5-May-2017

There are two versions of SourceTracker: SourceTracker and SourceTracker2

https://github.com/danknights/sourcetracker

https://github.com/biota/sourcetracker2

In the long term, I would recommend SourceTracker2, but it does not yet automatically generate nice figures like the first version does. SourceTracker2 will provide all results you need in a table, but the first version then makes pie charts and bar charts from the table for you. I would use SourceTracker2 and just make your own plots.

**General comments that apply to both versions:**

You need a mapping file and an OTU table.

The mapping file assigns each sample to a category and to "source" or "sink". It is tab-delimited. The first line contains the column headers, and the first column header must be "#SampleID".

The mapping file must contain an "Env" column that groups samples by some category. These categories will be used for the SourceTracker analysis.

The mapping file must also contain a "SourceSink" column that labels each sample as "source" or "sink".

OTU table: OTUs as rows, samples as columns. Tab-delimited, first line is a comment starting with "#", second line contains the column headers, first column header must be "#OTU ID".

see examples at <https://github.com/danknights/sourcetracker>

**Mapping File**

I generated the mapping file by copying and pasting all of my sample names into an excel file, manually filling the "Env" and "SourceSink" columns, and then saving as tab-delimited. Change linebreaks to Unix format if necessary.

**OTU Table**

I generated the OTU table from my phyloseq project with the following command:

write.table(t(otu\_table(merged)), file="AMF-otu-table-ST.tsv", quote=FALSE, sep='\t')

\* merged is the name of my phyloseq object

\* the t() wrapped around otu\_table(merged) transposes the rows/columns of the table to match the format required by SourceTracker

\* you still have to manually add the "#" to the first line and then "#OTU ID" to the second line.

**SourceTracker2**

To run SourceTracker2, run the following command (SourceTracker2 is installed globally on the server):

srun sourcetracker2 gibbs -i AMF-otu-table-ST-v2.tsv -m AMF-STmapping.txt -o AMF-st2 &

**Results are in mixing\_proportions.txt. Open in Excel and make plots.**

Manual settings for rarefaction level when running SourceTracker2:

srun sourcetracker2 gibbs -i AMF-otu-table-ST-v2.tsv -m AMF-STmapping.txt -o AMF-st2-2 --sink\_rarefaction\_depth 9900 --source\_rarefaction\_depth 20000 2>> st2\_2.log &

--> standard deviations of sink predictions much lower (better) for manual rarefaction settings, as expected.

No rarefaction at all:

srun sourcetracker2 gibbs -i AMF-otu-table-ST-v2.tsv -m AMF-STmapping.txt -o AMF-st2-3 --sink\_rarefaction\_depth 0 --source\_rarefaction\_depth 0 2>> st2\_3.log &

--> standard deviations even better for no rarefaction

Assignments of OTUs to sources is currently only available with the newest, development version of SourceTracker2. To use the development version, you must explicitly do so by using environmental modules as specified in the Cluster User Guide. In brief, run:

module load programs/SourceTracker-2.0.1-dev

before running `sourcetracker2` to use the development version. Make sure to use:

module switch programs/SourceTracker-2.0.1-dev programs/SourceTracker-2.0.1

to use the stable version again in the same session (if you log out and log in to the cluster, it automatically reverts to the stable version).

It is advisable to run SourceTracker with and without rare species and compare the results because rare species are very likely to be over-represented in the "Unknown" category. This python script for removing rare species may be helpful:

srun filter\_count\_table\_per\_sample.py AMF-otu-table-ST-v2.tsv 20

The '20' at the end of the command indicates that species with fewer than 20 counts in any one sample will be removed from the output table. In other words, only species with 20 or more counts in at least one sample will be included. To be very clear, this means 20 counts in one sample, not a total of 20 counts across all samples, hence the "per sample" in the name of the script.

**Old, not recommended instructions for running SourceTracker 1:**

To run SourceTracker (version 1), I used the following script downloaded from their github (i.e. not installed globally on the server):

srun Rscript sourcetracker\_for\_qiime.r -i AMF-otu-table-ST-v2.tsv -m AMF-STmapping.txt -v TRUE 2>> st.log &

* Requires editing your .bashrc file. For example:

cd ~

nano .bashrc

Add the following text:

export SOURCETRACKER\_PATH=/usr/local/src/SourceTracker-1.0

Ctl-X to close and save

log out of the server, log back in. should be good to go

SourceTracker (version 1) can also be used within an R interactive session. A tutorial for how to use SourceTracker in R can be found [here](https://github.com/danknights/sourcetracker/blob/master/example.r). To use on the cluster, first load the library with:

source('/usr/local/src/SourceTracker-1.0/src/SourceTracker.r')