Guidelines for implementing iMAP pipeline

## Requirements

The first step is to gather all materials needed for implementing the iMAP pipeline smoothly (Table S1).

Table S1: List of required materials for running iMAP pipeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Required** | **Description** | **Folder** | **Remarks** |
| **iMAP pipeline** | Bundled scripts for comprehensive microbiome analysis | iMAP | [Link](https://github.com/tmbuza/iMAP.git) |
| **Hardware** | Computer with multi-core processor: preferably 64-bit. |  |  |
|  | Remote Accessory Memory (RAM): 8 GB minimum. |  |  |
|  | Storage: Tens of gigabytes for small dataset otherwise a few terabytes |  |  |
| **Raw data** | Demultiplexed reads in FASTQ format with primers and barcodes removed | data/references |  |
| **Sample metadata** | A tab-separated file showing sample identifiers, categorical and numeric variables | data/metadata |  |
| **Mapping file** | A file that links sample IDs (1st column) to the names of forward (2nd column) and reverse (3rd column) data files |  |  |
| **Design files** | Files that assign samples to a specific variables or other categories |  |  |
| **Software** |  |  |  |
| *sekit* | For inspecting rawdata format and simple statistics | code | [Link](https://github.com/shenwei356/seqkit/releases/) |
| *FASTQc* | For creating base call quality score images and statistics | code | [Link](https://www.bioinformatics.babraham.ac.uk/projects/FASTQc/) |
| *bbmap\_bbduk* | For trimming poor quality reads | code | [Link](https://sourceforge.net/projects/bbmap/files/) |
| *multiqc* | For summarizing FASTQc output |  | [Link](https://github.com/ewels/MultiQC.git) |
| *Mothur* | For sequence processing and classifying the sequences and preliminary analysis | code | [Link](https://github.com/Mothur/Mothur/releases/) |
| **Statistical analysis and visualization** |  |  |  |
| *R* | For statistical analysis and visualization |  | [Link](https://cran.r-project.org/) |
| *Rstudio* | An IDE (integrated development environment) for R |  | [Link](https://www.rstudio.com/products/rstudio/download/) |
| *iTOL* | For display, annotation and management of phylogenetic trees |  | [Link](http://itol.embl.de/) |
| **Reference 16S rRNA gene alignments** |  |  |  |
| *SILVA* (nr) | Reference rRNA alignments | data/references | [Link](https://www.Mothur.org/wiki/Silva_reference_files) |
| **Reference 16S rRNA gene classifiers** |  |  |  |
| *SILVA*(no gap) | Degapped using *degap.seqs* function in *Mothur* | data/references | [Link](https://www.Mothur.org/wiki/Silva_reference_files) |
| *RDP* | Mothur-formatted | data/references | [Link](https://www.Mothur.org/wiki/RDP_reference_files) |
| *Greengenes* | Mothur-formatted | data/references | [Link](https://www.Mothur.org/wiki/Greengenes-formatted_databases) |
| *EzBioCloud* | Mothur-formatted | data/references | [Link](https://www.ezbiocloud.net/resources) |
| *Custom classifiesr* | Any manually built classifiers |  |  |
|  |  |  |  |

## Download iMAP repository

git clone https://github.com/tmbuza/iMAP.git  
cd iMAP  
  
# OR  
  
curl -LOk https://github.com/tmbuza/iMAP/archive/master.zip  
unzip master.zip  
mv iMAP-master iMAP  
rm -rf master.zip  
cd iMAP  
  
  
# OR  
  
  
wget --no-check-certificate https://github.com/tmbuza/iMAP/archive/master.zip   
unzip master.zip  
mv iMAP-master iMAP  
rm -rf master.zip  
cd iMAP

## Gather required materials

* Raw data
* Metadata
* Install software
* Download reference databases

# Mac  
bash ./code/requirements/iMAP\_requirements\_mac\_driver.bash  
  
# Linux  
  
bash ./code/requirements/iMAP\_requirements\_linux\_driver.bash

## Verify required folders and files

bash ./code/requirements/iMAP\_checkFiles\_driver.bash

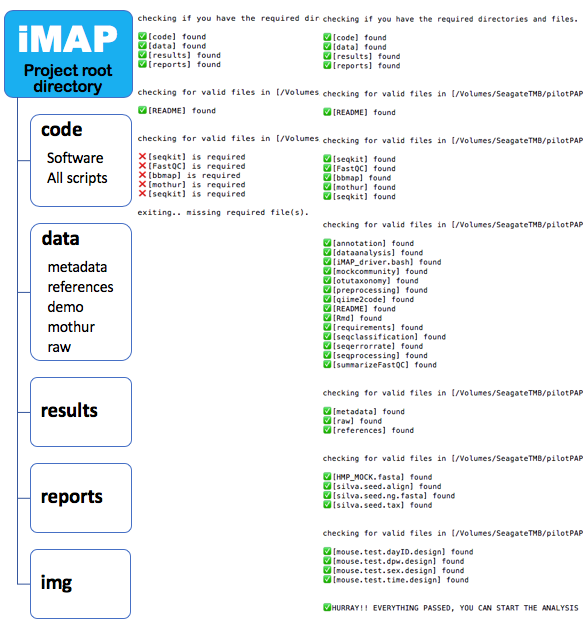


Figure S1: Major folders in the iMAP root directory. Folders and files marked with ✅ exist. Missing file marked ❌ must be found by the above script before proceeding.

## Bioinformatics analysis

### CLI: Command-line-interface

This is basically a method where users sequentially run individual or bundle scripts on CLI (Command -Line\_Interface) one at a time. We have bundled workflow-specific scripts into a driver to make the analysis easily implemented on CLI by just a single click.

bash ./code/requirements/iMAP\_requirement\_driver.bash  
bash ./code/requirements/iMAP\_checking\_driver.bash  
bash ./code/preprocessing/iMAP\_preprocessing\_driver.bash  
bash ./code/preprocessing/07\_multiqc\_fastqc\_summary.bash  
bash ./code/mockcommunity/iMAP\_mockcommunity\_driver.bash  
bash ./code/seqprocessing/iMAP\_seqprocessing\_driver.bash  
bash ./code/seqclassification/iMAP\_seqclassification\_driver.bash  
bash ./code/seqerrorrate/iMAP\_seqerrorrate\_driver.bash  
bash ./code/otutaxonomy/iMAP\_otutaxonomy\_driver.bash

### Running analysis using batch mode on CLI

The *iMAP\_driver.bash* is the master driver for running all analyses on CLI at once.

bash ./code/iMAP\_driver.bash

### Running analysis by submitting a job scheduling through PBS

Users must create a Portable Batch System (PBS) script that describes cluster resources to be used, parameters for the job and the commands to be executed. The following is a PBS script for running executing iMAP pipeline remotely. Note that you must provide the group allocation name (-A) but this may differ from one system to the other. Google for help just in case.

#### Batch mode

#!/bin/bash -f  
  
#PBS iMAPtest  
#PBS -A group allocation name  
#PBS -l nodes=1:ppn=8  
#PBS -l walltime=4000:00:00  
#PBS -l pmem=20gb  
#PBS -j oe  
#PBS -o iMAPtest.log  
#PBS -m abe  
#PBS -M tmb72@psu.edu  
  
cd $PBS\_O\_WORKDIR  
  
bash code/iMAP\_driver.bash

#### Multiple driver scripts

#!/bin/bash -f  
  
#PBS iMAPtest  
#PBS -A group allocation name  
#PBS -l nodes=1:ppn=8  
#PBS -l walltime=4000:00:00  
#PBS -l pmem=20gb  
#PBS -j oe  
#PBS -o iMAPtest.log  
#PBS -m abe  
#PBS -M tmb72@psu.edu  
  
cd $PBS\_O\_WORKDIR  
  
bash ./code/requirements/iMAP\_requirement\_driver.bash  
bash ./code/requirements/iMAP\_checking\_driver.bash  
bash ./code/preprocessing/iMAP\_preprocessing\_driver.bash  
bash ./code/preprocessing/07\_multiqc\_fastqc\_summary.bash  
bash ./code/mockcommunity/iMAP\_mockcommunity\_driver.bash  
bash ./code/seqprocessing/iMAP\_seqprocessing\_driver.bash  
bash ./code/seqclassification/iMAP\_seqclassification\_driver.bash  
bash ./code/seqerrorrate/iMAP\_seqerrorrate\_driver.bash  
bash ./code/otutaxonomy/iMAP\_otutaxonomy\_driver.bash