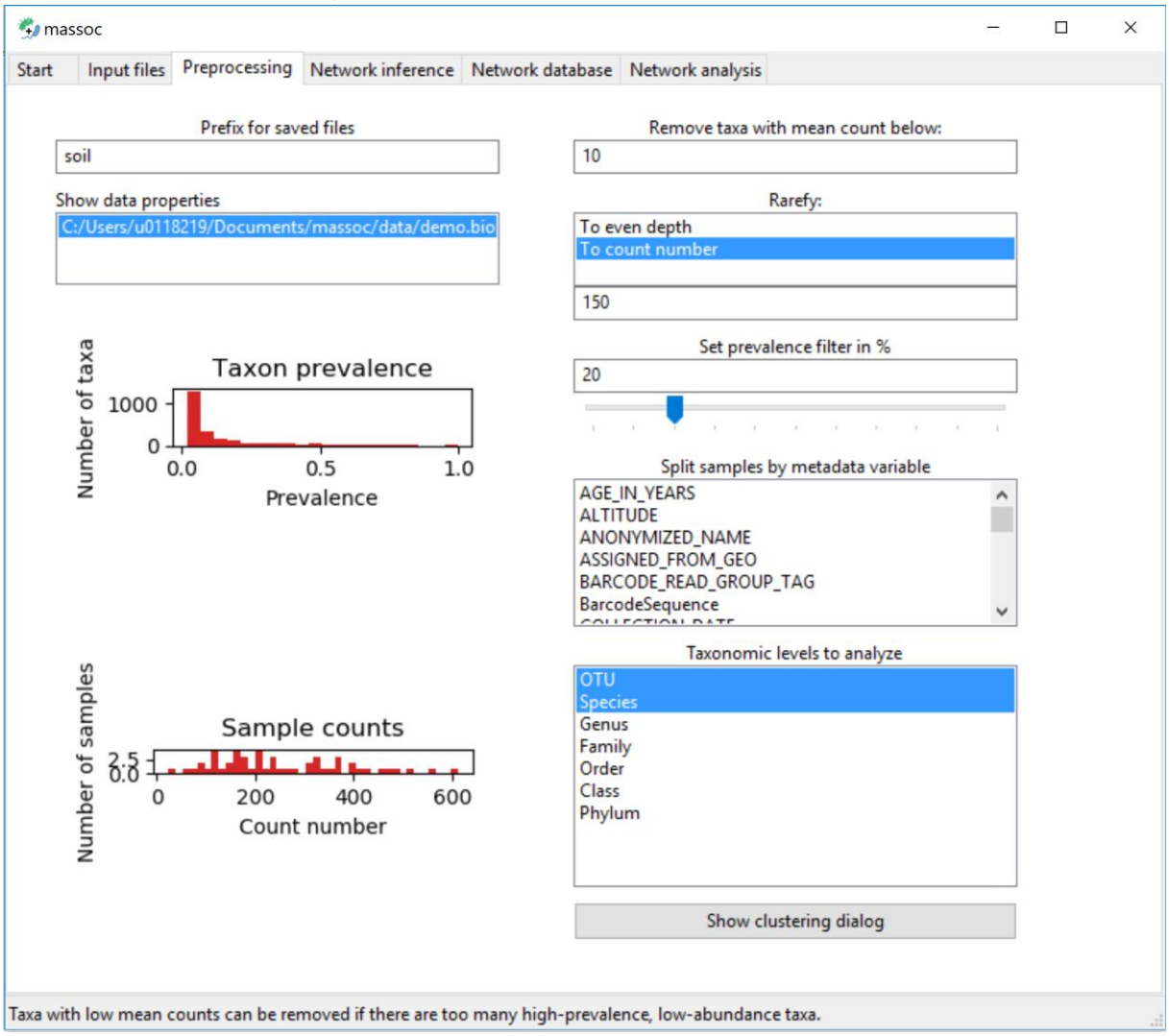


Next, select some preprocessing steps. The prefix is set to 'soil', so any intermediate files generated by *massoc* can be identified with this prefix.

Suitable preprocessing steps can be selected with the prevalence and count plots. In this case, there appear to be many taxa that only occur in a few samples, and many samples have very low counts. To filter taxa, the minimum mean count is set to 10, and a prevalence filter of 20% is applied.

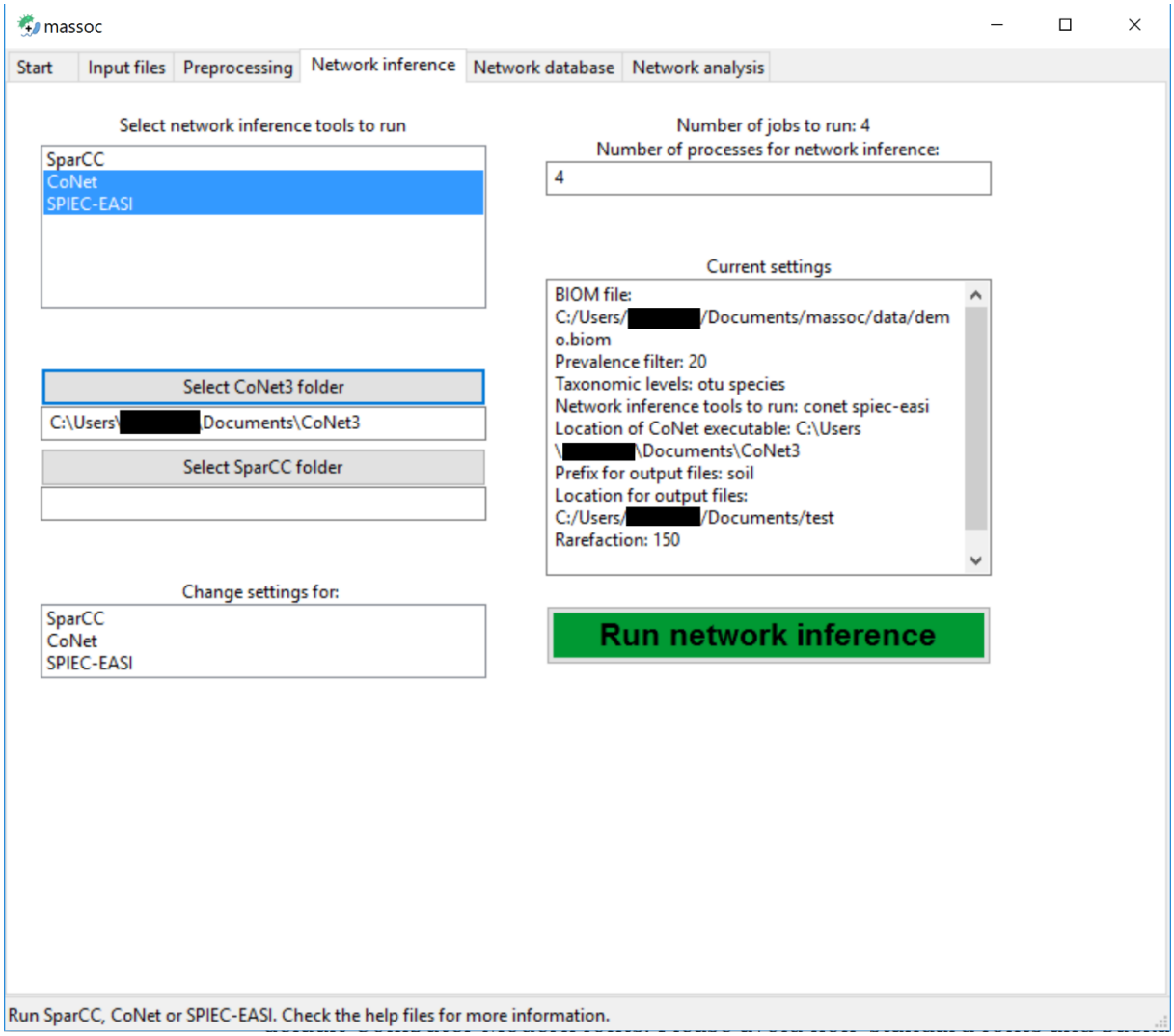
However, there are also some samples with low counts. Set the count number for rarefaction to 150; any samples below this number will not be included, and all remaining samples will be rarefied to 150 counts.

Finally, we can agglomerate taxa before network inference to check associations at different taxonomic levels. In this case, select OTU.



The next step is to infer networks. Select CoNet and SPIEC-EASI in the checkbox and make sure to specify the location of your CoNet3 folder. Here, we are just going to run CoNet with the settings specified in the CoNet.sh script included with *massoc*.

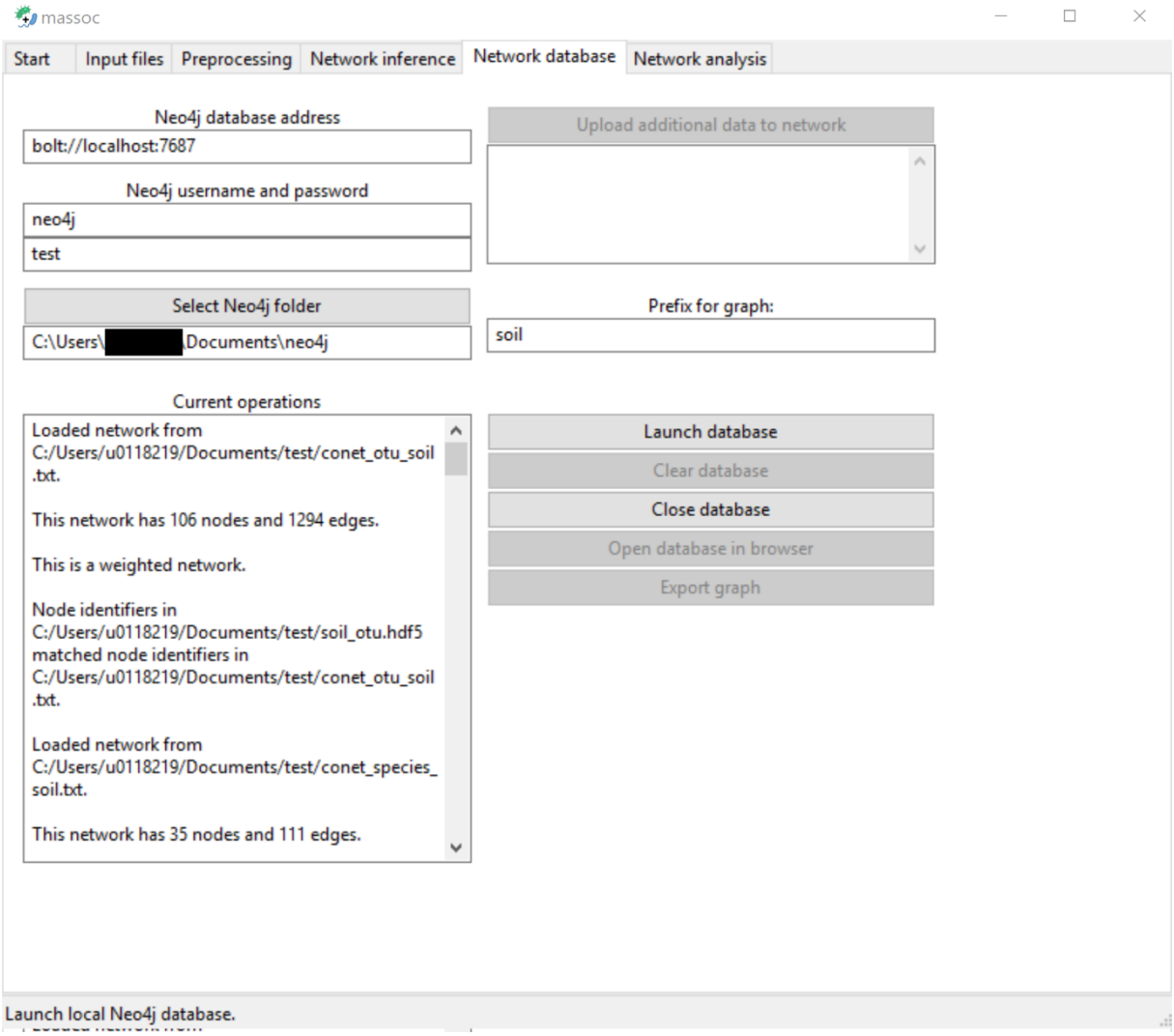
The review pane on the right should show some of the settings that have been selected so far. Run network inference by clicking the green button.



After you complete network inference, a text file of the edges will be written to disk. However, to obtain a rich GraphML file that is compatible with Cytoscape, it is necessary to first connect your network data to the original BIOM file. This demo assumes you already installed Neo4j Server on your machine. If not, please refer to https://neo4j.com/docs/operationsmanual/current/installation/.

After this step, a settings.json file will be written to the specified default directory. If you need to restore settings and intermediate files, load settings from this file and proceed with the next steps.

Specify your address, username and password for the Neo4j Server. Also select the folder that contains your Neo4j Server distribution. Afterwards, click *Launch database*. This will upload the BIOM files and network files to the database.



After the database launches successfully, you can carry out further operations as the buttons will no longer be grayed out. Try looking at the database in your browser by clicking the third button.

Note: it is possible that there is still a lingering Java process running if you previously opened the database, but did not close it safely. This will prevent *massoc* from uploading your data. If this happens, first try using the 'Close database' button. The internal settings may still have the ID of process stored and will therefore shut down the database.

If this does not work, open your task manager and look for a process called "Java™ Platform SE Binary". This process places a lock on the database to make sure it cannot be accidentally overwritten. Killing the process clears the lock so you can start a new database.

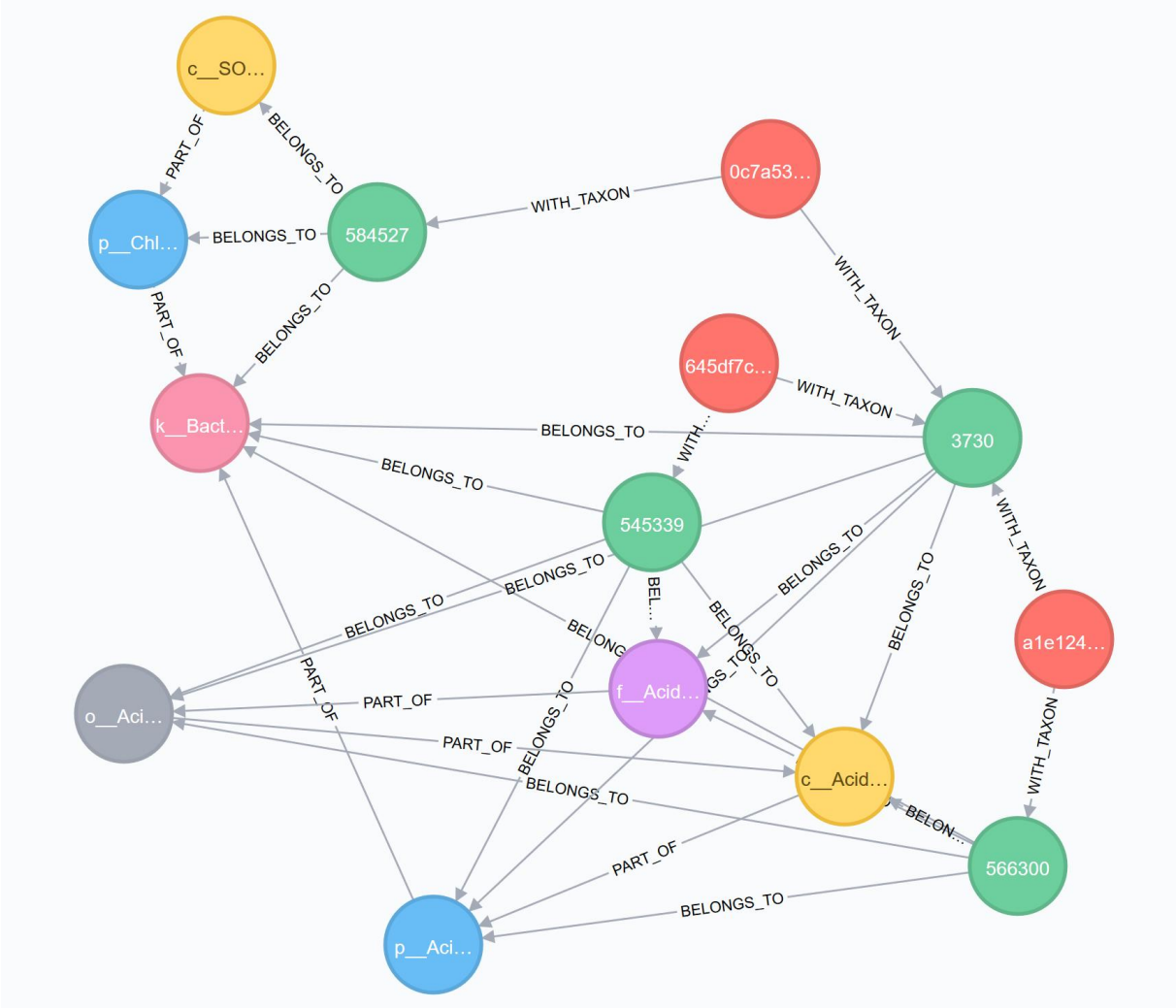
The graph database model used by *massoc* differs from most other microbial network inference tools. If you open the database in your browser, you can explore this database model in more detail.

The Neo4j database can be queried through the Cypher language. If we want to find a taxon and several associated nodes in the databse, we could use the following query in our local environment (by default, available through http://localhost:7474/browser/) :

```
MATCH p=(n)--(:Taxon)--(:Association)--(:Taxon {name: '3730'})--(b)
WHERE NOT n:Sample AND NOT b:Sample AND NOT n:Association AND NOT
b:Association RETURN p LIMIT 60
```

This query will show us other taxa that are connected to taxon 3730. By specifying that the links to those taxa should not be samples or association, we prevent the query from returning all other samples that these taxa were found in. The LIMIT specifies that we get no more than 60 nodes, otherwise the query can clog up the browser rather quickly.

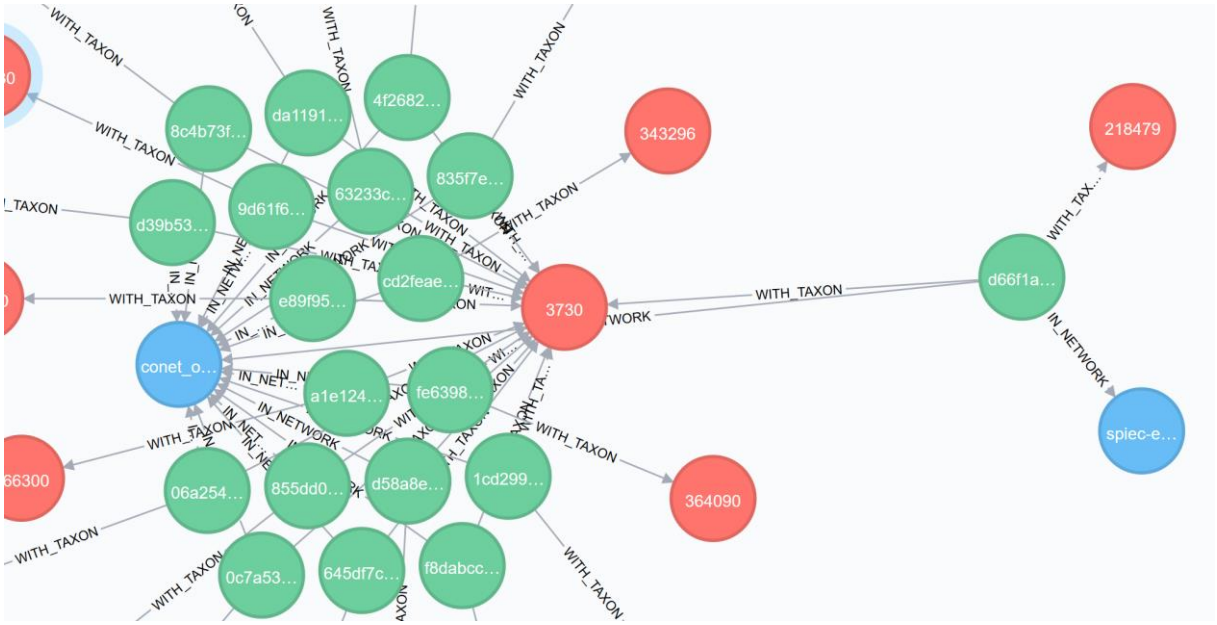
The screenshot below shows how the taxonomy and associations are stored: taxon 3730 has 3 associations (red nodes) with 3 other taxa (green nodes). Two of these taxa are Acidobacteria, while the other taxon belongs to the Chloracidobacteria.



In this case, it is not clear whether the associations were part of the SPIEC-EASI or CoNet networks. We can visualize this in more detail with the following query:

```
MATCH p=()--(:Association)--(:Taxon {name: '3730'}) RETURN p LIMIT 100
```

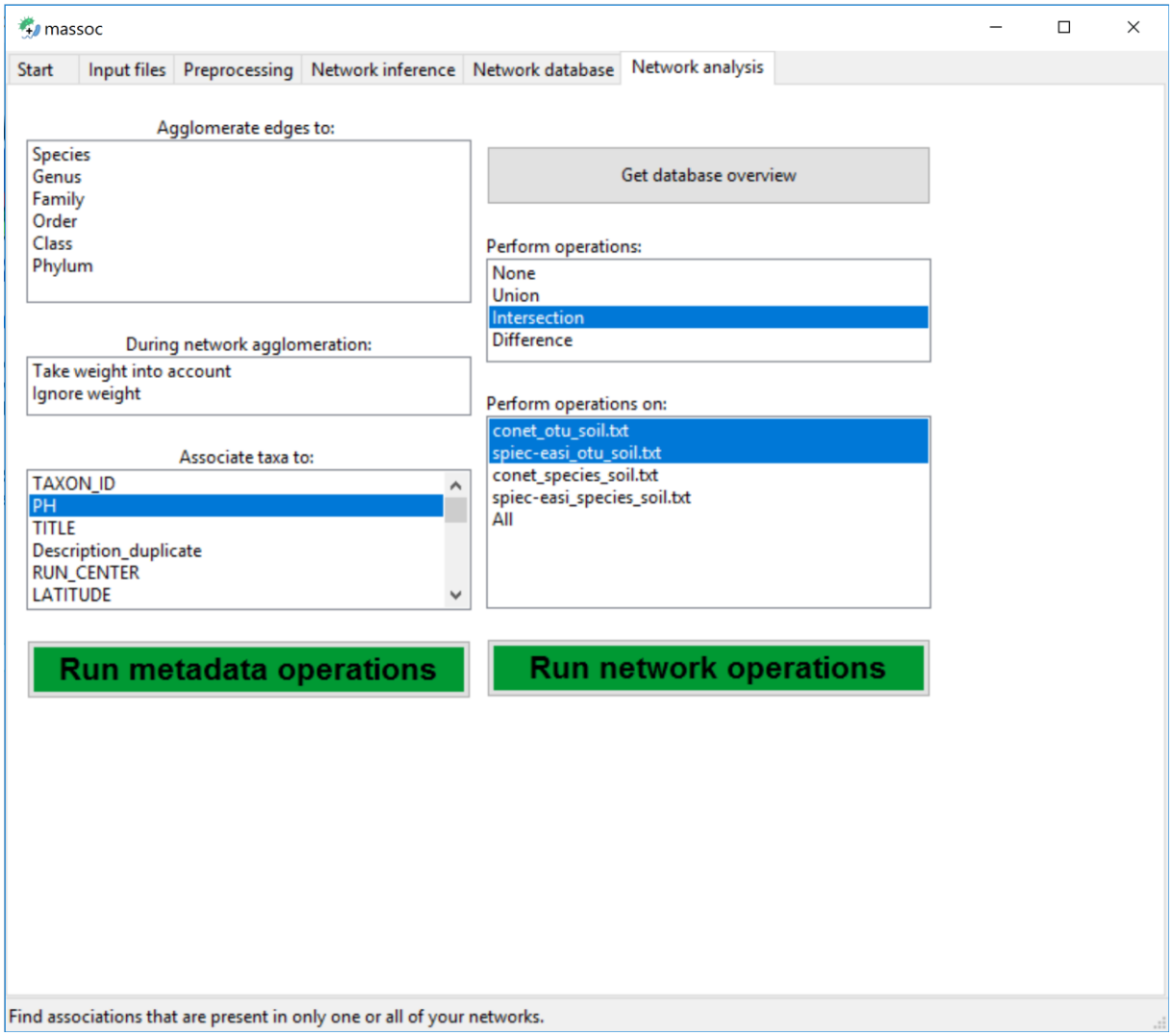
The screenshot shows that only one of the 19 associations was detected with both CoNet and SPIEC-EASI. This is the association with taxon 218479.



Internally, the Cypher language is used to create new nodes, find specific edges, or check whether a particular edges or node already exists. If you want to manipulate the database in more detail, you can do so through the browser or through a Neo4j driver. *massoc* uses the officially supported Neo4j Python driver, but there are drivers available for many languages.

While many of the operations *massoc* carries out in the analysis steps can also be carried out through a Neo4j driver, *massoc* provides a range utility functions that facilitate analysis of microbial networks. In the last step, we are going to extract subnetworks that only contain edges present in at least 2 networks.

First click the 'Get database overview' button. This will extract metadata from the database that you can use to select downstream procedures. Select PH from the metadata list and click on 'Run metadata operations'. Finally, select Intersection and both OTU-level networks. After you click the 'Run network operations' button, a graphml file with the intersection of both networks will be generated. You can open this file in Cytoscape; it should contain the intersection network, with Spearman correlations of pH to taxon abundance as node properties.



In case there is an issue with the program, *massoc* will write logs to a file called 'massoc.txt' one directory above the location of *massoc*. The log file will be refreshed periodically after it reaches a size of 500 KB.