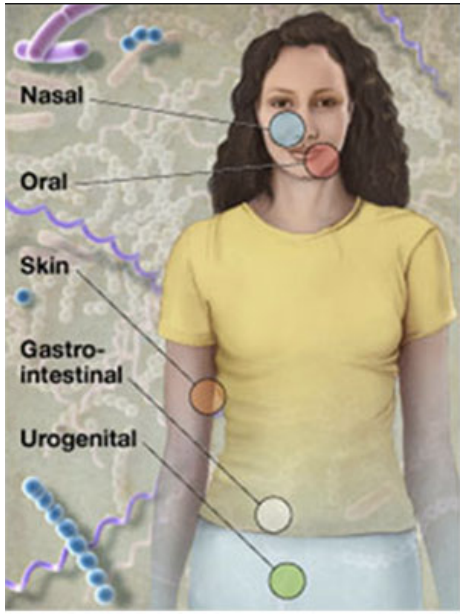
Workshop 11: Metagenomics Analysis

Shi, Baochen
Department of Pharmacology, UCLA

The Microbiome/Microbiota

The Human Microbiome



http://hmpdacc.org/

HMP 2008-2013, funded by NIH

- characterized the microbial communities at several different sites on the human body
- Initial 16S & mWGS metagenomic studies to generate an estimate of the complexity of the microbial community at each body site, providing initial answers to the questions of whether there is a "core" microbiome at each site
- Demonstration projects to determine the relationship between disease and changes in the human microbiome

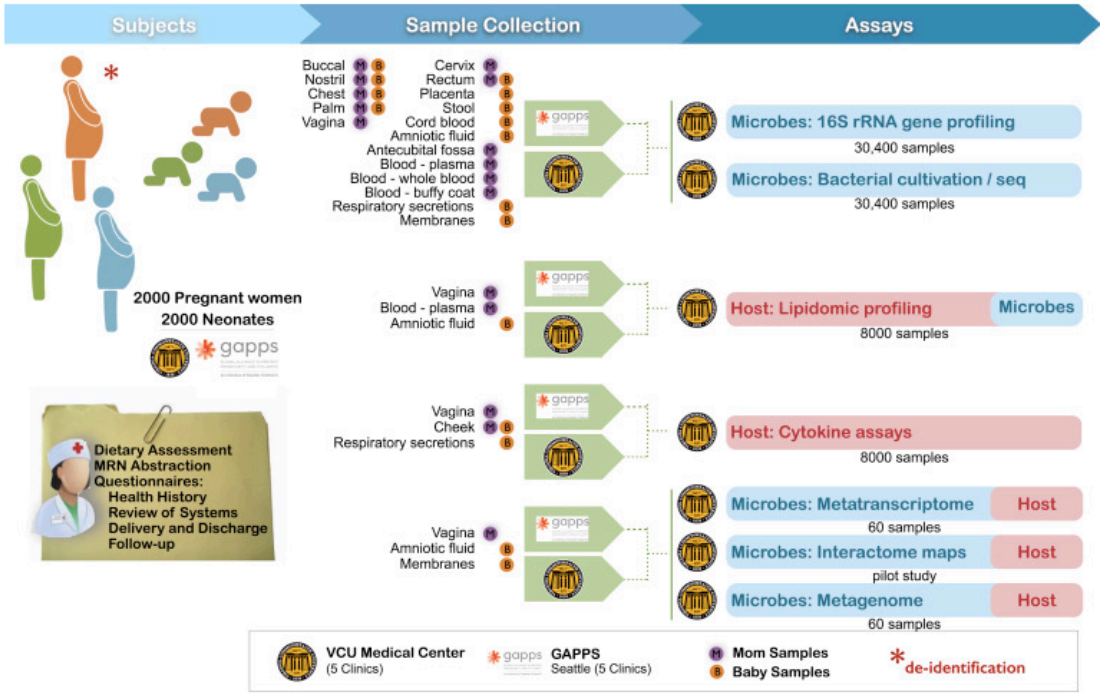
- Development of a reference set of 3,000 isolate microbial genome sequences
- Development of new tools and technologies for computational analysis
- establishment of a data analysis and coordinating center (DACC), and resource repositories

The Human Microbiome

Cell Host Microbe. 2014 Sep 10;16(3):276-89

iHMP creates integrated longitudinal datasets of biological properties from 3 different cohort studies of microbiome-associated conditions using multiple "omics" technologies.

Project 1: Pregnancy & Preterm Birth

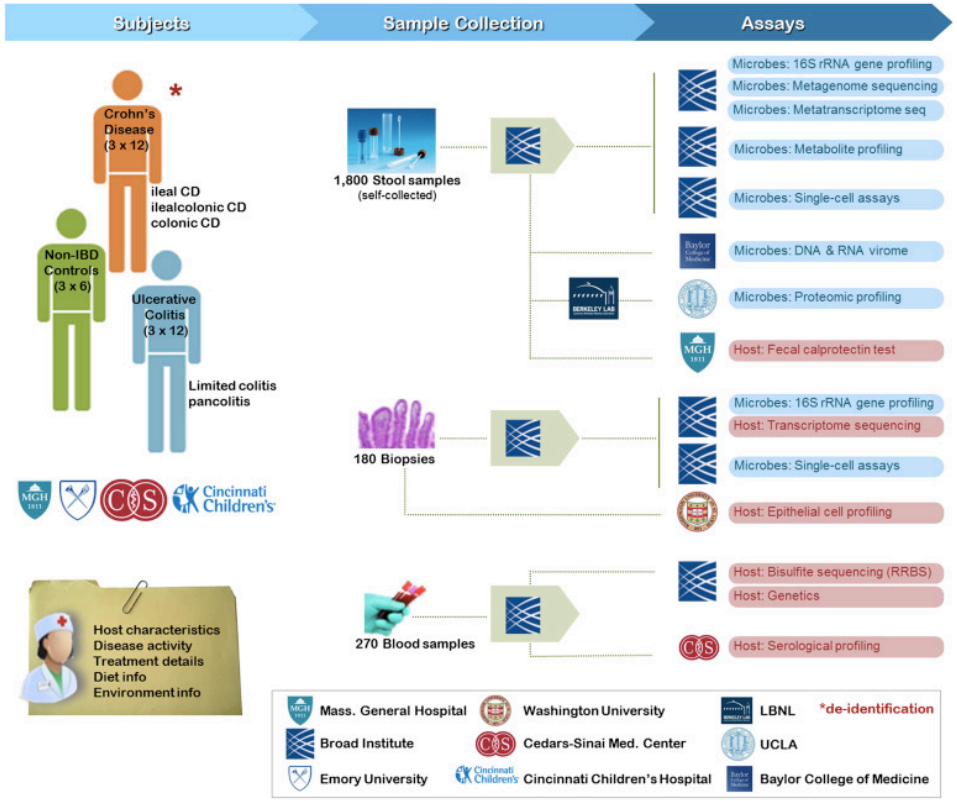


The Human Microbiome

Cell Host Microbe. 2014 Sep 10;16(3):276-89

iHMP creates integrated longitudinal datasets of biological properties from 3 different cohort studies of microbiome-associated conditions using multiple "omics" technologies.

Project 2: Onset of Inflammatory Bowel Disease (IBD)

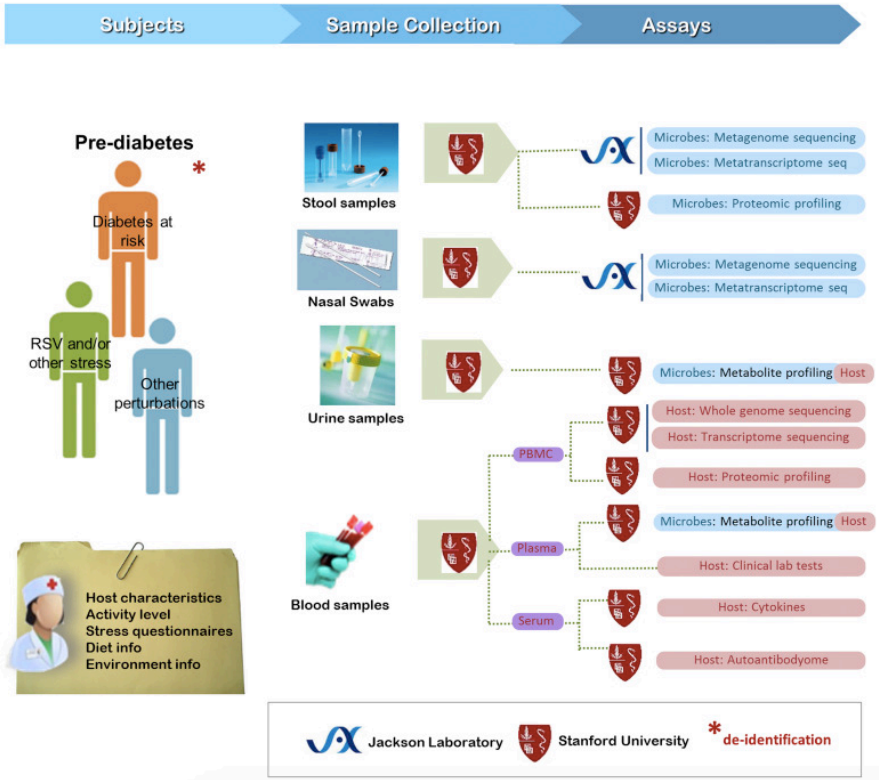


The Human Microbiome

Cell Host Microbe. 2014 Sep 10;16(3):276-89

iHMP creates integrated longitudinal datasets of biological properties from 3 different cohort studies of microbiome-associated conditions using multiple "omics" technologies.

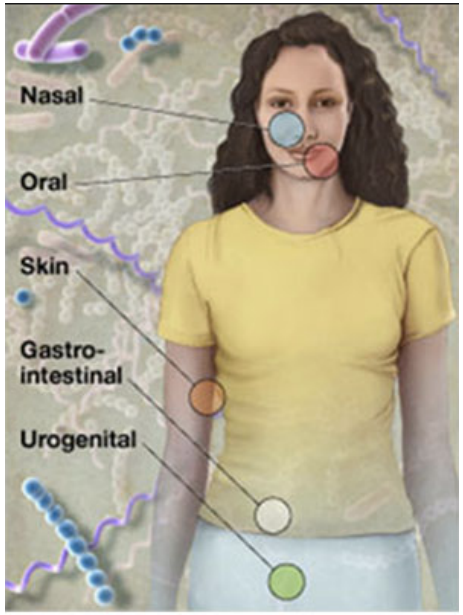
Project 3: Onset of Type 2 Diabetes

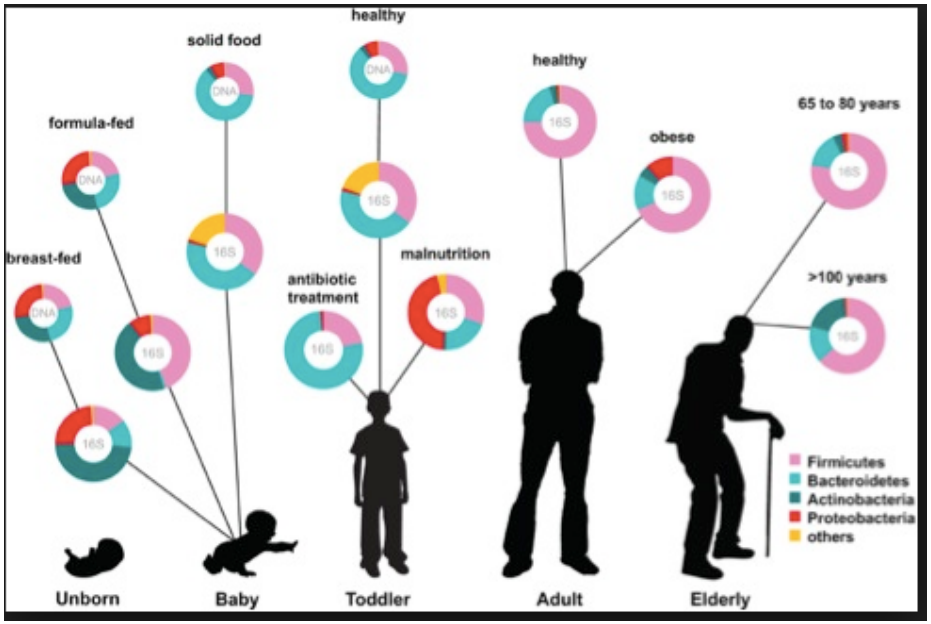


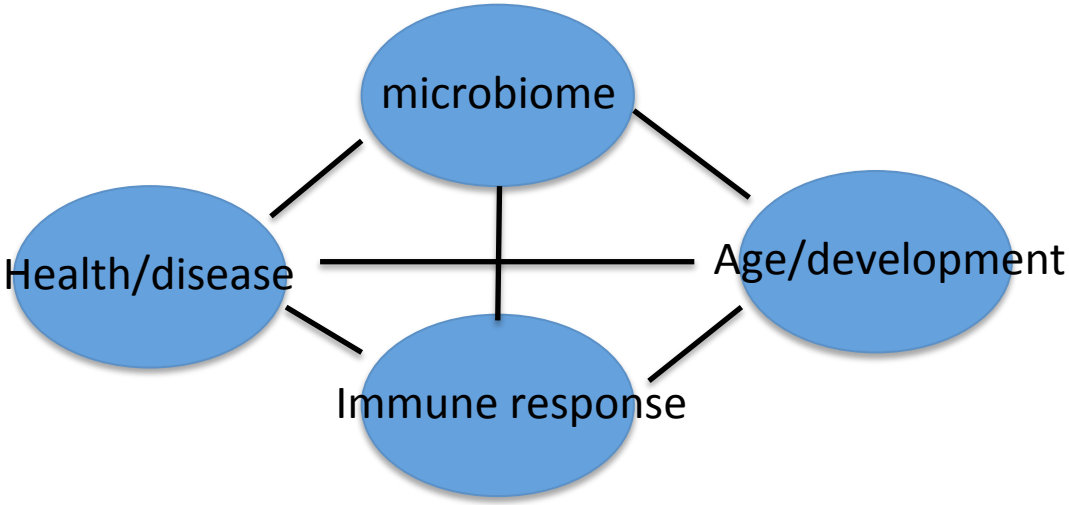
The Human Microbiome

- The human gut microbiome in health and disease
- mWGS

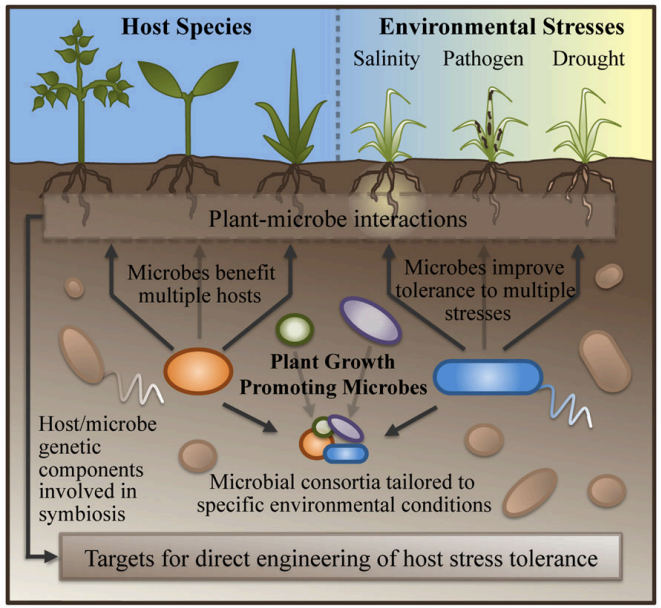
The Human Microbiome



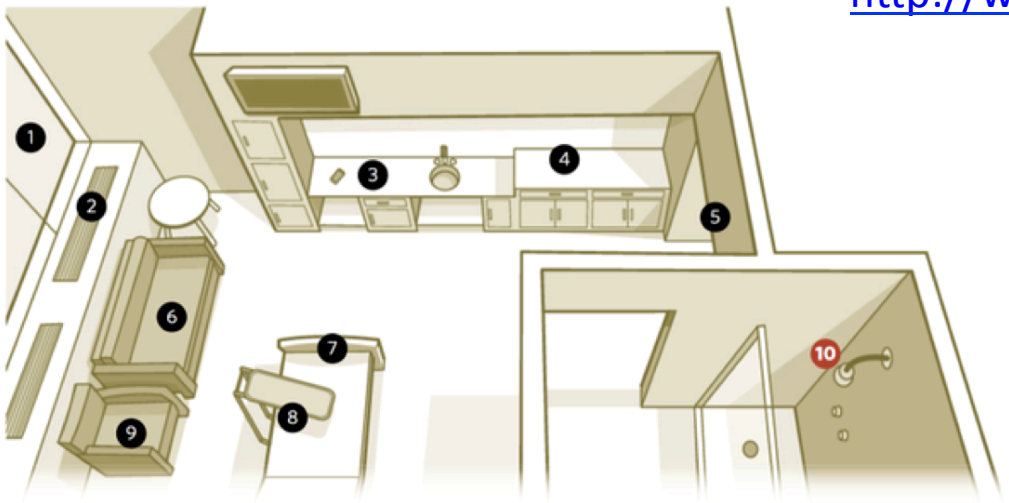




The Earth/Environmental Microbiome



http://www.earthmicrobiome.org/

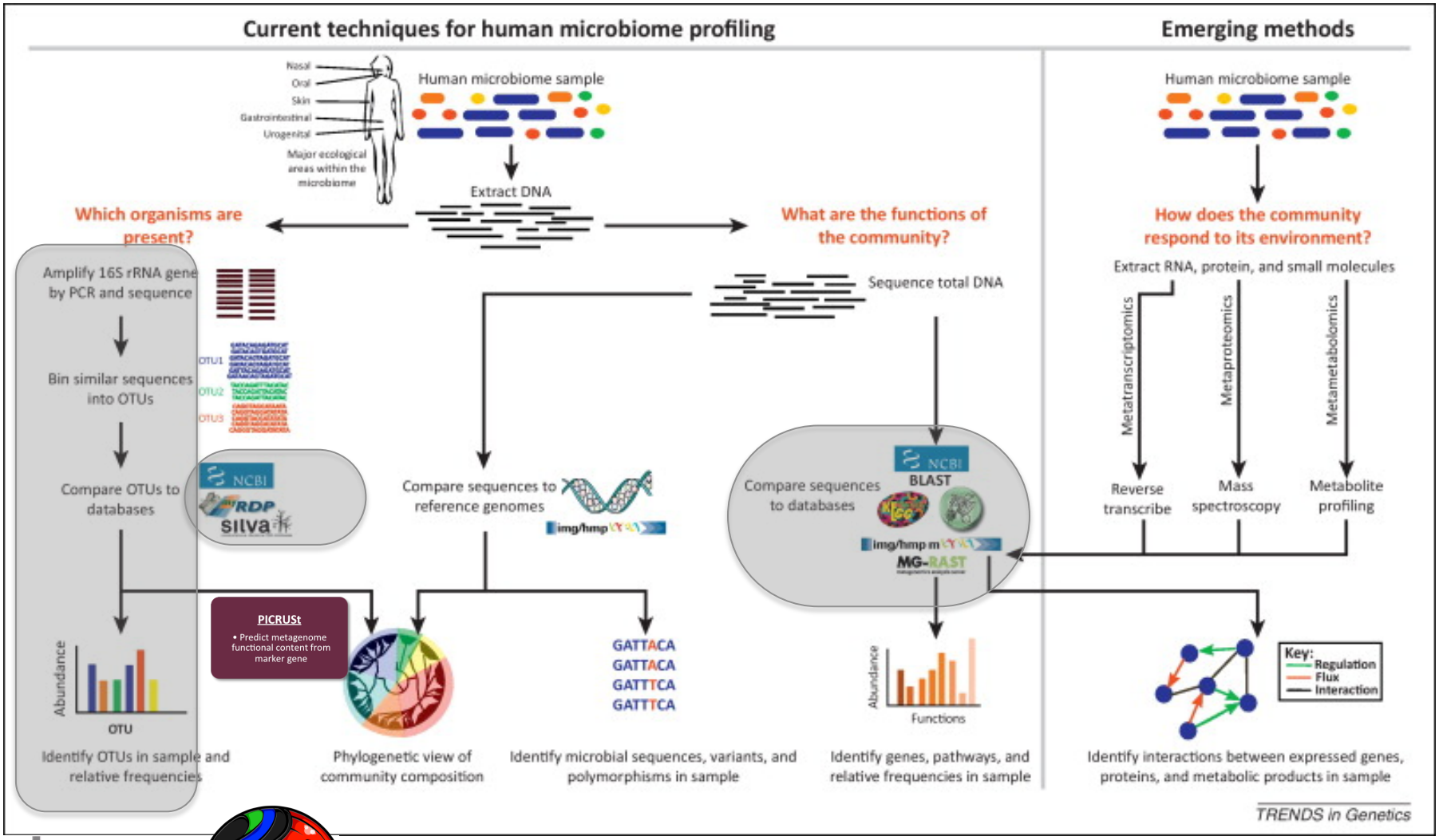


earth microbiome project

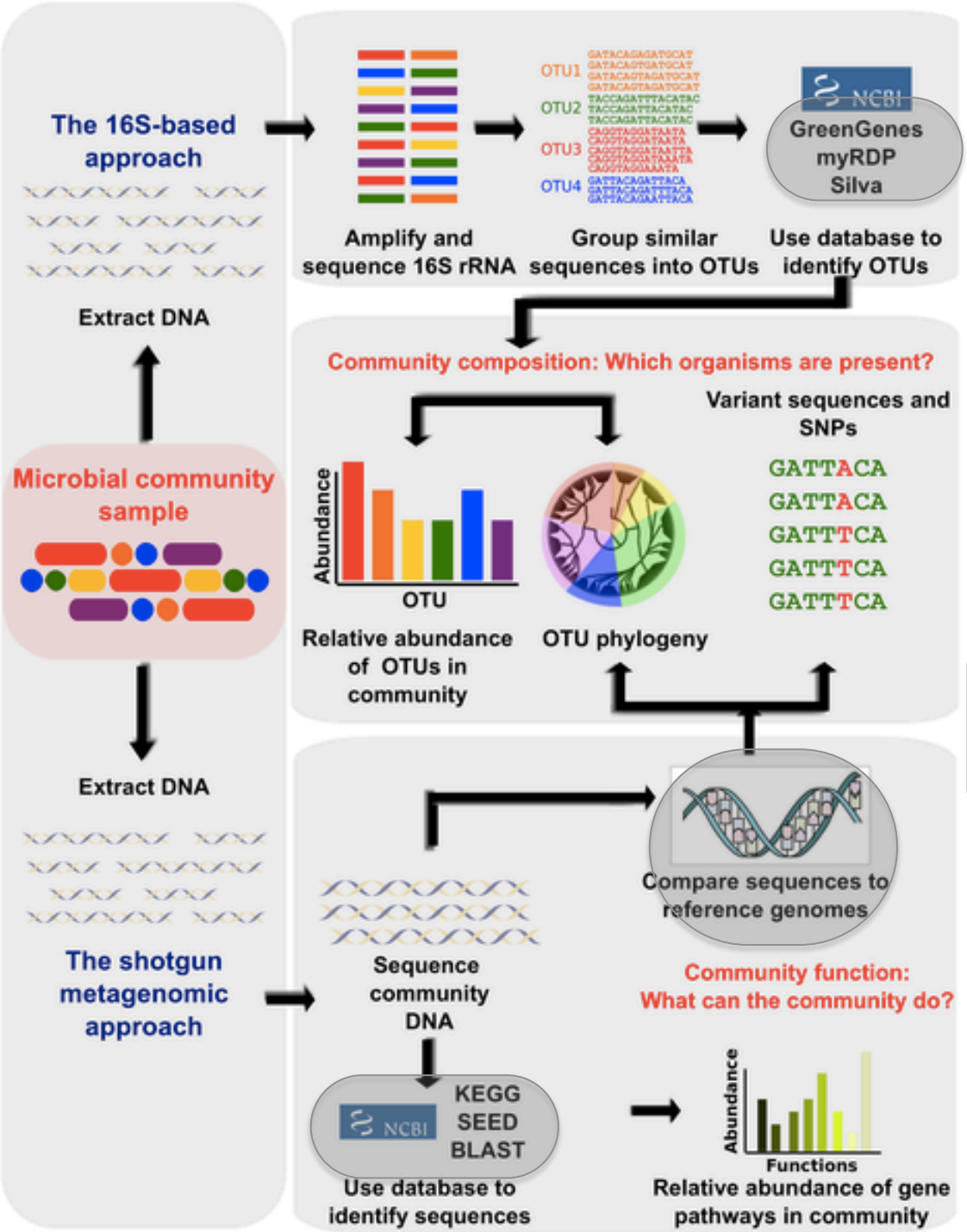
Workshop 11: Metagenomics Analysis

Shi, Baochen
Department of Pharmacology, UCLA

The Microbiome study



Xochitl Morgan et al. Trends in Genetics 2013



qiime
Quantitative Insights Into Microbial Ecology

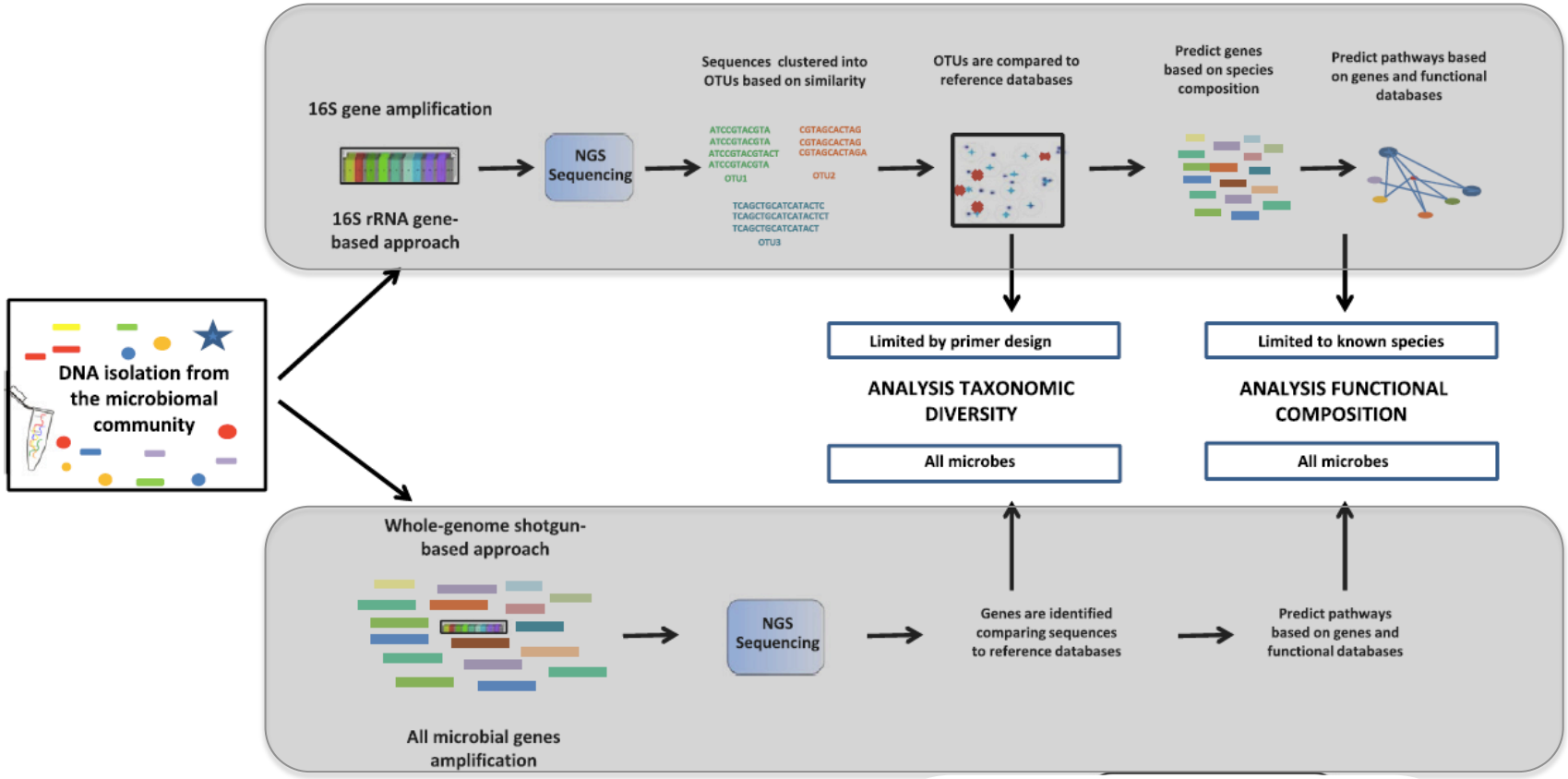
vs.

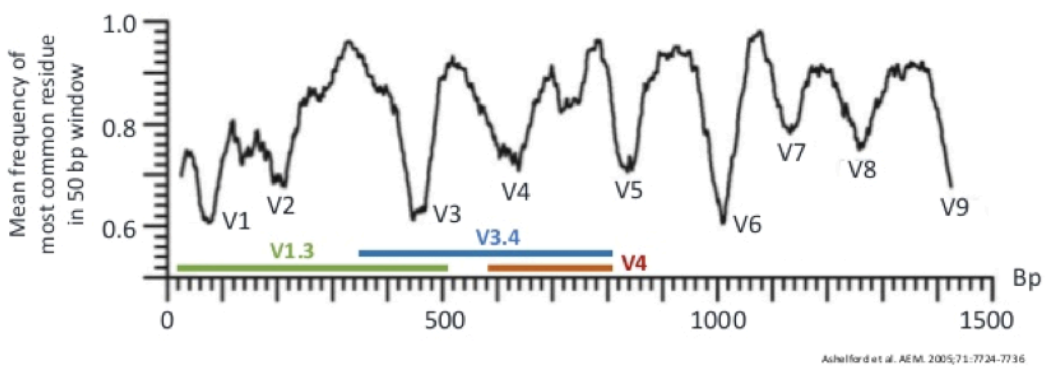
Mothur

PICRUSt

- Predict metagenome functional content from marker gene

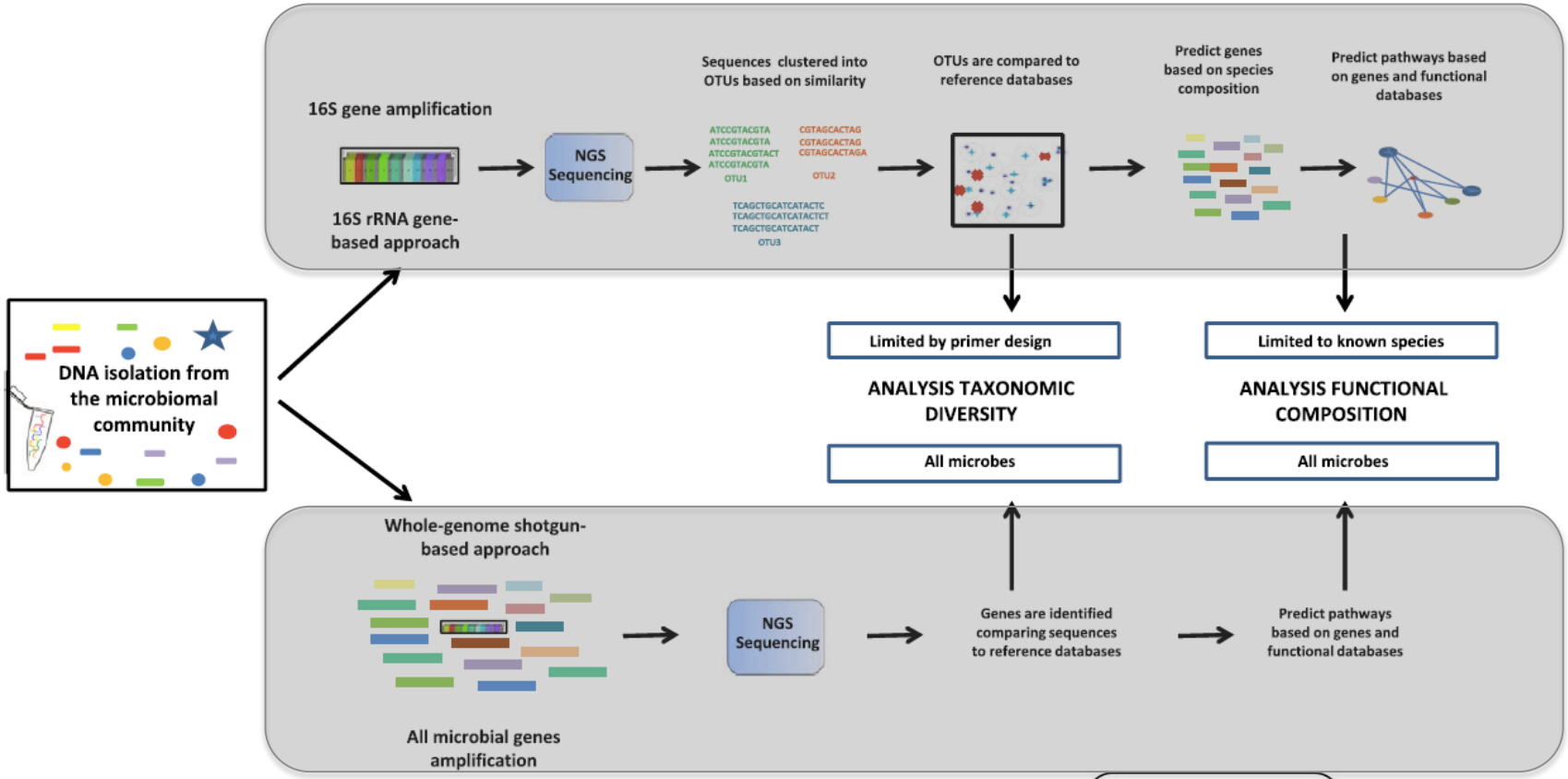
MG-RAST
metagenomics analysis server

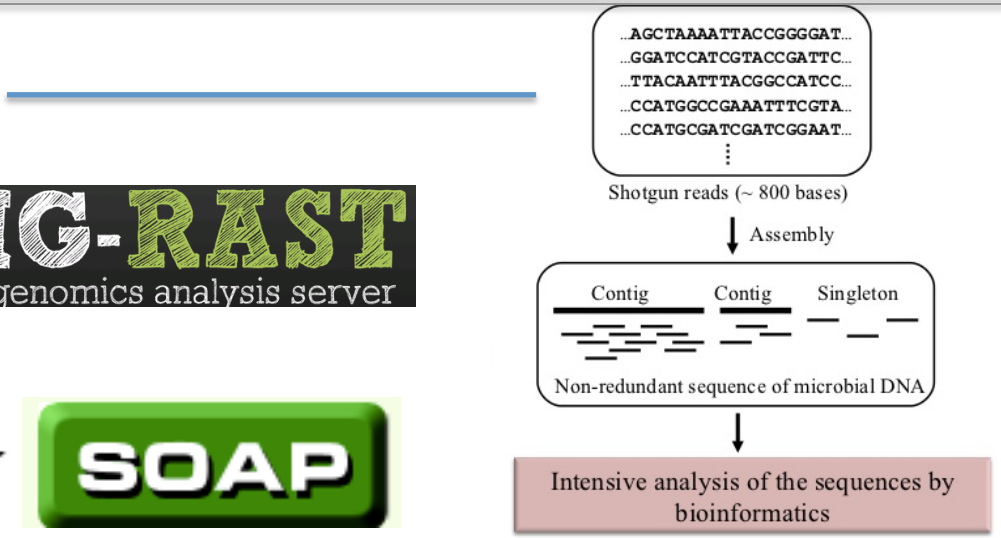




rdp

silva
high quality ribosomal RNA databases





Reference-based method

MG-RAST metagenomics analysis server

de novo assembly

Ray

SOAP

Outlines

The workshop: 2 hours per day over 3 days.

Day 1.
i) how to perform the 16S rRNA-based analysis using bioinformatics pipeline QIIME

Day 2.
i) Introduce statistical analyses in QIIME

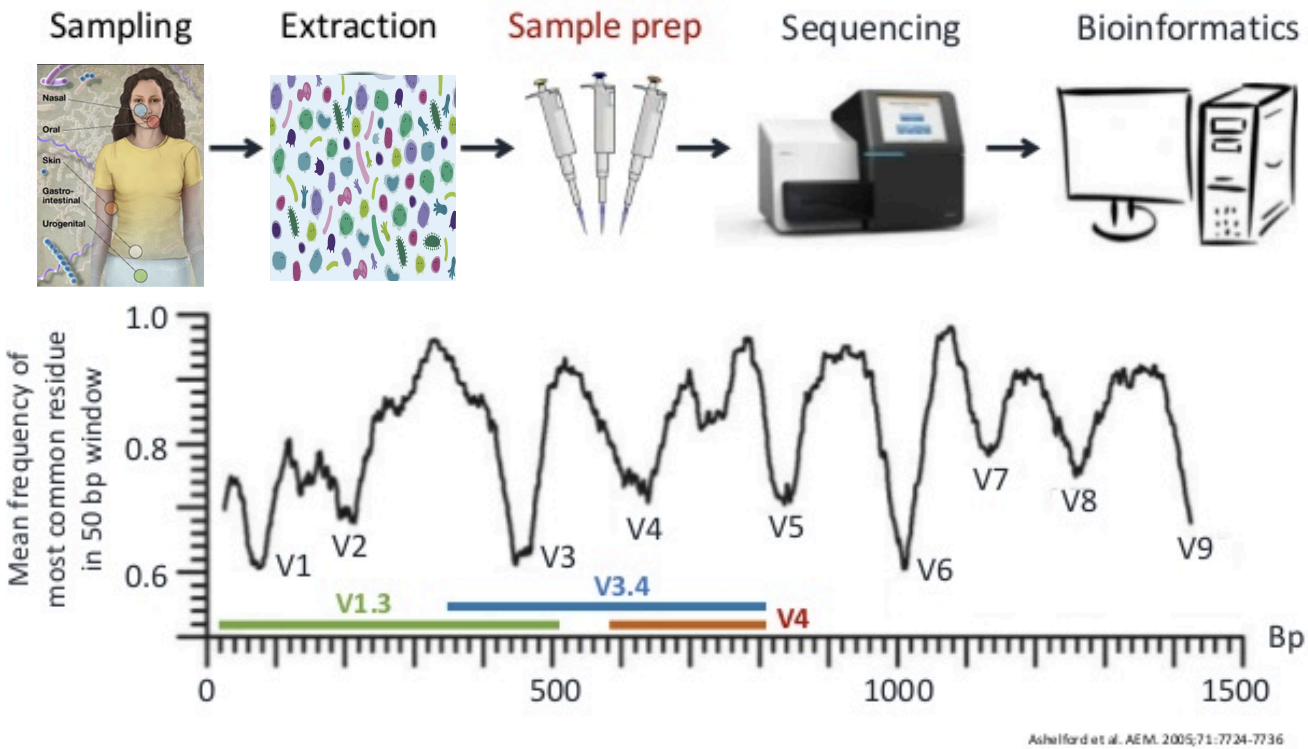
Day 3.

i) Functional analyses of the microbiome

ii) Open Q&A

Outlines

Typical workflow



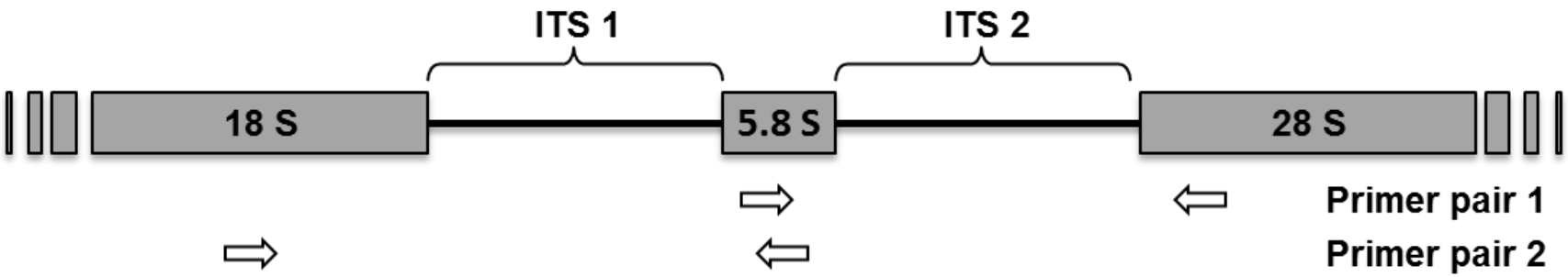
Universal, unbiased, distinguishable —>16S rRNA gene

V1-V3: human skin/vaginal microbiome

V3-V5: human gut microbiome

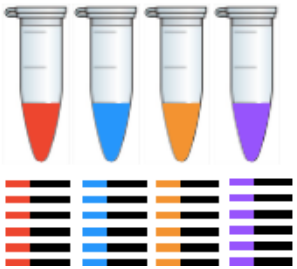
V4: human gut microbiome or soil microbiome

Outlines



fungal ITS sequences

Extract DNA and **PCR amplify** with barcoded primer



Pool amplicons

Pyrosequence amplicons using 454's GS FLX instrument

Screen, assign sequences to samples using barcode

```
>AGTGAGAGAAGCAGGGTCGTAATGTT . . .
>AGTGCGATGCGTAGGGTCGTAATGCG . . .
>AGTGCGATGCGTAGGGTCGTAATGTA . . .
>AGTGGATGCTCTAGGGTCGTAATGCA . . .
>AGTGTCACGGTGAGGGTCGTAATGGG . . .
>AGTGGATGCTCTAGGGTCGTAATGTT . . .
>AGTGTCACGGTGAGGGTCGTAATGCC . . .
>AGTGAGAGAAGCAGGGTCGTAATCAC . . .
. . .
```

qiime
Quantitative Insights Into Microbial Ecology

Outlines

We will demonstrate 16S amplicon analysis using QIIME

a) Run QIIME on Hoffman2 or local installation

b) Sequence data preparation

c) Operational Taxonomic Units (OTU) picking, Taxonomic assignment & inferring phylogeny

d) microbiome diversity analyses

Useful UNIX Commands: Covered in Workshop 1 & 2

- Where am I? pwd
- Current directory ./
- Home directory ~/
- Change directory cd ~/data
- Move up one level cd ..
- List files in folder ls
- Look at a file less fileName
- Copy a file cp ~/data/file ~/otherdir/
- Delete a file rm fileName
- Delete a directory rmdir ~/dirName/
- Move a file mv ~/data/file ~/otherdir/file
- Secure copy scp user@host1:dir/file user@host2:dir/file
- Compress a file gzip –c file > file.gz
- Uncompress a file gunzip file.gz
- Make a new folder mkdir data2
- Count lines in a file wc –l fileName

Questions/In Doubt/Lost?

- Unix manual: man functionX
- Google is your best friend!
- Any of your friendly QCB fellows

a) Run on Hoffman2

idre UCLA Institute for Digital Research and Education

A shared account on mac for students:

login: workshop
password: NGS_Analysis

Login in hoffman2:

```
ssh biosbc@hoffman2.idre.ucla.edu
qrsh
```

Qiime is in :

```
cd /u/local/apps
```

Load Qiime:

```
module load qiime
```

Logging into Hoffman2

ssh biosbc@hoffman2.idre.ucla.edu

```
$ ssh biosbc@hoffman2.idre.ucla.edu
biosbc@hoffman2.idre.ucla.edu's password:
Welcome to the Hoffman2 Cluster!
```

change to your user ID

a) Run on Hoffman2

idre UCLA Institute for Digital Research and Education

A shared account on mac for students:

login: workshop
password: NGS_Analysis

Login in hoffman2:

```
ssh biosbc@hoffman2.idre.ucla.edu
qrsh
```

Qiime is in :

```
cd /u/local/apps
```

Load Qiime:

```
module load qiime
```

Useful Tools on Hoffman2

ls /u/local/apps/

```
abaqus
abaqusdocs
accelrys
Accelrys
activeperl
ActiveTcl
adina
Adobe
affymetrix
AFNI
aida
amber
ampliconnoise
anaconda
annovar
ansys_inc
ant
antlr
armadillo
arpack
atlas
atomeye
atompaw
autoconf
automake
bamtools
bcftools
bcl2qfastq
beagle
beaglecall
beagle_utilities
bedtools
bfast
bioscope
blas
blast
blast+
blat
blcr
blitz
bmapuclatools
boost
boost-jam
bowtie
bowtie2
bwa
caffe
casava
cdbtools
cd-hit
cernlib
cern_root
clearcut
clhep
cln
cmake
common
comsol
conan
condor
consed
cp2k
cpmd
ctffind
cufflinks
cytoscape
ddd
ddplot
dejagnu
dirac
diskusage
dl_poly
dx
eclipse
eigen
eigensoft
emacs
espresso
fastphase
fasttree
fastx_toolkit
fbat
fe-safe
ffmpeg
fftw2
fftw3
flex
fltk
freebayes
freesurfer
freetds
fribidi
fsl
gamess
gatk
gatk-queue
gaussian
gaussview
gcta
gdal
gdc
geant4
genetorrent
geos
gflags
gftp
ginac
git_OLD
globalarrays
glog
gmp
gnuplot
gpac
grace
grads
graphicsmagick
graphviz
gromacs
gsl
hadoop
handbrake
harminv
haskell
hdf
hdf5
homer
hyperworks
iaida
idl
idr
igraph
ilog
IM-IMA
impute
imsl
infernal
inspect
installation
intel
isight
jags
jam
java
jaxodraw
jmol
julia
kepler
lammps
lapack
lapack++
ldope
leveldb
libgtextutils
libtool
libxc
lmdb
loni_pipeline
lumerical
lynx
mach
mafft
manorm
maple
maq
maqview
mathematica
matio
matlab
mats
mecab
meep
meld
merlin
metis
mfold
microbiomeutil
migrate
minirosetta
mira
molcas
molden
molpro
mono
mopac
mothur
mpc
mpfr
mpiblast
mplayer
mrt
mumps
muscle
mygroup
mysql
namd
ncl
nco
netbeans
netcdf
newbler
ngsplot
nlopt
numpy
nwchem
ocaml
octave
omssa
oommf
openbabel
opencv
openfoam
openmotif
opensees
osiris
osu-micro-benchmarks
papi
paraview
parmetis
parsec
parsinsert
pcre
pdt
penncnv
perl_modules
petsc
pgsql
phantompeakqualtools
phase
picard-tools
plink
pop-c++
povray
pplcer
preseq
proj.4
protobuf
pypy
pypy3
python
qcachegrind
qchem
qgfe
qhull
qiime
qiime_data
qrupdate
qsub
qt
quantiSNP
queue.logs
R
raxml
rdp_classifier
relion
remcom
repeatmasker
RepeatModeler
resmap
rosetta
rstudio
rsync
rtax
ruby
sage
samtools
scalapack
scilab
scons
sentaurus
shapeit
shapelib
skampi
slots
snappy
solar
sortmerna
soxr
splicetrap
spm
stata
stressapptest
structure
subversion
suitesparse
sundials
superlu_dist
swarm
sysstat
szip
t500
tau
tcl
tcsh
tecplot360
texlive
tmux
tophat
toscastructure
totalview
trans-ABySS
treemix
trilinos
trinity
tt
tvmet
uclust
udunits
usearch
usr_lib_GLoverride
valgrind
vasp
vcftools
vegas
ViennaRNA
vim
visit
vmd
votca
vtk
wannier90
weblogo
wolfram
x264
x86_open64
xerces-c
xfdtd
xfig_OLD
xilinx-vivado
xmd
xmedcon
xpdf_OLD
yaml
yasm
zlib
```

a) Run on Hoffman2

idre UCLA Institute for Digital Research and Education

A shared account on mac for students:

login: workshop
password: NGS_Analysis

Login in hoffman2:

```
ssh biosbc@hoffman2.idre.ucla.edu
qrsh
```

Qiime is in :

```
cd /u/local/apps
```

Load Qiime:

```
module load qiime
```

On Hoffman2 Interactively

#grab a Hoffman2 session:

qrsh

or

qrsh -now n -l i,mem=5G,time=02:00:00,exclusive=TRUE -pe shared 6

Request exact configuration

Reserve the whole node

time limit of the session
(default is 2 hours, 24hrs max)

Request 6 cores

Memory requested per core (default is 1G)

a) MacQIIME local installation

QIIME have a lot of dependencies .
The quickest way to get started using QIIME VirtualBox or MacQIIME
http://qiime.org/1.9.0/install/install.html

MacQIIME:
a1) Download MacQIIME (http://www.wernerlab.org/software/macqiime/macqiime-installation)

```
tar -xvf MacQIIME_*.tgz
```

a2) install MacQIIME
Copy macqiime folder to root, and then copy "macqiime" script to /usr/bin/

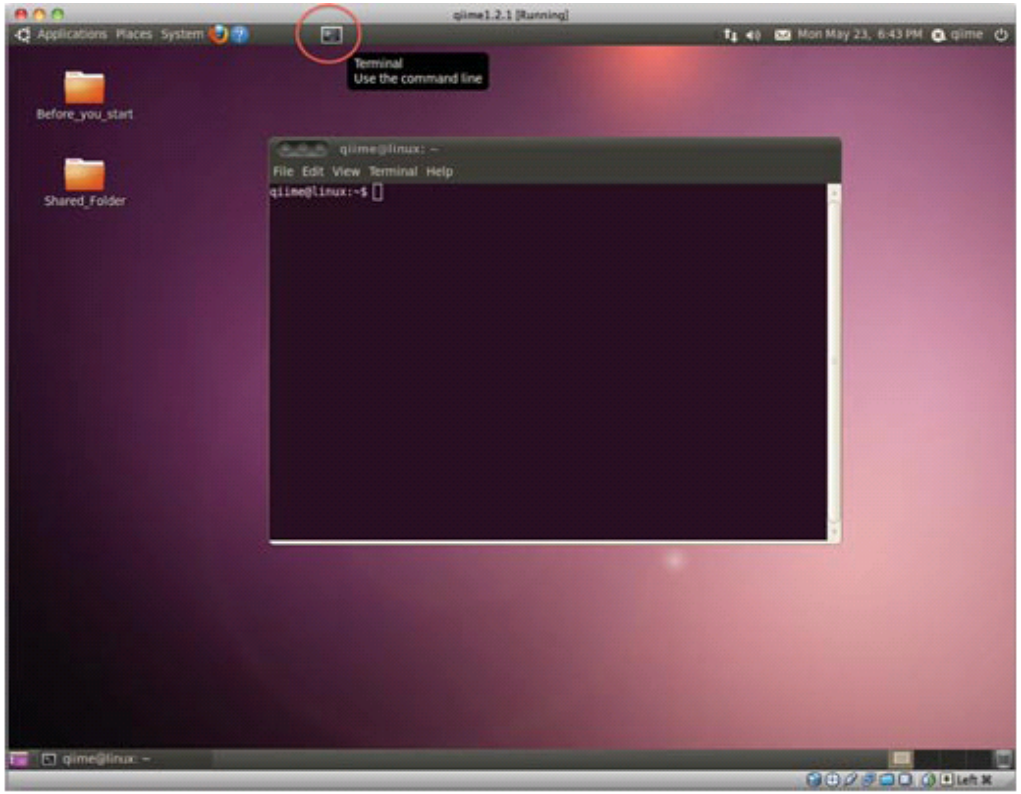
```
./install.s
```

source the environment variables

```
source /macqiime/configs/bash_profile.txt
```

Tutorial, test data & test script:
/macqiime/QIIME/

a) QIIME VirtualBox installation



QIIME VirtualBox (Linux, windows, Mac):

a1) Download and install oracle VirtualBox (http://qiime.org/1.9.0/install/virtual_box.html)

a2) Download the QIIME Virtual Box

a3) Create a new virtual machine

Outlines

We will demonstrate steps for

a) Run QIIME on Hoffman2 or local installation

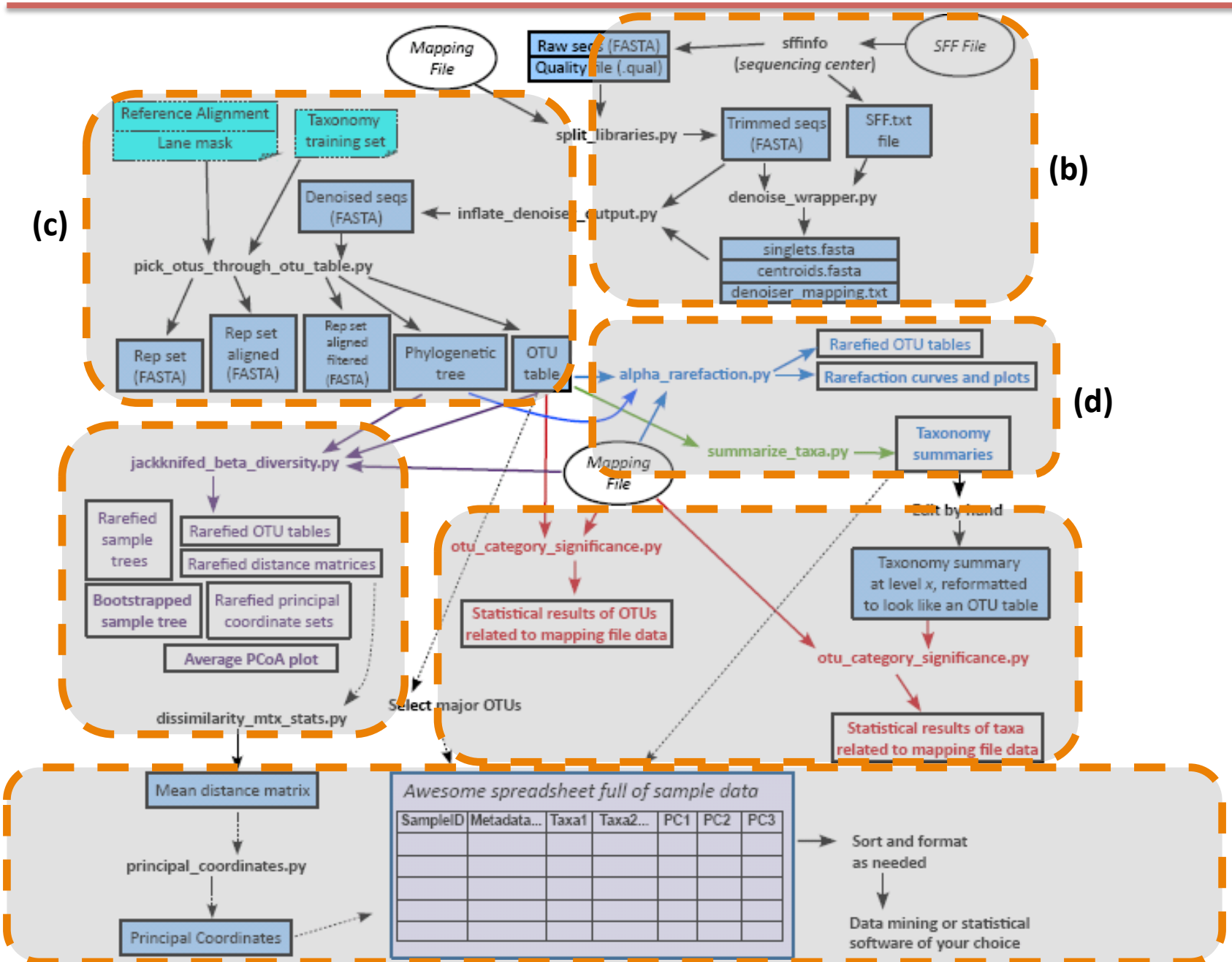
b) Sequence data preparation

c) Operational Taxonomic Units (OTU) picking, Taxonomic assignment & inferring phylogeny

d) microbiome diversity analyses

Quantitative Insights Into Microbial Ecology

Flowchart


Mapping File
Raw seqs (FASTA)
Quality file (.qual)
sffinfo (sequencing center)
SFF File
(b)
split_libraries.py
Trimmed seqs (FASTA)
SFF.txt file
denoise_wrapper.py
inflate_denoiser_output.py
singlets.fasta
centroids.fasta
denoiser_mapping.txt
(c)
Reference Alignment
Lane mask
Taxonomy training set
Denoised seqs (FASTA)
pick_otus_through_otu_table.py
Rep set (FASTA)
Rep set aligned (FASTA)
Rep set aligned filtered (FASTA)
Phylogenetic tree
OTU table
(d)
alpha_rarefaction.py
Rarefied OTU tables
Rarefaction curves and plots
summarize_taxa.py
Taxonomy summaries
Mapping File
jackknifed_beta_diversity.py
Rarefied sample trees
Rarefied OTU tables
Rarefied distance matrices
Bootstrapped sample tree
Rarefied principal coordinate sets
Average PCoA plot
dissimilarity_mtx_stats.py
Edit by hand
otu_category_significance.py
Statistical results of OTUs related to mapping file data
Taxonomy summary at level x, reformatted to look like an OTU table
otu_category_significance.py
Statistical results of taxa related to mapping file data
Select major OTUs
Mean distance matrix
principal_coordinates.py
Principal Coordinates
Awesome spreadsheet full of sample data
SampleID
Metadata...
Taxa1
Taxa2...
PC1
PC2
PC3
Sort and format as needed
Data mining or statistical software of your choice

Flowchart

1. SFF (raw 454 data, optional)
2. fasta/qual files
3. demultiplexing/quality filtering

(b) Sequence data preparation

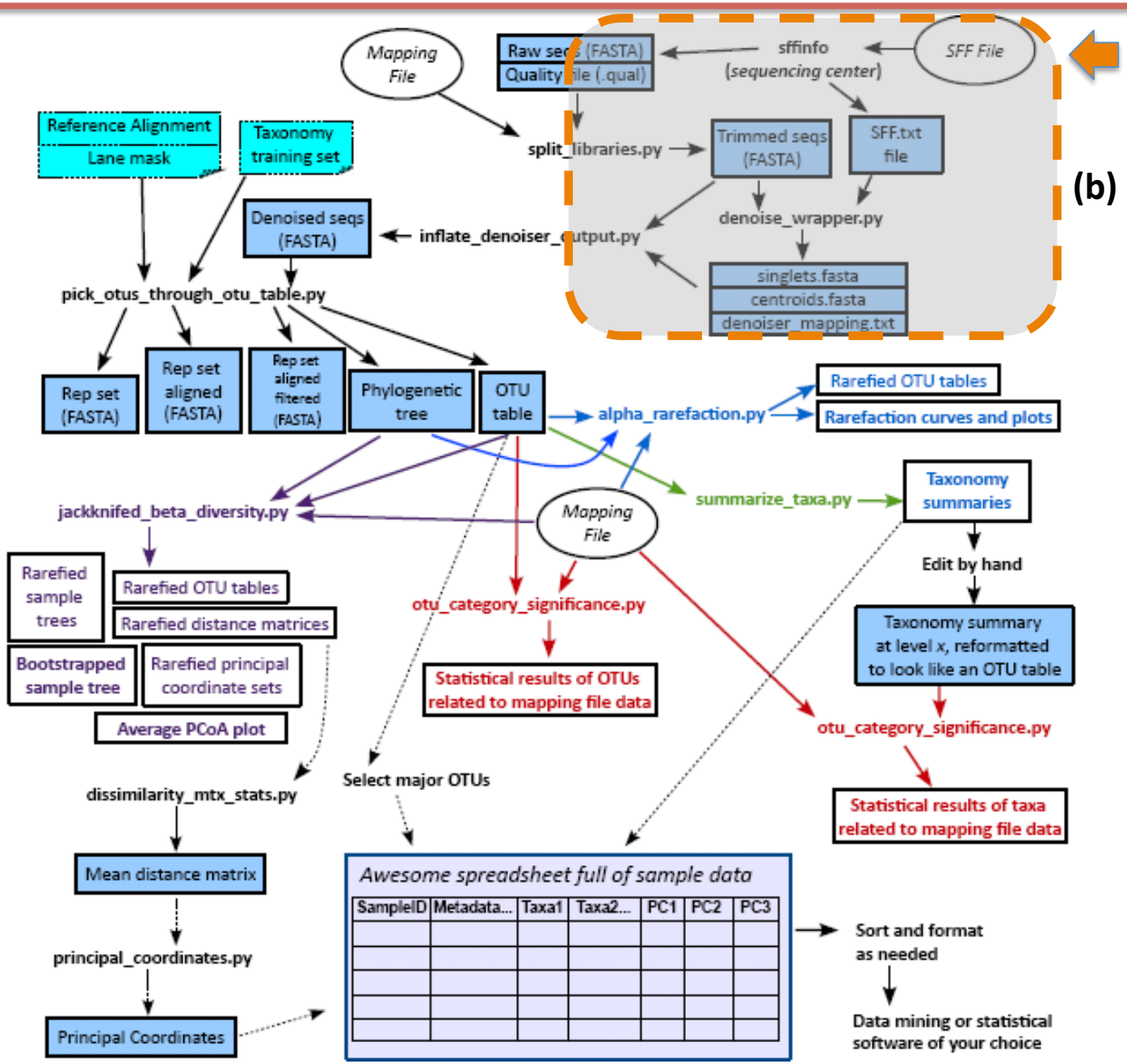
4. OTU picking
5. representative sequences
6. taxonomic assignments/tree building

(c) Operational Taxonomic Units (OTU) picking, Taxonomic assignment & inferring phylogeny

7. OTU table and downstream processing

(d) microbiome diversity analyses

Flowchart


Mapping File
Raw seqs (FASTA)
Quality file (.qual)
sffinfo
(sequencing center)
SFF File
Reference Alignment
Lane mask
Taxonomy training set
split_libraries.py
Trimmed seqs (FASTA)
SFF.txt file
(b)
Denoised seqs (FASTA)
inflate_denoiser_output.py
denoise_wrapper.py
singlets.fasta
centroids.fasta
denoiser_mapping.txt
pick_otus_through_otu_table.py
Rep set (FASTA)
Rep set aligned (FASTA)
Rep set aligned filtered (FASTA)
Phylogenetic tree
OTU table
alpha_rarefaction.py
Rarefied OTU tables
Rarefaction curves and plots
summarize_taxa.py
Taxonomy summaries
jackknifed_beta_diversity.py
Mapping File
Edit by hand
Rarefied sample trees
Rarefied OTU tables
Rarefied distance matrices
otu_category_significance.py
Taxonomy summary at level x, reformatted to look like an OTU table
Bootstrapped sample tree
Rarefied principal coordinate sets
Statistical results of OTUs related to mapping file data
Average PCoA plot
otu_category_significance.py
Select major OTUs
dissimilarity_mtx_stats.py
Statistical results of taxa related to mapping file data
Mean distance matrix
Awesome spreadsheet full of sample data
SampleID
Metadata...
Taxa1
Taxa2...
PC1
PC2
PC3
Sort and format as needed
principal_coordinates.py
Data mining or statistical software of your choice
Principal Coordinates

b) Data preparation

We will start out with raw sequencing data generated on 454

- Sequences (.fna): 454 generated FASTA file.
- Quality Scores (.qual): 454 generated quality score file
- experimental data about the samples (Mapping File) generated by user.

#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
#Example mapping file for the QIIME analysis package.					
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

On Hoffman2, copy all the files you need to scratch folder

```
cd $Scratch
pwd

cd /u/home/b/biosbc/
cd /u/local/apps/qiime/1.8.0/examples/qiime_tutorial/
cp /u/local/apps/qiime/1.8.0/examples/qiime_tutorial/*  /u/home/b/biosbc/
```

change to your ID

Quantitative Insights Into Microbial Ecology

b) Data preparation

Standard flowgram format (SFF) is a binary file format used to encode results of 454 pyrosequencing

```
process_sff.py -i Fasting_Example.sff -f -o output_dir
```

-f --make_flowgram

- Sequences (.fna): 454 generated FASTA file.
- Quality Scores (.qual): 454 generated quality score file

b) Data preparation

Mapping File

#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
#Example mapping file for the QIIME analysis package.					
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

assign the multiplexed reads to samples based on their nucleotide barcode (*demultiplexing*)

```
split_libraries.py -m Fasting_Map.txt -f Fasting_Example.fna -q Fasting_Example.qual -o split_library_output
```

Quantitative Insights Into Microbial Ecology

b) Data preparation

- **split_library_log.txt** : summary of demultiplexing and quality filtering, including the number of reads detected for each sample and a brief summary of any reads that were removed due to quality considerations.
- **histograms.txt** : tab-delimited file shows the number of reads at regular size intervals before and after splitting.
- **seqs.fna** : fasta formatted

This step also performs quality filtering based on the characteristics of each sequence, removing any low quality or ambiguous reads.

```
split_libraries.py -m Fasting_Map.txt -f Fasting_Example.fna -q Fasting_Example.qual -o split_library_output
```

-l, --min_seq_length
Minimum sequence length, in nucleotides [default: 200]
-t, --trim_seq_length
Calculate sequence lengths after trimming primers and barcodes [default: False]
-s, --min_qual_score
Min average qual score allowed in read [default: 25]

Quantitative Insights Into Microbial Ecology

b) Data preparation

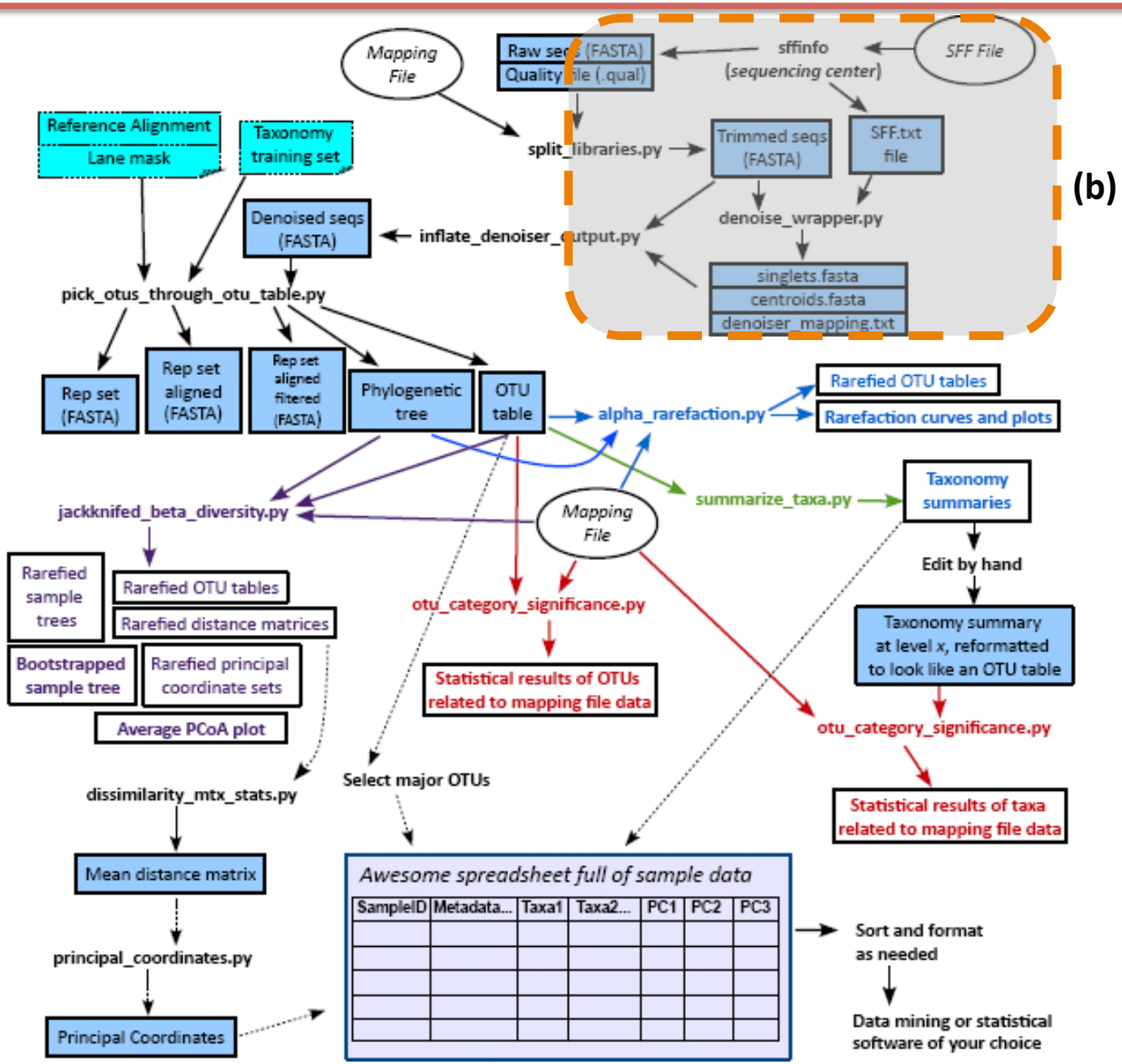
Illumina runs:
demultiplexing of Fastq sequence data where barcodes and sequences are contained in two separate fastq files.

Demultiplex and quality filter (at Phred >= Q20) one lane of Illumina fastq data

```
split_libraries_fastq.py –i  fastq.gz –b barcode.fastq.gz --rev_comp_mapping_barcodes –o outdir/ -m Fasting_Map.txt -q 19
```

Quantitative Insights Into Microbial Ecology

Flowchart


Mapping File
Raw seqs (FASTA)
Quality file (.qual)
sffinfo
(sequencing center)
SFF File
Reference Alignment
Lane mask
Taxonomy training set
split_libraries.py
Trimmed seqs (FASTA)
SFF.txt file
(b)
Denoised seqs (FASTA)
inflate_denoiser_output.py
denoise_wrapper.py
singlets.fasta
centroids.fasta
denoiser_mapping.txt
pick_otus_through_otu_table.py
Rep set (FASTA)
Rep set aligned (FASTA)
Rep set aligned filtered (FASTA)
Phylogenetic tree
OTU table
alpha_rarefaction.py
Rarefied OTU tables
Rarefaction curves and plots
summarize_taxa.py
Taxonomy summaries
jackknifed_beta_diversity.py
Mapping File
Edit by hand
Rarefied sample trees
Rarefied OTU tables
Rarefied distance matrices
otu_category_significance.py
Taxonomy summary at level x, reformatted to look like an OTU table
Bootstrapped sample tree
Rarefied principal coordinate sets
Statistical results of OTUs related to mapping file data
Average PCoA plot
otu_category_significance.py
Select major OTUs
dissimilarity_mtx_stats.py
Statistical results of taxa related to mapping file data
Mean distance matrix
Awesome spreadsheet full of sample data
SampleID
Metadata...
Taxa1
Taxa2...
PC1
PC2
PC3
Sort and format as needed
principal_coordinates.py
Data mining or statistical software of your choice
Principal Coordinates