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
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
FASTQc	For creating base call quality score images and statistics	code	Link
bbmap_bbduk	For trimming poor quality reads	code	Link
multiqc	For summarizing FASTQc output		Link
Mothur	For sequence processing and	code	Link

	classifying the sequences and preliminary analysis		
Statistical analysis and visualization			
R	For statistical analysis and visualization		Link
Rstudio	An IDE (integrated development environment) for R		Link
iTOL	For display, annotation and management of phylogenetic trees		Link
Reference 16S rRNA gene alignments			
SILVA (nr)	Reference rRNA alignments	data/references	Link
Reference 16S rRNA gene classifiers			
SILVA(no gap)	Degapped using *degap.seqs* function in *Mothur*	data/references	Link
RDP	Mothur-formatted	data/references	Link
Greengenes	Mothur-formatted	data/references	Link
EzBioCloud	Mothur-formatted	data/references	Link
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

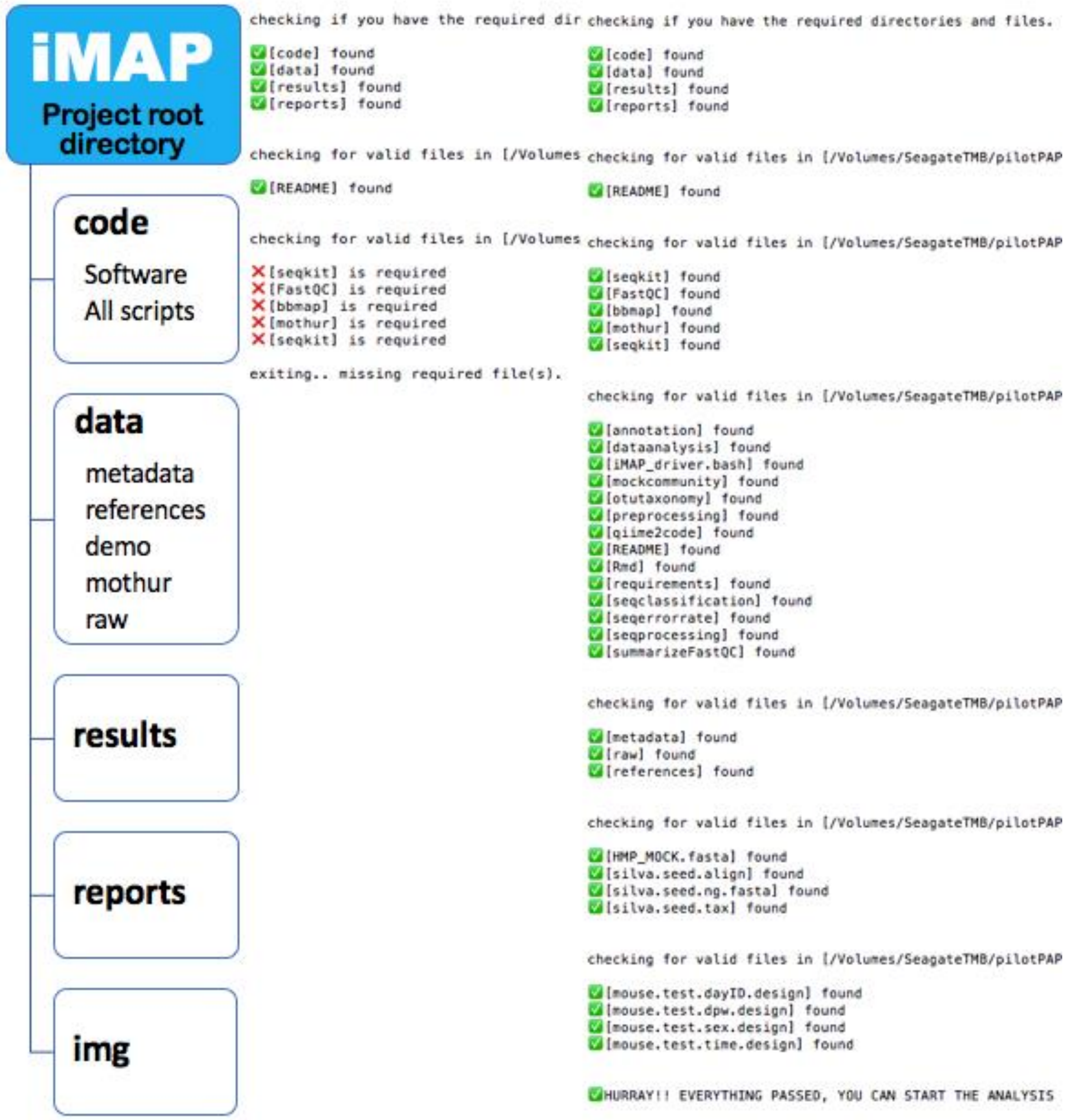
```
# Mac
bash ./code/requirements/iMAP_requirements_mac_driver.bash

# Linux

bash ./code/requirements/iMAP_requirements_linux_driver.bash
```

Verify required folders and files

```
bash ./code/requirements/iMAP_checkFiles_driver.bash
```



```
checking if you have the required dir

✅[code] found
✅[data] found
✅[results] found
✅[reports] found

checking for valid files in [/Volumes

✅[README] found

checking for valid files in [/Volumes

❌[seqkit] is required
❌[FastQC] is required
❌[bbmap] is required
❌[mothur] is required
❌[seqkit] is required

exiting.. missing required file(s).
```

```
checking if you have the required directories and files.

✅[code] found
✅[data] found
✅[results] found
✅[reports] found

checking for valid files in [/Volumes/SeagateTMB/pilotPAP

✅[README] found

checking for valid files in [/Volumes/SeagateTMB/pilotPAP

✅[seqkit] found
✅[FastQC] found
✅[bbmap] found
✅[mothur] found
✅[seqkit] found

checking for valid files in [/Volumes/SeagateTMB/pilotPAP

✅[annotation] found
✅[dataanalysis] found
✅[iMAP_driver.bash] found
✅[mockcommunity] found
✅[otutaxonomy] found
✅[preprocessing] found
✅[qiime2code] found
✅[README] found
✅[Rmd] found
✅[requirements] found
✅[seqclassification] found
✅[seqerrorrate] found
✅[seqprocessing] found
✅[summarizeFastQC] found

checking for valid files in [/Volumes/SeagateTMB/pilotPAP

✅[metadata] found
✅[raw] found
✅[references] found

checking for valid files in [/Volumes/SeagateTMB/pilotPAP

✅[HMP_MOCK.fasta] found
✅[silva.seed.align] found
✅[silva.seed.ng.fasta] found
✅[silva.seed.tax] found

checking for valid files in [/Volumes/SeagateTMB/pilotPAP

✅[mouse.test.dayID.design] found
✅[mouse.test.dpw.design] found
✅[mouse.test.sex.design] found
✅[mouse.test.time.design] found

✅HURRAY!! EVERYTHING PASSED, YOU CAN START THE ANALYSIS
```

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
```