Metadata Samples procedure for CCFA Microbiome Initiative Bioinformatics Center

Introduction

This document guides researchers through creating a map.txt file. This file contains information regarding the format of the sample data contained in the sample data files uploaded to the researchers project directory. It is used to automatically check the syntax of the sample data contained in the project directory and for notification of potential issues requiring manual intervention (via email to the researcher).

> **Note:** The following example describes sequence data using the Qiime map.txt format. However, any sample specific metadata should be included by listing in a name - value pair format (e.g sample_id → "anything").

map.txt format

The first line of the map.txt format is the header which defines the columns of the data. It must be tab delimited, begin with '#SampleID', and contain any variable which change within the data. The final column in the header line is 'Description'. An example of this map.txt format can be found on the qiime.org website. Below is an example and brief description of the qiime map.txt format:

```
#SampleID BarcodeSequence LinkerPrimerSequence Treatment DOB Description
#Example mapping file for the QIIME analysis package. These 9 samples are from a study of the effects of
#exercise and diet on mouse cardiac physiology (Crawford, et al, PNAS, 2009).
PC.354 AGCACGAGCCTA YATGCTGCCTCCCGTAGGAGT Control 20061218 Control_mouse__I.D._354
PC.355 AACTCGTCGATG YATGCTGCCTCCCGTAGGAGT Control 20061218 Control_mouse__I.D._355
PC.356 ACAGACCACTCA YATGCTGCCTCCCGTAGGAGT Control 20061126 Control_mouse__I.D._356
PC.481 ACCAGCGACTAG YATGCTGCCTCCCGTAGGAGT Control 20070314 Control_mouse__I.D._481
PC.593 AGCAGCACTTGT YATGCTGCCTCCCGTAGGAGT Control 20071210 Control_mouse__I.D._593
PC.607 AACTGTGCGTAC YATGCTGCCTCCCGTAGGAGT Fast 20071112 Fasting_mouse__I.D._607
PC.634 ACAGAGTCGGCT YATGCTGCCTCCCGTAGGAGT Fast 20080116 Fasting_mouse__I.D._634
PC.635 ACCGCAGAGTCA YATGCTGCCTCCCGTAGGAGT Fast 20080116 Fasting_mouse__I.D._635
PC.636 ACGGTGAGTGTC YATGCTGCCTCCCGTAGGAGT Fast 20080116 Fasting_mouse__I.D._636
```

Header and Description lines for the QIIME format

Lines which start with '#' (pound) are either header or description lines. The header line is mandatory, must appear first, and must be tab delimited. It must start with '#SampleID', followed by 'BarcodeSequence', followed by 'LinkerPrimerSequence'. Finally, the column header 'Description' must appear at the end of the line. Again, all header fields must be tab separated. One or more lines

beginning with '#' can be added after the header line which describes the data. These lines can be free form.

## Data lines for the QIIME format

Lines containing data begin with the SampleID field and are alphanumeric and '.' (period). All other data fields allow alphanumeric and the set of characters ._%+- ;:,/ Tabs must be used to separate the fields between the data columns.

This 'map.txt' file ***must*** then be uploaded to the project directory using either the Web UI or Secure FTP as described in the Data Upload.pdf guide.

Map.txt output format

Most researchers will probably use a spreadsheet program to create the map.txt format. Please export the file from the spreadsheet to text before uploading to the server (if possible).

Directory Structure
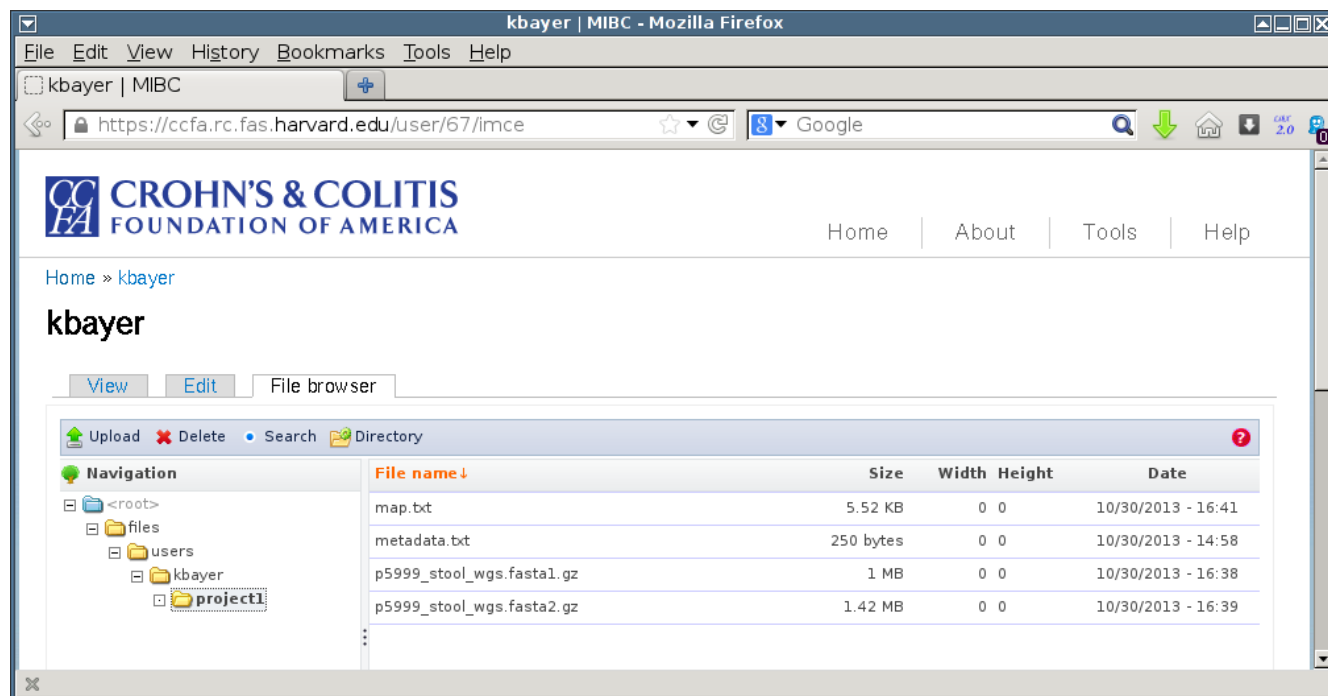
Inside each project folder, the researcher ***must*** upload three types of files:

- Sample data file(s)
- metadata.txt file

- map.txt file

The metadata.txt file type is described in the following document:

1. Metadata: https://ccfa.rc.fas.harvard.edu/docs/Metadata.pdf