Pipeline for Alcohol consumption analysis

Gray text: Custom name or part that changes with the user

Text on yellow background: Downloaded files, project-specific files or files I gave

General non-black characters Fixed format, should not be modified

Red text : Eye-catching, non-code

>> : at docker container Instructions entered within the order

$> : Instructions entered in the physical environment

PS Connection last introduced: [supercomputers](https://docs.google.com/document/d/1AUo0gO0j7083l-ulGyvCzKbmnkIi2CKtaWJ0DsoKQLI/edit?usp=sharing)

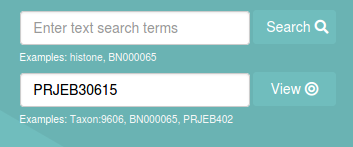
Start Archive:[Weis起始.zip](https://drive.google.com/file/d/1HqkD-o43cu1nX4RW7eyRJjcVEHB-vwml/view?usp=share_link)

**1、Download sequence**

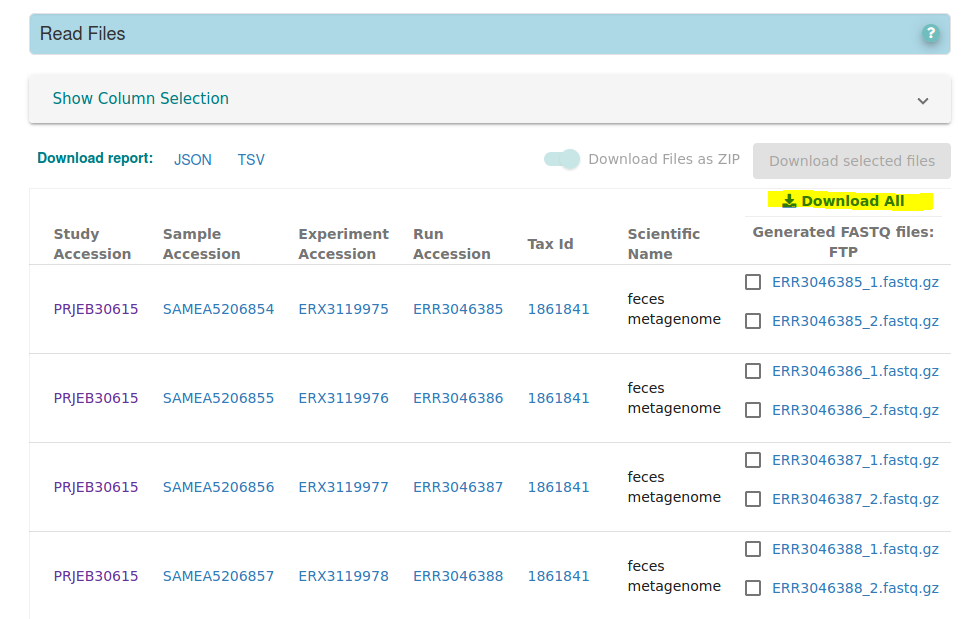
**1. Go to the ENA website**

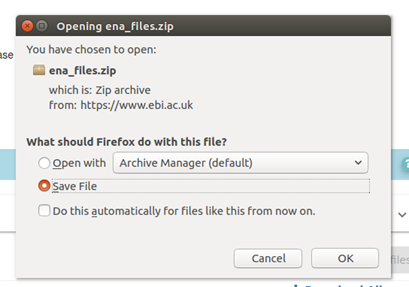
<https://www.ebi.ac.uk/ena/browser/home>

**2. Enter the item number in the upper right corner PRJEB30615**



**3. Download the FASTQ file (default folder /home/shenglab/Downloads)**





(Now sh file)

**4. Create a project folder in /data2/shenglab => weis\_qiime2 (All analysis is performed in data2.)**

$> docker run -it --name shenglab -v /:/tmp ubuntu bash # (The '/' to the left of the colon is the root directory, which means that the entire computer has entered the virtual environment, so you need to be careful with this operation)

>> cd /tmp # (Now the /tmp of the virtual environment corresponds to the root directory of the physical environment, so you should always pay attention to where your current working directory is)

>> mkdir /tmp/data2/shenglab1/hills\_q

**5. Once the data has been downloaded, Move to /data2/shenglab/weis\_qiime2/filtered\_fastq**

$> mkdir ena\_files (Please at Downloads and put sh file inside P.S. Pay attention to the working directory)

$> cd /ena\_files

$> chmod a+x ena.sh

$> sh ena.sh cd ./ena\_files sh

>> mv /tmp/home/shenglab1/Downloads/fastq\_files /tmp/data2/shenglab1/hills\_q/filtered\_fastq

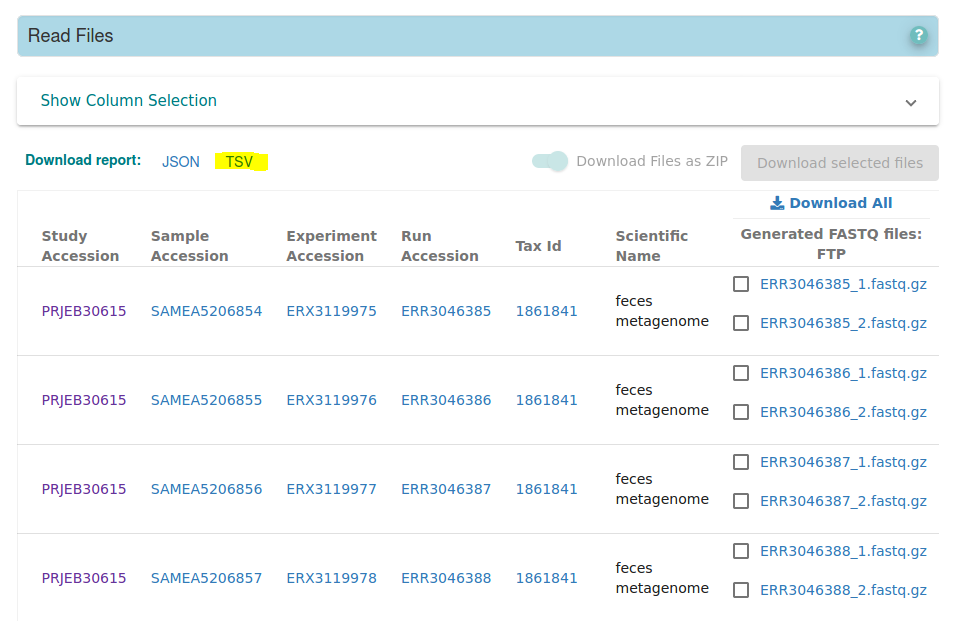
**2. Make metadata (on personal computer).**

**1.**

Create a folder (named: Weis).

**2.**

Go to the ENA website and download the report to Weis (File name: filereport\_read\_run\_PRJEB30615\_tsv.txt)



**3.**

Add the following files to Weis:

(1) metadata\_crowling.R

(2) supplementary data.xlsx

(3) Crawling\_KO\_number\_LEVEL23.R

(4) table1.R

(5) qiime2\_postanalysis.R

(6) Wilcoxon\_and\_volcano\_plot.R

(7) Wilcoxon\_and\_volcano\_plot-2.R

(8) before\_stamp.R

Open Rstudio (you need to install R, Rstudio first) and create the project

File => New Project => Existing Directory => Link to the Weis folder (After that, all from Weis. Rproj file into the project)

**4.**

Import content，取得(1) hillburns\_metadata\_Clean.csv

**5.**

upload(1) hillburns\_metadata\_Clean.csv , to Drive，and download to supercomputer (By default is saved /home/shenglab/Downloads)

**6. Move to /data2/shenglab/weis\_qiime2**

>> mv /tmp/home/shenglab1/Downloads/hillburns\_metadata\_Clean.csv /tmp/data2/shenglab1/hills\_q/hillburns\_metadata\_Clean.csv

>> mv /tmp/home/shenglab1/Downloads/KO\_LEVEL23\_2022\_1214.tsv /tmp/data2/shenglab1/hills\_q/KO\_LEVEL23\_2022\_1214.tsv

>> mv /tmp/home/shenglab1/Downloads/hills1.sh /tmp/data2/shenglab1/hills\_q/hills1.sh

**7. Remove container**

$> docker stop $(docker ps -a -q --filter="name=shenglab")

$> docker rm $(docker ps -a -q --filter="name=shenglab")

**3. Prepare materials**

**1.**

$> docker run -it --name shenglab -v /data2/shenglab1/hills\_q:/tmp ubuntu bash

>> cd /tmp

**2. make manifest (qiime2 has a specified format, Read more** [**Importing data — QIIME 2 2022.2.0 documentation**](https://docs.qiime2.org/2022.2/tutorials/importing/)**)**

#Changed to make it single ended

>> echo -e "sample-id\tabsolute-filepath" > manifest.tsv

>> tail -n +2 hillburns\_metadata\_Clean.csv | while IFS=, read -r ID rest

do

col1=$(echo "${ID}" | tr -d "\r\n" | sed -e 's/^"//' -e 's/"$//')

echo -e "${col1}\t"'$PWD'"/filtered\_fastq/${col1}.fastq" >> manifest.tsv

done

**3. Remove container**

$> docker stop $(docker ps -a -q --filter="name=shenglab")

$> docker rm $(docker ps -a -q --filter="name=shenglab")

**4. QIIME2**

**1. Install qiime2**

>> docker pull quay.io/qiime2/core

**2.**

>> docker run -it --name shenglab -v /data2/shenglab1/hills\_q:/tmp quay.io/qiime2/core /bin/bash

>> cd /tmp

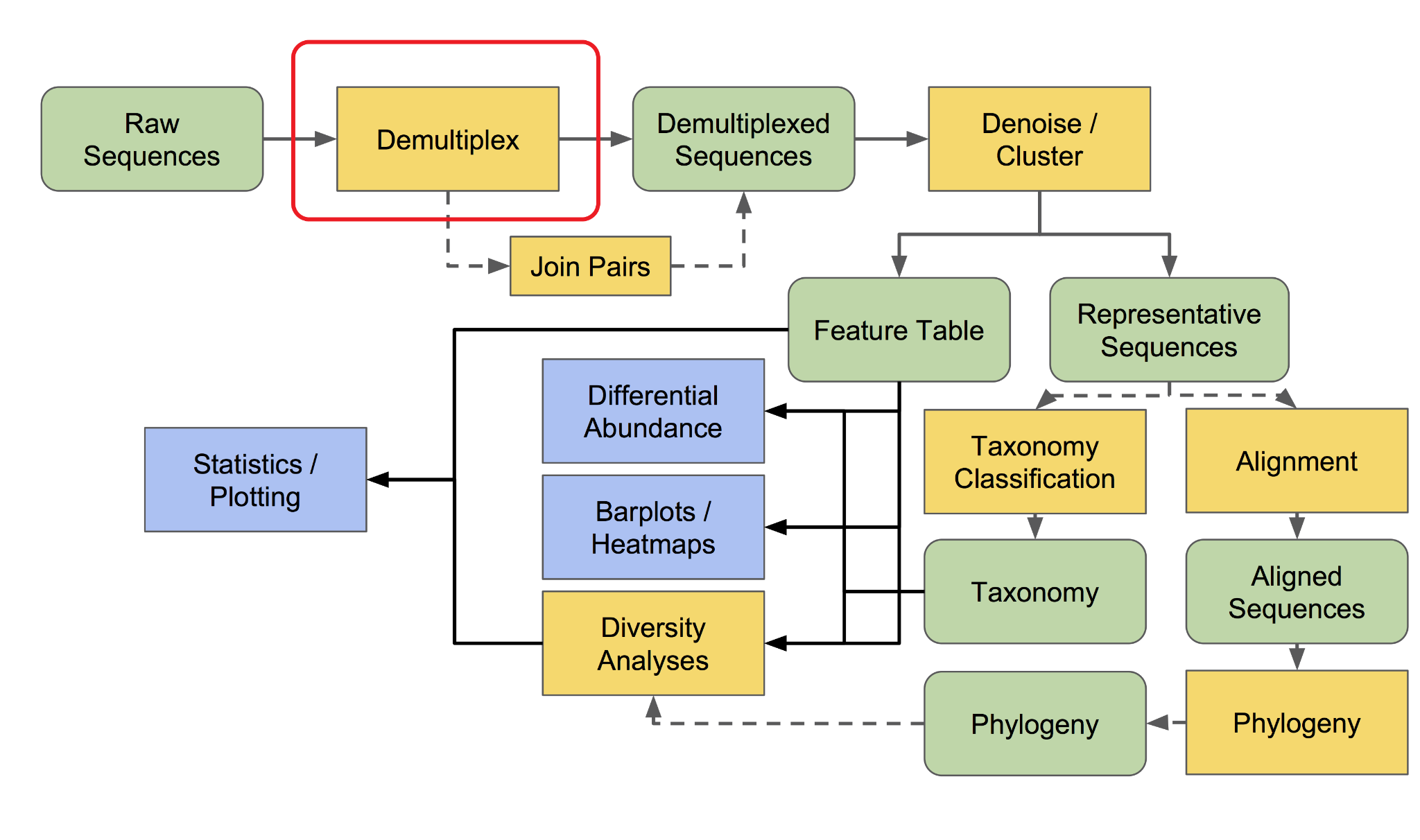
**3. metadata turn csv to tsv**

>> pip install csvkit

>> csvformat -T hillburns\_metadata\_Clean.csv > hillburns\_metadata\_Clean.tsv

**START OF HILLS1.SH**

**4. Import data into the QIIME2 common format .qza**



#For Single ended

>> qiime tools import \

--type 'SampleData[SequencesWithQuality]' \

--input-path manifest.tsv \

--input-format SingleEndFastqManifestPhred33V2 \

--output-path demultiplexed-seqs.qza

**5. QIIME2 Visualization file format .qzv (Place the file in** [**QIIME 2 View**](https://view.qiime2.org/) **to display the visualization)**

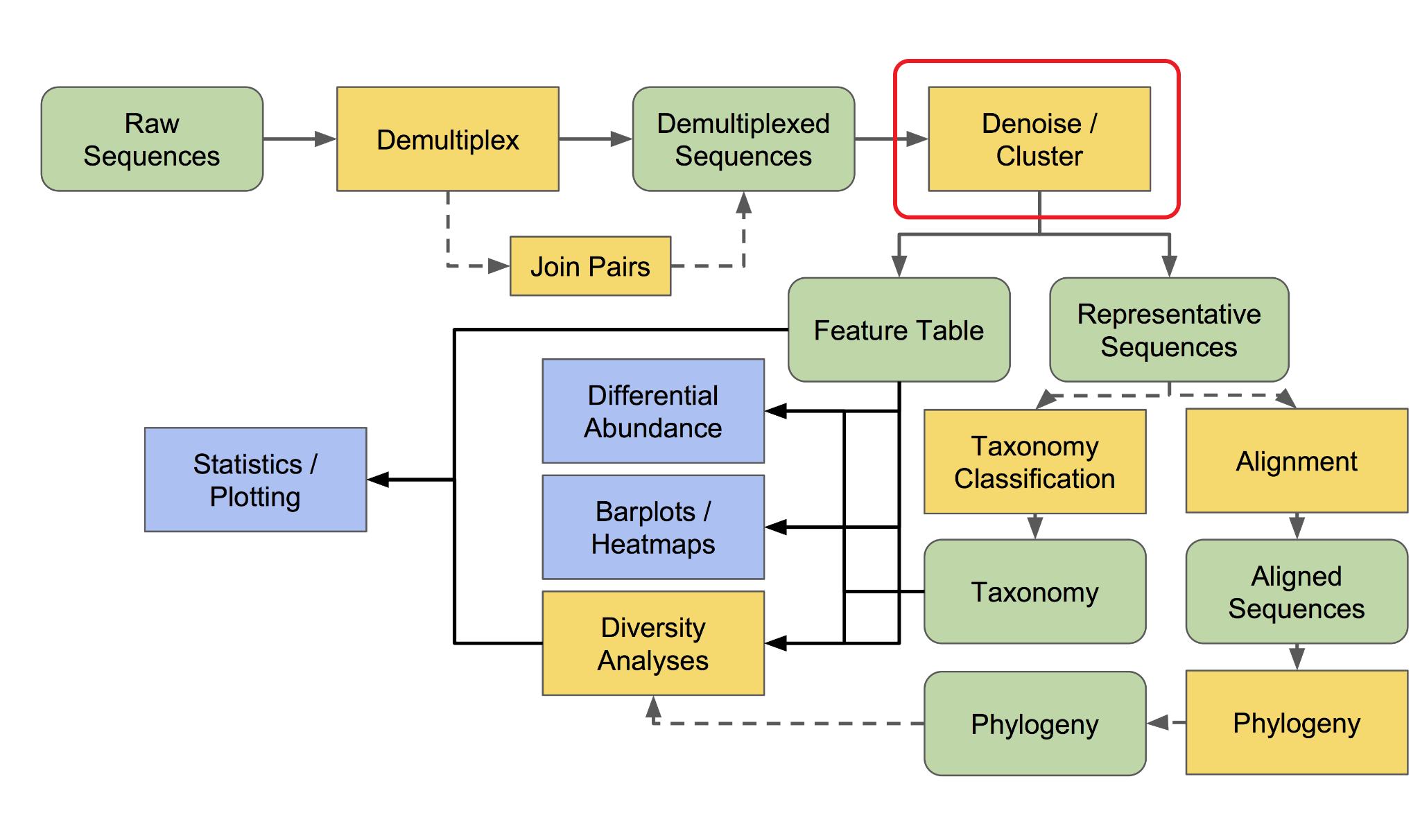
>> qiime demux summarize \

--i-data demultiplexed-seqs.qza \

--o-visualization demultiplexed-seqs.qzv

**6. dada2 (It will be a little long for a few hours; Observe trimmed-seqs.qzv to remove the poor-quality fractions)**

[**https://forum.qiime2.org/t/merging-quality-control-and-overlapping/12618/8**](https://forum.qiime2.org/t/merging-quality-control-and-overlapping/12618/8)



>> qiime dada2 denoise-single \

--i-demultiplexed-seqs demultiplexed-seqs.qza \

--p-trim-left 20 \

--p-trunc-len 150 \

--p-n-threads 0 \

--o-representative-sequences rep-seqs.qza \

--o-table table.qza \

--o-denoising-stats stats-dada2.qza

**7. DADa2 process information visualization**

>> qiime metadata tabulate \

--m-input-file stats-dada2.qza \

--o-visualization stats-dada2.qzv

**8. Visualization of DADA2 results**

>> qiime feature-table summarize \

--i-table table.qza \

--m-sample-metadata-file hillburns\_metadata\_Clean.tsv \

--o-visualization table.qzv

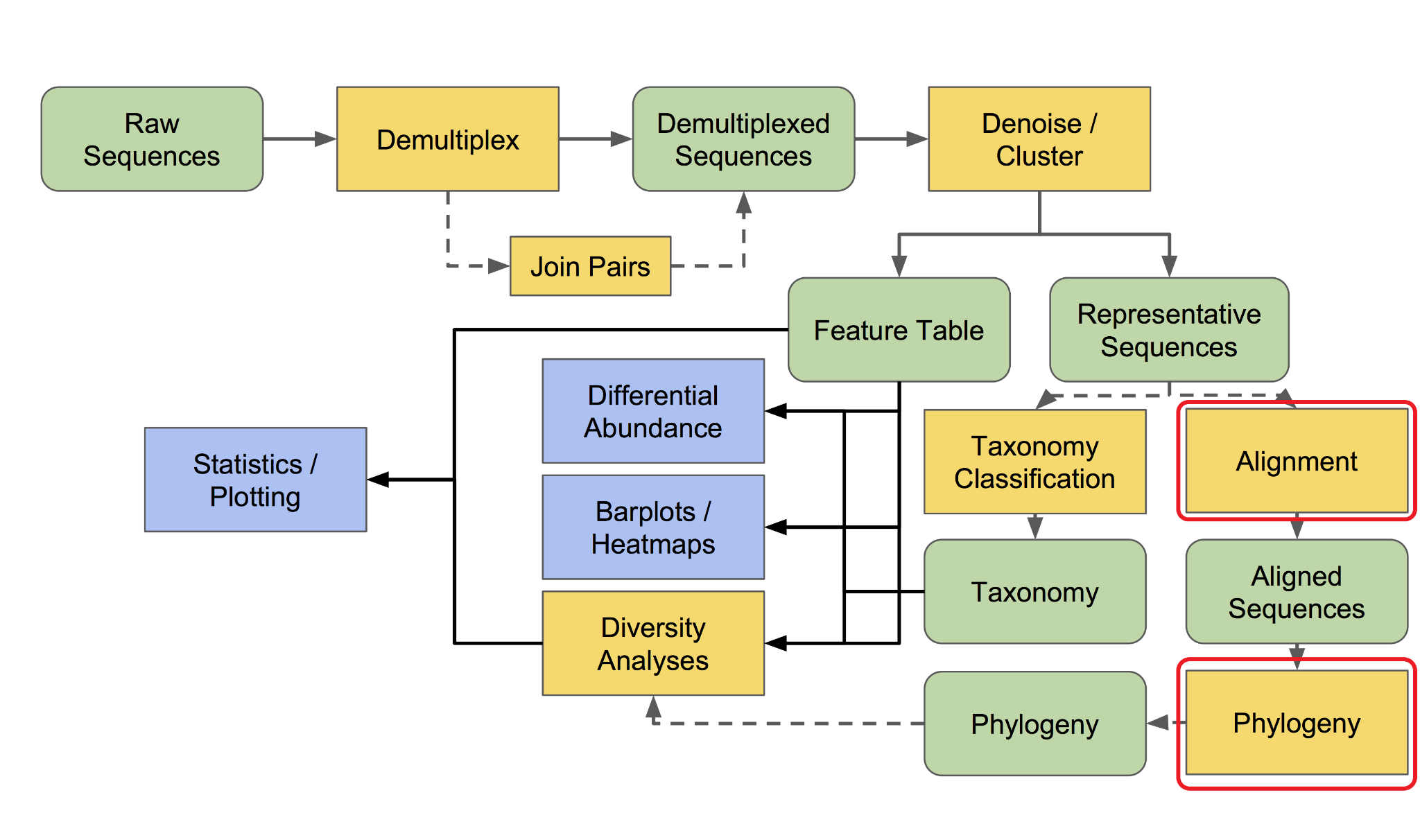
**9. Sequence visualization**

>> qiime feature-table tabulate-seqs \

--i-data rep-seqs.qza \

--o-visualization rep-seqs.qzv

**10. Generating a Phylogenetic Tree (Rooted/Rootless)**



>> qiime phylogeny align-to-tree-mafft-fasttree \

--i-sequences rep-seqs.qza \

--o-alignment aligned-rep-seqs.qza \

--o-masked-alignment masked-aligned-rep-seqs.qza \

--o-tree unrooted-tree.qza \

--o-rooted-tree rooted-tree.qza

**11. normalization**

The following (rarefy / relative-frequency) two methods are selected, due to relative-frequency, some methods are limited, we take rarefy as an example rarefaction To observe table.qzv determines the sampling depth. Observe the minimum sampling frequency

>> qiime feature-table rarefy \

--i-table table.qza \

--p-sampling-depth 1412 \

--p-with-replacement \

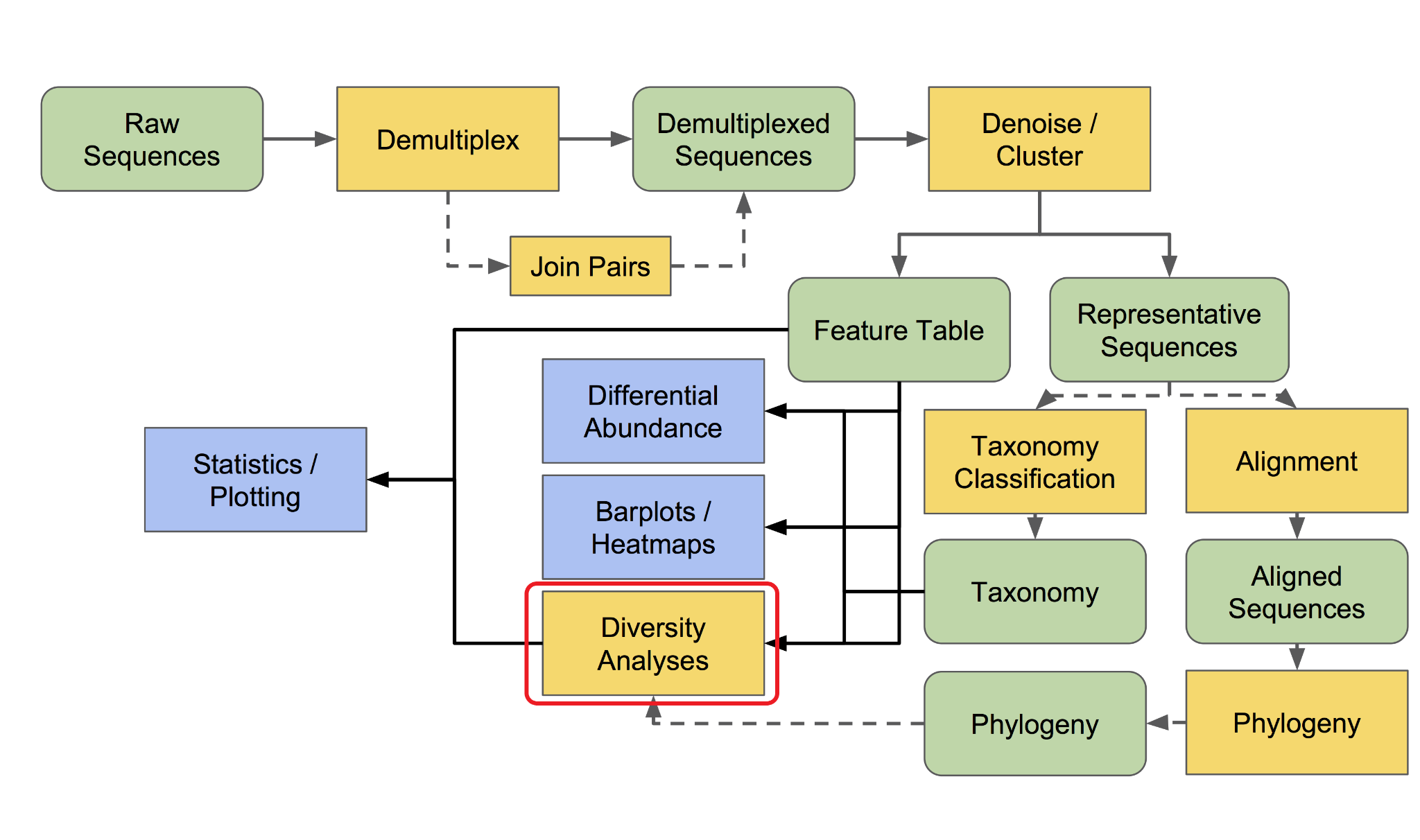
--o-rarefied-table rarefied\_table.qza

>> qiime feature-table relative-frequency \

--i-table table.qza \

--o-relative-frequency-table rel-feature-table.qza

**12. alpha diversity (All available metric references** [**Alpha and Beta Diversity Explanations and Commands**](https://forum.qiime2.org/t/alpha-and-beta-diversity-explanations-and-commands/2282)**)**



>> mkdir metrics-results

>> for METRIC in observed\_features chao1 shannon simpson

do

qiime diversity alpha \

--i-table rarefied\_table.qza \

--p-metric ${METRIC} \

--o-alpha-diversity metrics-results/${METRIC}\_vector.qza

done

**13. alpha diversity visualization**

>> mkdir alpha-diversity-visualization

>> for METRIC in observed\_features chao1 shannon simpson

do

qiime diversity alpha-group-significance \

--i-alpha-diversity metrics-results/${METRIC}\_vector.qza \

--m-metadata-file hillburns\_metadata\_Clean.tsv \

--o-visualization alpha-diversity-visualization/${METRIC}-group-significance.qzv

done

**14. beta diversity distance matrix (phylogenetic)**

>> for METRIC in unweighted\_unifrac weighted\_unifrac

do

qiime diversity beta-phylogenetic \

--i-table rarefied\_table.qza \

--i-phylogeny rooted-tree.qza \

--p-metric ${METRIC} \

--o-distance-matrix metrics-results/${METRIC}\_distance\_matrix.qza

done

**15. beta diversity distance matrix (non-phylogenetic)**

>> for METRIC in canberra

do

qiime diversity beta \

--i-table rarefied\_table.qza \

--p-metric ${METRIC} \

--o-distance-matrix metrics-results/${METRIC}\_distance\_matrix.qza

done

**16. beta diversity statistical verification**

>> mkdir beta-diversity-visualization

>> for METRIC in unweighted\_unifrac weighted\_unifrac canberra

do

qiime diversity beta-group-significance \

--i-distance-matrix metrics-results/${METRIC}\_distance\_matrix.qza \

--m-metadata-file hillburns\_metadata\_Clean.tsv \

--m-metadata-column alcohol\_frequent \

--p-method permanova \

--p-pairwise \

--o-visualization beta-diversity-visualization/${METRIC}-disease-status-significance.qzv

done

**17. beta diversity PCoA matrix**

>> for METRIC in unweighted\_unifrac weighted\_unifrac canberra

do

qiime diversity pcoa \

--i-distance-matrix metrics-results/${METRIC}\_distance\_matrix.qza \

--o-pcoa metrics-results/${METRIC}\_pcoa\_matrix.qza

done

**18. beta diversity PCoA visualization**

>> for METRIC in unweighted\_unifrac weighted\_unifrac canberra

do

qiime emperor plot \

--i-pcoa metrics-results/${METRIC}\_pcoa\_matrix.qza \

--m-metadata-file hillburns\_metadata\_Clean.tsv \

--o-visualization beta-diversity-visualization/${METRIC}\_emperor.qzv

done

**19. Observe table.qzv to determine the maximum depth of the alpha sparse curve (I do it around the median of frequency)**

>> qiime diversity alpha-rarefaction \

--i-table table.qza \

--i-phylogeny rooted-tree.qza \

--p-min-depth 1 \

--p-max-depth 45500 \

--p-steps 50 \

--p-metrics observed\_features chao1 shannon simpson \

--m-metadata-file hillburns\_metadata\_Clean.tsv \

--o-visualization alpha-rarefaction.qzv

**20. Training Species Classifier V4-V5 (520F, 926R Primers) (If there is a ready-made one on the network and the same Primers, you can catch it directly, you don't need to train it yourself, because it takes a long time, and not every computer has that much resources).**

**Related Teaching:**

<https://docs.qiime2.org/2022.2/tutorials/feature-classifier/>

<https://docs.qiime2.org/2022.2/data-resources/>

<https://forum.qiime2.org/t/processing-filtering-and-evaluating-the-silva-database-and-other-reference-sequence-data-with-rescript/15494>

<https://forum.qiime2.org/t/pre-trained-silva-classifier-v3-v4-qiime-2021-4/20470>

>> wget 'https://data.qiime2.org/2022.2/common/silva-138-99-seqs.qza'

>> wget 'https://data.qiime2.org/2022.2/common/silva-138-99-tax.qza'

>> qiime feature-classifier extract-reads \

--i-sequences silva-138-99-seqs.qza \

--p-f-primer GTGCCAGCMGCCGCGGTAA \

--p-r-primer GGACTACHVGGGTWTCTAAT \

--o-reads ref-seqs-520f-926r.qza

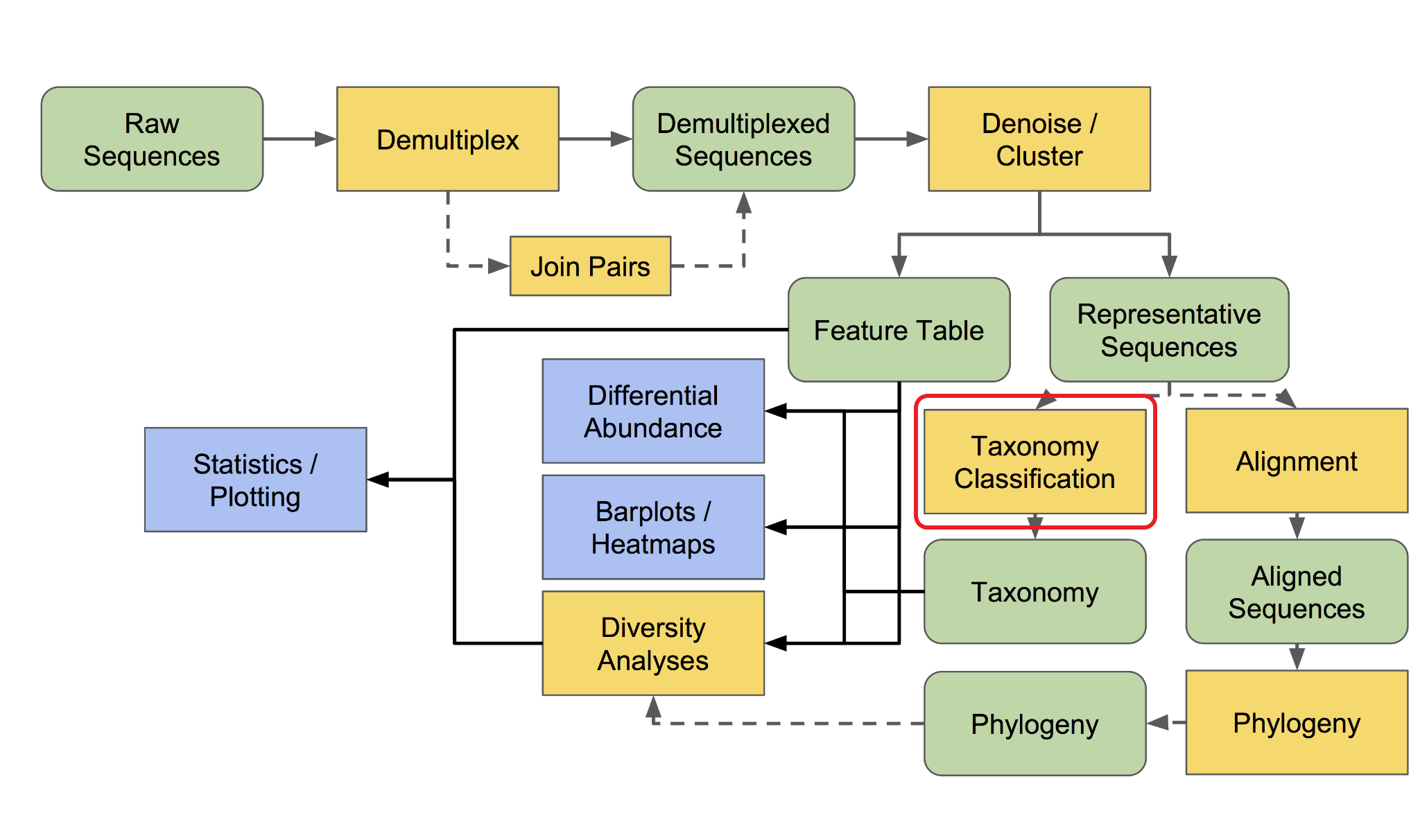
>> qiime feature-classifier fit-classifier-naive-bayes \

--i-reference-reads ref-seqs-520f-926r.qza \

--i-reference-taxonomy silva-138-99-tax.qza \

--o-classifier silva-138-99-520f-926r-classifier.qza

**21. taxonomic analysis**



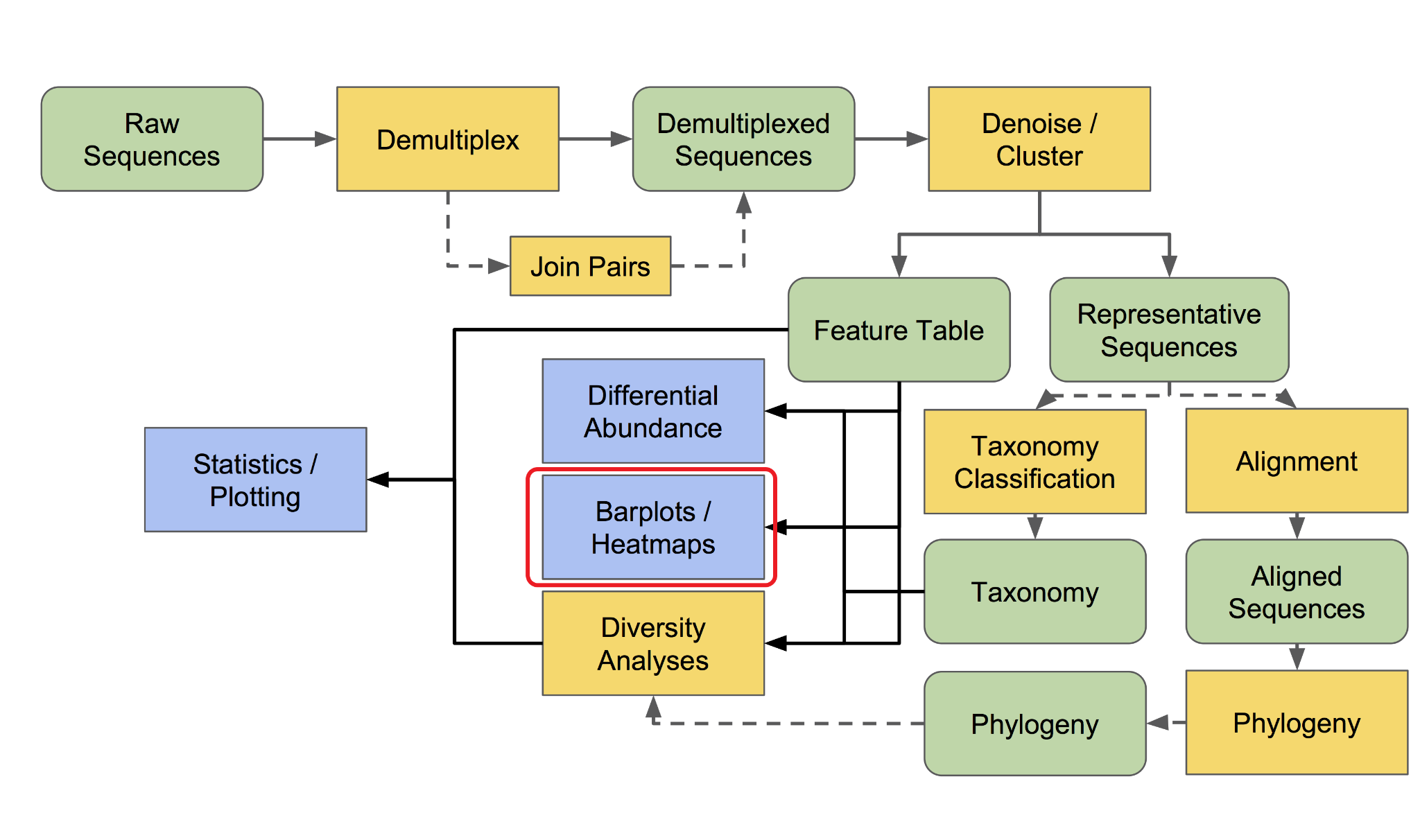
>> qiime feature-classifier classify-sklearn \

--i-classifier silva-138-99-520f-926r-classifier.qza \

--i-reads rep-seqs.qza \

--o-classification taxonomy.qza

**22. taxonomy barplot visualization**



>> qiime taxa barplot \

--i-table table.qza \

--i-taxonomy taxonomy.qza \

--m-metadata-file hillburns\_metadata\_Clean.tsv \

--o-visualization taxa-bar-plots.qzv

**23. make taxonomy frequency form**

>> mkdir phyloseq

>> taxon=( domain-kingdom phylum class order family genus species )

for i in {1..7}

do

qiime taxa collapse \

--i-table table.qza \

--i-taxonomy taxonomy.qza \

--p-level ${i} \

--o-collapsed-table phyloseq/${taxon[${i}-1]}-table.qza

done

**24. make taxonomy relative frequency form**

>> for taxon in domain-kingdom phylum class order family genus species

do

qiime feature-table relative-frequency \

--i-table phyloseq/${taxon}-table.qza \

--o-relative-frequency-table phyloseq/rel-${taxon}-table.qza

done

**25. taxonomy frequency Heatmap visualization**

>> for taxon in order family genus

do

qiime feature-table heatmap \

--i-table phyloseq/${taxon}-table.qza \

--o-visualization heatmap\_${taxon}.qzv \

--p-color-scheme YlOrRd \

--p-cluster both

done

**26. export feature corresponding taxonomy of the table**

>> qiime tools export \

--input-path taxonomy.qza \

--output-path phyloseq/taxonomy

>> mv phyloseq/taxonomy/taxonomy.tsv phyloseq/taxonomy.tsv

>> rm -r phyloseq/taxonomy

**27. export feature frequency form**

>> qiime tools export \

--input-path table.qza \

--output-path phyloseq/feature-table

>> biom convert -i phyloseq/feature-table/feature-table.biom -o phyloseq/feature-table.tsv --to-tsv

>> cp phyloseq/feature-table/feature-table.biom phyloseq/feature-table.biom

>> rm -r phyloseq/feature-table

**28. export feature relative frequency form**

>> qiime tools export \

--input-path rel-feature-table.qza \

--output-path phyloseq/rel-feature-table

>> biom convert -i phyloseq/rel-feature-table/feature-table.biom -o phyloseq/rel-feature-table.tsv --to-tsv

>> rm -r phyloseq/rel-feature-table

**29. export taxonomy frequency form**

>> for taxon in domain-kingdom phylum class order family genus species

do

qiime tools export \

--input-path phyloseq/${taxon}-table.qza \

--output-path phyloseq/${taxon}-table

biom convert -i phyloseq/${taxon}-table/feature-table.biom -o phyloseq/${taxon}-table.tsv --to-tsv

rm -r phyloseq/${taxon}-table

done

**30. export taxonomy relative frequency form**

>> for taxon in domain-kingdom phylum class order family genus species

do

qiime tools export \

--input-path phyloseq/rel-${taxon}-table.qza \

--output-path phyloseq/rel-${taxon}-table

biom convert -i phyloseq/rel-${taxon}-table/feature-table.biom -o phyloseq/rel-${taxon}-table.tsv --to-tsv

rm -r phyloseq/rel-${taxon}-table

done

**31. Export Order**

>> qiime tools export \

--input-path rep-seqs.qza \

--output-path phyloseq/rep-seqs

>> mv phyloseq/rep-seqs/dna-sequences.fasta phyloseq/dna-sequences.fna

>> rm -r phyloseq/rep-seqs

**32. Export rootless trees**

>> qiime tools export \

--input-path unrooted-tree.qza \

--output-path phyloseq

>> mv phyloseq/tree.nwk phyloseq/unrooted\_tree.nwk

**33. Export Rooted Trees**

>> qiime tools export \

--input-path rooted-tree.qza \

--output-path phyloseq

>> mv phyloseq/tree.nwk phyloseq/rooted\_tree.nwk

**34. Prepare the information into the phyloseq folder**

>> cp hillburns\_metadata\_Clean.tsv phyloseq/hillburns\_metadata\_Clean.tsv

>> cp rep-seqs.qza phyloseq/rep-seqs.qza

>> cp table.qza phyloseq/feature-table.qza

>> cp rel-feature-table.qza phyloseq/rel-feature-table.qza

>> cp unrooted-tree.qza phyloseq/unrooted-tree.qza

>> cp rooted-tree.qza phyloseq/rooted-tree.qza

>> cp taxonomy.qza phyloseq/taxonomy.qza

>> cp rarefied\_table.qza phyloseq/rarefied\_table.qza

# Place two folders /data2/shenglab/weis\_qiime2/phyloseq and /data2/shenglab/weis\_qiime2/metrics-results copy to your computer's Weis folder

End of HILLS1.SH

**36. Remove container**

$> docker stop $(docker ps -a -q --filter="name=shenglab")

$> docker rm $(docker ps -a -q --filter="name=shenglab")

**5、PICRUSt2**

**1.**

$> wget https://github.com/picrust/picrust2/archive/v2.5.2.tar.gz (at Downloads)

移至/data2/shenglab1/weis\_ PRJEB30615

$> docker run -it --name shenglab\_ubt -v /:/tmp ubuntu bash

>> cd /tmp

>> mv /tmp/home/shenglab1/Downloads/v2.5.2.tar.gz /tmp/data2/shenglab1/hills\_q/v2.5.2.tar.gz

$> docker stop $(docker ps -a -q --filter="name=shenglab\_ubt")

$> docker rm $(docker ps -a -q --filter="name=shenglab\_ubt")

**2.**

$> docker run -it --name shenglab\_pc2 -v /data2/shenglab1/hills\_q:/tmp continuumio/anaconda3 bash

>>cd /tmp

>> tar xvzf v2.5.2.tar.gz

>> cd picrust2-2.5.2/

>> conda env create -f picrust2-env.yaml

>> conda activate picrust2

>> pip install --editable .

>> cd /tmp

**1. (Do this for each method)**

Go back to your computer and execute Crawling\_KO\_number\_LEVEL23.R contents, acquisition KO\_LEVEL23\_2022\_1214.tsv

Deliver to a supercomputer in the cloud

Move to /data2/shenglab/weis\_qiime2

$> docker run -it --name shenglab\_ubt -v /:/tmp ubuntu bash

>> cd /tmp

>> mv /tmp/home/shenglab1/Downloads/KO\_LEVEL23\_2022\_1214.tsv /tmp/data2/shenglab1/hills\_q/KO\_LEVEL23\_2022\_1214.tsv

$> docker stop $(docker ps -a -q --filter="name=shenglab\_ubt")

$> docker rm $(docker ps -a -q --filter="name=shenglab\_ubt")

**5. Import data**

>> picrust2\_pipeline.py \

-s phyloseq/dna-sequences.fna \

-i phyloseq/feature-table.biom \

-o picrust2\_output\_K

**6.**

>> cd ./picrust2\_output\_K

>> hsp.py \

-i 16S \

-t out.tre -o marker\_predicted\_and\_nsti.tsv.gz \

-p 1 \

-n

**7.**

>> hsp.py \

-i KO -t out.tre \

-o KO\_predicted.tsv.gz \

-p 1

**8.**

>> metagenome\_pipeline.py \

-i ../phyloseq/feature-table.biom \

-m marker\_predicted\_and\_nsti.tsv.gz \

-f KO\_predicted.tsv.gz \

-o KO\_metagenome\_out \

--strat\_out

**9.**

>> pathway\_pipeline.py \

-i KO\_metagenome\_out/pred\_metagenome\_contrib.tsv.gz \

-o KEGG-Pathways \

--no\_regroup \

--map /tmp/picrust2-2.5.2/picrust2/default\_files/pathway\_mapfiles/KEGG\_pathways\_to\_KO.tsv

**10.**

>> add\_descriptions.py \

-i KEGG-Pathways/path\_abun\_unstrat.tsv.gz \

--custom\_map\_table /tmp/picrust2-2.5.2/picrust2/default\_files/description\_mapfiles/KEGG\_pathways\_info.tsv.gz \

-o KEGG-Pathways/path\_abun\_unstrat\_descrip.tsv.gz

**11. Add a description of KEGG's data**

>> add\_descriptions.py \

-i KEGG-Pathways/path\_abun\_unstrat.tsv.gz \

--custom\_map\_table /tmp/KO\_LEVEL23\_2022\_1214.tsv \

-o KEGG-Pathways/L2\_3\_path\_abun\_unstrat\_descrip.tsv.gz

**12. Unzip the final result**

>> gzip KEGG-Pathways/L2\_3\_path\_abun\_unstrat\_descrip.tsv.gz --decompress --stdout > KEGG-Pathways/L2\_3\_path\_abun\_unstrat\_descrip.tsv

# data2/shenglab/weis\_qiime2/picrust2\_output\_K/KEGG-Pathways/L2\_3\_path\_abun\_unstrat\_descrip.tsv Copy to your computer's Weis folder

**13. Remove container (Make sure you don't need it anymore before removing it, otherwise you have to run it again in step 3)**

$> docker stop $(docker ps -a -q --filter="name=shenglab\_pc2")

$> docker rm $(docker ps -a -q --filter="name=shenglab\_pc2")

**6. R Drawing & Extraneous Analysis**

**1. Table 1 Narrative statistics**

execute table1.R

**2. alpha and beta diversity analysis The drawing results are reproduced on R, adding a self-style drawing style**

execute qiime2\_alpha\_plot.R and qiime2\_beta\_plot.R

**3. You can also use the DADA2 file to build a phyloseq format on R to do your own analysis, and here you will demonstrate diversity analysis (not necessary).**

# It should be noted that different software has different methods of normalization when doing diversity analysis, so the results will be somewhat different

execute qiime2\_postanalysis.R

**4. Group verification**

execute Wilcoxon\_and\_volcano\_plot.R

**7、STAMP**

**1. Change the pathway data to a STAMP compliant format**

execute before\_stamp.R acquisition pathway\_stamp.tsv

**2. Install STAMP**

STAMP\_2\_1\_3.exe

**3. STAMPanalysis**

File > load data > pathway\_stamp.tsv and phyloseq/hillburns\_metadata\_Clean.tsv Throw it in

Select Variables in the upper right corner Alcohol\_frequent

Select on the left Two groups

Statistics test: White’s nonparametric t-test

Multiple test correction: Storey FDR

q-value filter (Ticked together will only show significant variables)

Select below Extended error bar