Making a mapping file

Introduction

QIIME requires that we have a mapping file for several of the steps in the workflow. A mapping file provides a mapping of sample IDs to information about those samples (metadata). This information can be simple demographic information such as the sex and age of a patient, or it can be information about the library preparation such as extraction method, or barcode. Really metadata can be any information about the samples being analyzed but in practice we tend to restrict these data to biologically meaningful information. This metadata is in turn used to contextualize our understanding of the data and guide analysis. For example if we know a patients age we can examine the contribution of age to the structure of the microbiome. Most high-throughput analyses make use of mapping files, so an understanding of what they are is a valuable skill. In this protocol we will construct a QIIME ready protocol.

Objectives

1. Gain experience using popular command line text editors.
2. Construct a QIIME mapping file.

Protocol

1. First we need to copy the template.

```
$ bash
$ cp
/nfs1/Teaching/CGRB/525_f15/data/mapping_file_templat
e.txt ./
# The ./ is a relative path to the directory that we
are currently in.
```

2. Rename the file using the `mv` command. This command will move file but also can change the name of files in place as below.

```
$ mv mapping_file_template.txt
<your_new_file_name.txt>
```

3. Next we should edit this file:

```
$ vim my_mapping_file.txt
```

4. This file should contain a tab delimited header that looks like this:

```
#SampleID BarcodeSequence LinkerPrimerSequence Site
Group Description
```

There are two lines in this file after the header line that you will need to delete. These lines are here to provide an example.

5. Enter edit mode in vim by typing 'a'. Now you can add data to the file such as your library info by typing. Make sure this file is tab separated.

```
#SampleID BarcodeSequence LinkerPrimerSequence Site
Group Description
1.1 CGATTAGAGTAT GGACTACHVGGGTWTCTAAT Buccal 1 1.1
1.2 GGGCCTTATATA GGACTACHVGGGTWTCTAAT Ear 1 1.2
```

Note that the mapping file has some very specific restrictions on what characters can and can't be used. These restrictions are incredibly important to follow as they reflect how the software will parse the data downstream. For example, if the software expects a tab delimited file and we give it a space delimited file the software might improperly split our data causing havoc! So whenever there is information about how to construct mapping files for input into software, you would be wise to throughly read and understand them before creating your mapping file.

The creators of QIIME list these on their website I have included their list here.

1. The first column header must be “#SampleID”, and the data in this column must contain unique (short and meaningful) sample identifiers containing only alphanumeric and period (”.”) characters. Leading and trailing spaces will raise a warning when using validate_mapping_file.py.
2. The second column header must be “BarcodeSequence”, where each value in that column corresponds to the barcode used for each sample. Only IUPAC DNA characters are acceptable. Leading and trailing spaces will raise a warning when using validate_mapping_file.py.
3. The third column header must be “LinkerPrimerSequence”, where each value in that column corresponds to the primer used to amplify that sample. Only IUPAC DNA characters are acceptable. Leading and trailing spaces will raise a warning when using validate_mapping_file.py.
4. All subsequent column headers (except the last one) are metadata headers. For example, a “Smoker” column would include either “Yes” or “No”. Note that the data in each column is assumed to be categorical unless specified otherwise. Categorical data columns must include at least 2 unique values per column. All

metadata must be composed of only alphanumeric, underscore ("_"), period (“.”), minus sign (“-”), plus sign (“+”), percentage (“%”), space (” ”), semicolon (”;”), colon (”:”), comma (”,”), and/or forward slash (“/”) characters. For missing data, write “NA”; do not leave blanks.

5. The last column of the mapping file must be named “Description”. Information in this column includes information that is unique to each sample, such as the medications taken by the patient, or any other descriptive information. The same character restrictions that apply to the metadata columns in guideline four apply to sample descriptions. Sample/Run Description should be kept brief, if possible. Information that applies to all samples in a mapping file should go in the run description section, which is defined as lines starting with a “#” character, immediately following the header line (See example format below.) Information that is specific to a particular sample should go in the “Description” column.
6. There should be no empty lines or comment lines (starting with #) throughout the metadata, with the exception of any additional run description lines that immediately follow the initial header line.
7. Quotes (”) will be stripped from the mapping file (header and data fields) when it is parsed by most scripts in QIIME. For validate_mapping_file.py, these will be flagged with a warning.
8. Stripping of leading and trailing whitespace is only performed on table cells (including sample IDs), not on the column headers. If quote characters (”) are present, these are removed first, followed by whitespace stripping.

6. Once you are done editing the file press the `esc` key to exit edit mode. Just to double check that you have tabs in the file by entering:

```
:set list
```

The output should look something like this:

```
#SampleID^IBarcodeSequence^ILinkerPrimerSequence^ISit
e^IGroup^IDescription$
1.1^ICGATTAGAGTAT^IGGACTACHVGGGTWTCTAAT^IBuccal^I1^I1
.1$
1.2^IGGGCCTTATATA^IGGACTACHVGGGTWTCTAAT^INose^I1^I1.2
$
```

7. To exit vim you have a couple options for both you must first hit the `esc` key. To save changes type `:wq` to disregard and exit the changes type `:q!`.

8. Now we will use a QIIME script to validate the mapping file.

```
#Note the -m option specifies the path to the mapping
file to be validated.
$ validate_mapping_file.py -m my_mapping_file.txt
```

As long as you get no errors you are good to go. If you do get error check to see that these errors were resolved in the corrected version QIIME outputs. You will need to rerun the above step to determine if the corrected file passes QIIME's requirements. If it does then you are good to go, if the revised version still fails then flag down an instrutor or TA.

You are now ready to take start the QIIME workflow.