

Metagenome Report

Project: tutorial

Date: 07/18/2017

Contents

Introduction	1
Quality Control	2
DNA Samples Quality Control	3
DNA Samples Tables of Filtered Reads	3
DNA Samples Plots of Filtered Reads	5
Taxonomic Profiling of Metagenomic Reads	5
Species Count Table	6
Ordination	7
Heatmap	8
Barplot	9
Functional Profiling of Metagenomic Reads	10
Pathway Abundance	10
Features	13
Data Processing Workflow Information	15
Software Versions	15
Tasks Run	15

Introduction

The data was run through the standard workflow for whole metagenome shotgun sequencing.

Quality Control

This report section contains information about the quality control processing for all 6 single-end fastq input files. These files were run through the [KneadData](#) QC pipeline. Reads were first trimmed then filtered against contaminate reference databases: hg38_demo and rRNA_demo. Reads were filtered sequentially with those reads passing the first filtering step used as input to the next filtering step. This chain of filtering removes reads from all references in serial. The tables and plots are annotated as follows:

- raw : Untouched fastq reads.
- trim : Number of reads remaining after trimming bases with Phred score < 20. If the trimmed reads is < 70% of original length then it is removed altogether.
- hg38_demo : Number of reads after depleting against reference database hg38_demo.
- rRNA_demo : Number of reads after depleting against reference database hg38_demo and rRNA_demo.

DNA Samples Quality Control

DNA Samples Tables of Filtered Reads

	DNA reads			
	Raw	Trim	hg38_demo	rRNA_demo
HD32R1_subsample	35,200	35,104	35,031	34,935
HD42R4_subsample	27,249	27,183	27,117	27,063
HD48R4_subsample	30,251	30,144	30,076	29,991
LD96R2_subsample	108,134	108,064	107,976	107,891
LV16R4_subsample	86,293	86,209	86,122	86,039
LV20R4_subsample	61,076	61,013	60,948	60,854

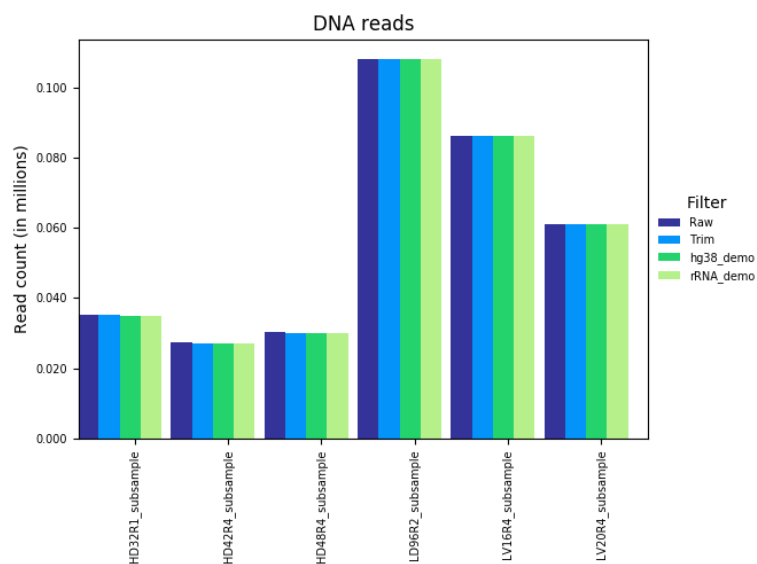
A data file exists of this table: [qc_counts_table.tsv](#)

DNA microbial read proportion				
	hg38_demo / Trim	hg38_demo / Raw	rRNA_demo / Trim	rRNA_demo / Raw
HD32R1_subsample	0.99792	0.99520	0.99519	0.99247
HD42R4_subsample	0.99757	0.99516	0.99559	0.99317
HD48R4_subsample	0.99774	0.99422	0.99492	0.99141
LD96R2_subsample	0.99919	0.99854	0.99840	0.99775
LV16R4_subsample	0.99899	0.99802	0.99803	0.99706
LV20R4_subsample	0.99893	0.99790	0.99739	0.99637

Proportion of reads remaining after removing host reads relative to the number of: i) quality-trimmed reads, and ii) raw unfiltered reads.

A data file exists of this table: [microbial_counts_table.tsv](#)

DNA Samples Plots of Filtered Reads



Taxonomic Profiling of Metagenomic Reads

This report section contains information about the taxonomy for all DNA samples. These samples were run through [MetaPhlAn2](#).

Species abundances are passed through a basic filter requiring each species to have at least 0.01 % abundance in at least 10 % of all samples.

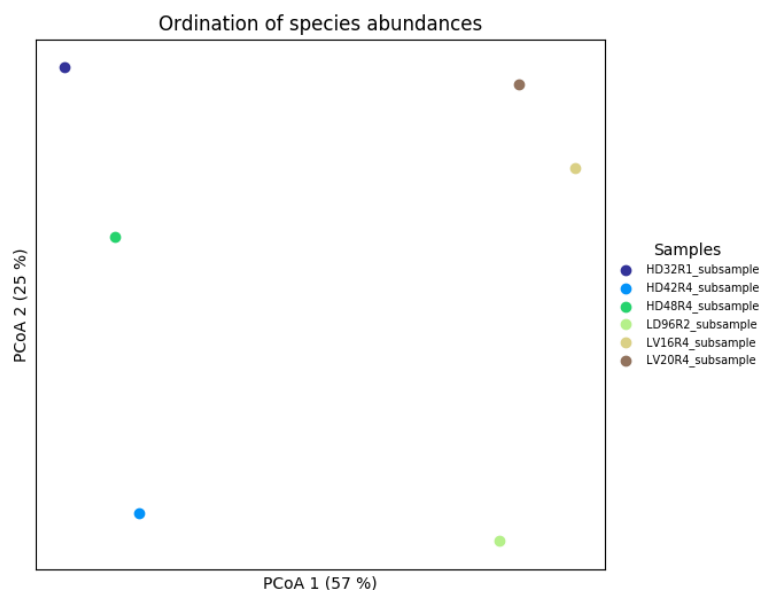
A total of 25 species were identified. After basic filtering 25 species remained.

Species Count Table

Total species per sample		
	Total	After filter
HD32R1_subsample	17	17
HD42R4_subsample	16	16
HD48R4_subsample	20	20
LD96R2_subsample	13	13
LV16R4_subsample	16	16
LV20R4_subsample	19	19

A data file exists of this table: [species_counts_table.tsv](#)

Ordination

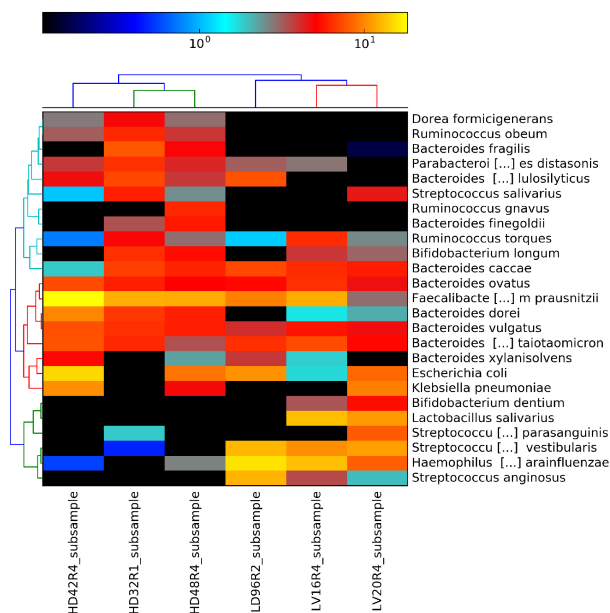


Principal coordinate analysis of variance among samples, based on Bray-Curtis dissimilarities between species profiles of samples. Filtered species' relative abundances were arcsin-square root transformed to approximate a normal distribution and down-weight the effect of highly abundant species on Bray-Curtis dissimilarities. Numbers in parenthesis on each axis represent the amount of variance explained by that axis.

Heatmap

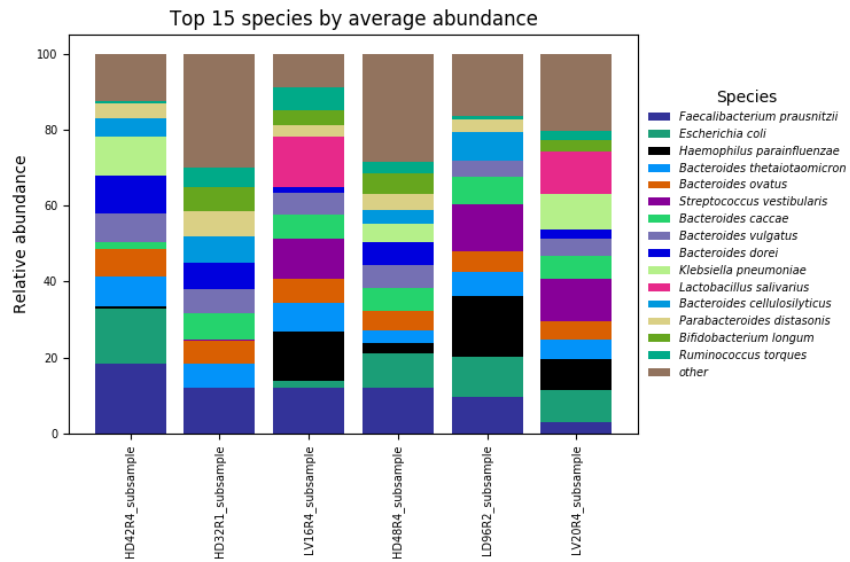
The top 25 species based on average relative abundance are shown in the heatmap. The heatmap was generated with [Hclust2](#).

Top 25 species by average abundance



Hierarchical clustering of samples and species, using top 25 species with highest mean relative abundance among samples. Abundances were log10 transformed prior to clustering, and the “average linkage” clustering on the Euclidean distance metric was used to cluster samples. The species dendrogram is based on pairwise (Spearman) correlation between species. Samples are columns and species are rows. The color bar represents relative abundances on a log10 scale.

Barplot



Stacked barplot of 15 most abundant species among samples. Samples in the plot were sorted on the species with the highest mean abundances among samples, in decreasing order.

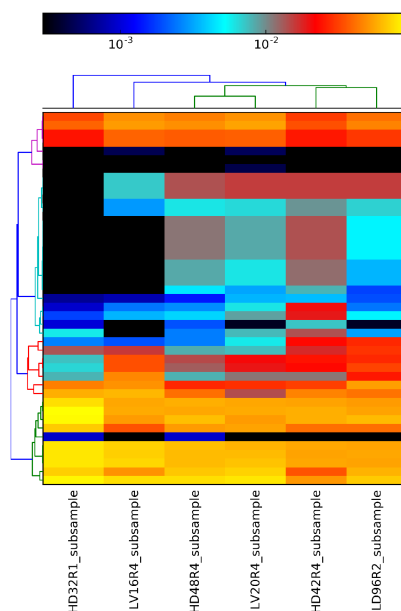
Functional Profiling of Metagenomic Reads

This report section contains preliminary exploratory figures that summarize HUMAnN2 functional profiling of all samples. HUMAnN2 performs species-specific and species-agnostic quantification of gene families, EC enzyme modules, and pathways, using the UniRef and MetaCyc databases. For more information on functional profiling and the databases used, see websites for [HUMAnN2](#), [UniRef](#), and [MetaCyc](#).

Pathway Abundance

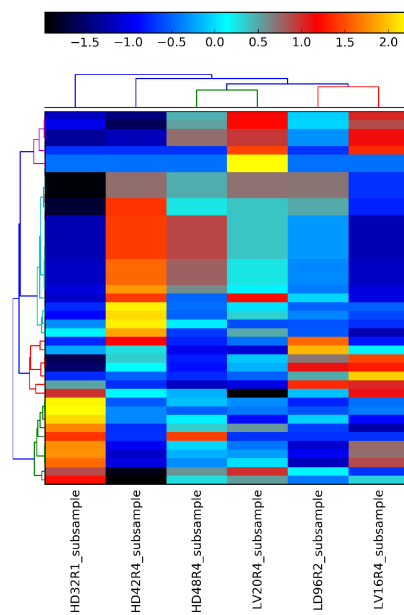
Hierarchical clustering of samples and pathways, using top 50 pathways with highest mean relative abundance among samples. The ‘average linkage’ clustering on the Euclidean distance metric was used to cluster samples. The pathway dendrogram is based on pairwise (Spearman) correlation between pathways. Samples are columns and pathway are rows. The heatmaps were generated with [Hclust2](#).

Top 50 pathways by average abundance



Abundances were log10 transformed prior to clustering. The color bar represents relative abundances on a log10 scale.

Top 50 pathways by average abundance



Abundances were z-score transformed prior to clustering. The color bar represents relative abundances on a z-score scale.

Top 50 pathways by average abundance (partial table)

	Average	Variance
PWY-7221: guanosine ribonucleotides de novo biosynthesis	0.0736	0.000131
PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)	0.0655	8.71e-05
PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing)	0.0645	7.29e-05
PWY-6387: UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing)	0.0644	8.19e-05
PWY-5695: urate biosynthesis/inosine 5'-phosphate degradation	0.0637	0.000163
PWY-5686: UMP biosynthesis	0.063	0.000182
PWY-6700: queuosine biosynthesis	0.0593	0.000171
PWY-6385: peptidoglycan biosynthesis III (mycobacteria)	0.0592	6.8e-05
NONMEVIPP-PWY: methylerythritol phosphate pathway I	0.0486	0.000123
PWY-7228: superpathway of guanosine nucleotides de novo biosynthesis I	0.0456	5.53e-05
PWY0-1296: purine ribonucleosides degradation	0.0439	0.000235
PWY-6125: superpathway of guanosine nucleotides de novo biosynthesis II	0.0412	5.68e-05
PWY-6163: chorismate biosynthesis from 3-dehydroquinate	0.0391	0.000108
PWY-7220: adenosine deoxyribonucleotides de novo biosynthesis II	0.032	3.47e-05
PWY-7222: guanosine deoxyribonucleotides de novo biosynthesis II	0.032	3.47e-05
PWY-7560: methylerythritol phosphate pathway II	0.0214	9.53e-05
PWY-6270: isoprene biosynthesis I	0.0214	7.44e-05
PWY-6609: adenine and adenosine salvage III	0.0185	0.000198
GLUTORN-PWY: L-ornithine biosynthesis	0.0153	5.08e-05
PWY-5897: superpathway of menaquinol-11 biosynthesis	0.0113	3.43e-05

The table is too large to include the full table in this document. A partial table is shown which includes only 20 rows. Please see the data file for the full table: [top_average_pathways_names.tsv](#)

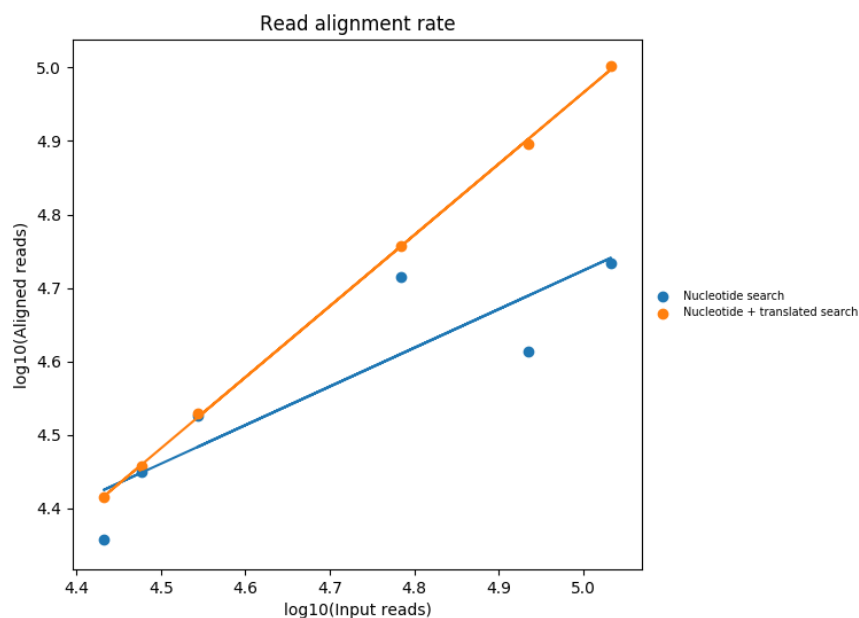
Detailed functions of the top 3 pathways can be found on the following MetaCyc pages:

- [PWY-7221: guanosine ribonucleotides de novo biosynthesis](#)
- [PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II \(lysine-containing\)](#)
- [PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I \(meso-diaminopimelate containing\)](#)

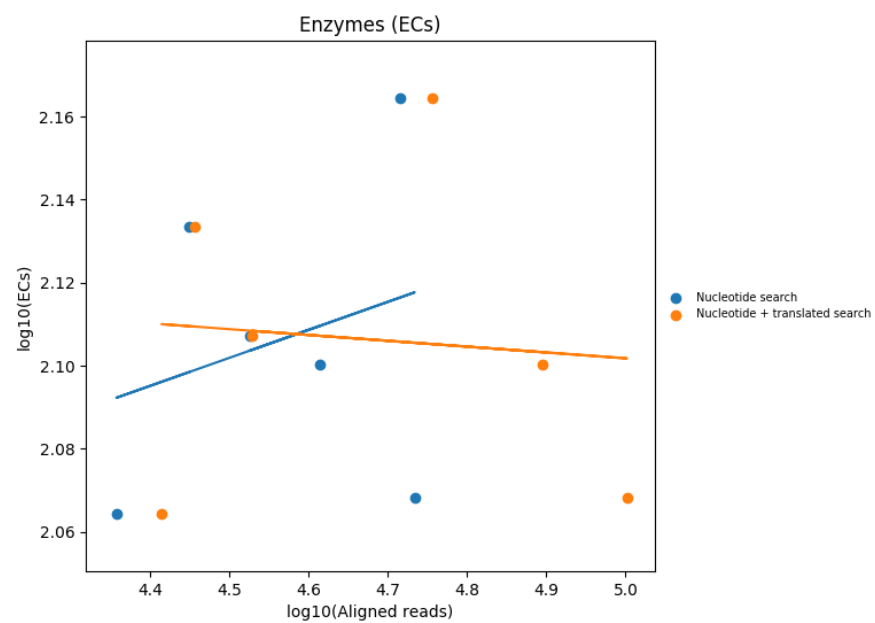
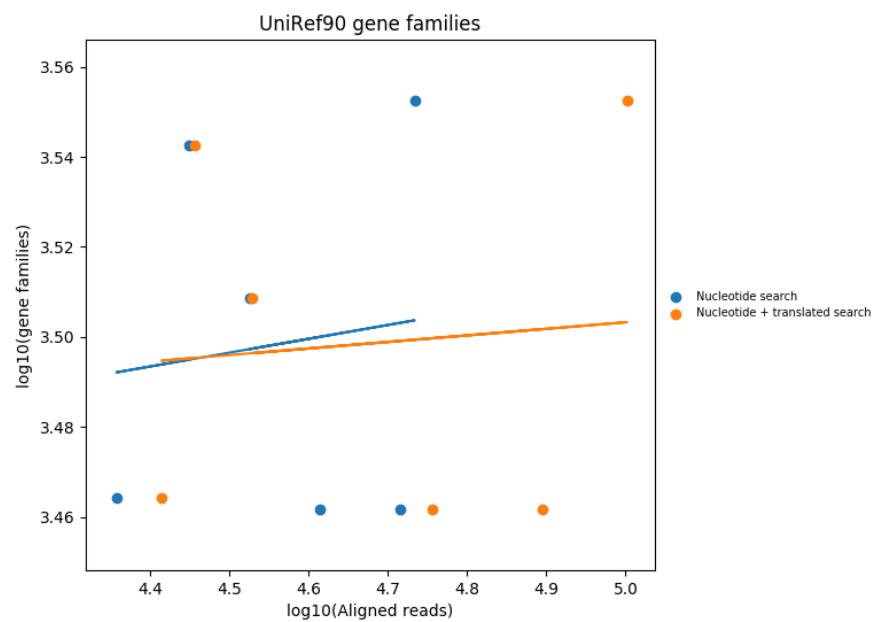
To learn more about other pathways, search for the pathway by name on the [MetaCyc](#) website.

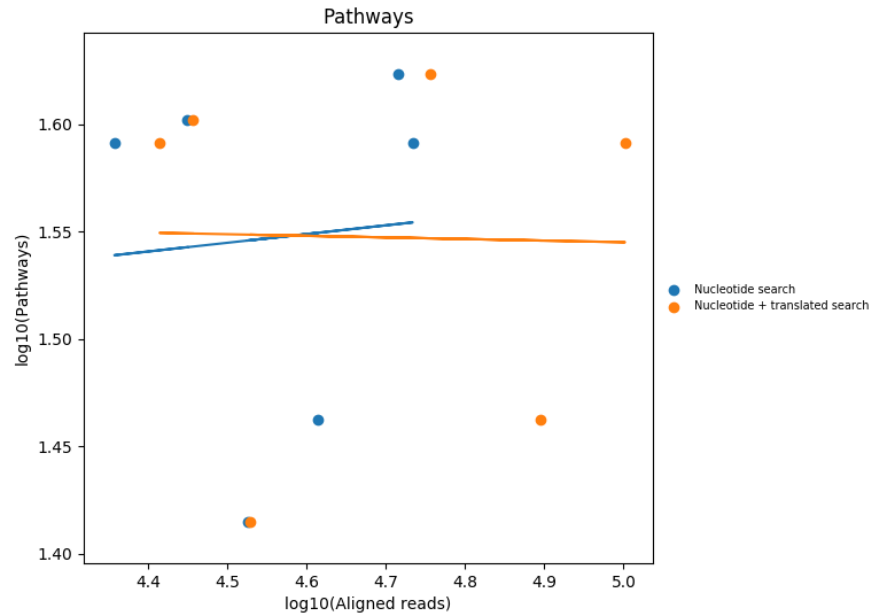
Features

Feature detection as a function of sequencing depth. Effect of sample sequencing depth on the ability to detect microbiome functional features in metagenomic sequence data. HUMAnN2 functional profiling of DNA quality filtered reads was performed on individual samples in species-specific mode (blue), i.e. nucleotide alignment against pangenomes of species identified in the sample with MetaPhlAn2, and in combined species-specific and -agnostic (orange) mode, in which reads not matching any pangenome reference sequences were subjected to translated searching against the UniRef90 database. Each profiled sample is represented by a orange and blue point in each plot. Linear regression fit is represented by straight lines in each plot.



Number of aligned reads in species-specific (nucleotide search) and species-agnostic (translated search) HUMAnN2 mode as a function of input reads.





Detection of UniRef90 gene families, enzyme modules, and pathways as a function of aligned reads.

Data Processing Workflow Information

Software Versions

- kneaddata v0.6.1
- MetaPhlAn version 2.6.0 (19 August 2016)
- humann2 v0.11.1

Tasks Run

- kneaddata -input LD96R2_subsample.fastq.gz -output main -threads 1
-output-prefix LD96R2_subsample -reference-db Homo_sapiens_demo
-reference-db SILVA_demo -serial

- metaphlan2.py LD96R2_subsample.fastq -input_type fastq -output_file LD96R2_subsample_taxonomic_profile.tsv -samout LD96R2_subsample_bowtie2.sam -nproc 1 -no_map -tmp_dir main
- sample2markers.py -ifn_samples LD96R2_subsample_bowtie2.sam -input_type sam -output_dir strainphlan -nprocs 1
- humann2 -input LD96R2_subsample.fastq -output main -o-log LD96R2_subsample.log -threads 1 -taxonomic-profile LD96R2_subsample_taxonomic_profile.tsv
- humann2_regroup_table -input LD96R2_subsample_genefamilies.tsv -output LD96R2_subsample_ecs.tsv -groups uniref90_level4ec
- humann2_renorm_table -input LD96R2_subsample_ecs.tsv -output LD96R2_subsample_ecs_relab.tsv -units relab -special n
- humann2_renorm_table -input LD96R2_subsample_genefamilies.tsv -output LD96R2_subsample_genefamilies_relab.tsv -units relab -special n
- humann2_renorm_table -input LD96R2_subsample_pathabundance.tsv -output LD96R2_subsample_pathabundance_relab.tsv -units relab -special n
- kneaddata -input HD42R4_subsample.fastq.gz -output main -threads 1 -output-prefix HD42R4_subsample -reference-db Homo_sapiens_demo -reference-db SILVA_demo -serial
- metaphlan2.py HD42R4_subsample.fastq -input_type fastq -output_file HD42R4_subsample_taxonomic_profile.tsv -samout HD42R4_subsample_bowtie2.sam -nproc 1 -no_map -tmp_dir main
- sample2markers.py -ifn_samples HD42R4_subsample_bowtie2.sam -input_type sam -output_dir strainphlan -nprocs 1
- humann2 -input HD42R4_subsample.fastq -output main -o-log HD42R4_subsample.log -threads 1 -taxonomic-profile HD42R4_subsample_taxonomic_profile.tsv
- humann2_regroup_table -input HD42R4_subsample_genefamilies.tsv -output HD42R4_subsample_ecs.tsv -groups uniref90_level4ec
- humann2_renorm_table -input HD42R4_subsample_ecs.tsv -output HD42R4_subsample_ecs_relab.tsv -units relab -special n
- humann2_renorm_table -input HD42R4_subsample_genefamilies.tsv -output HD42R4_subsample_genefamilies_relab.tsv -units relab -special n

- `humann2_renorm_table -input HD42R4_subsample_pathabundance.tsv -output HD42R4_subsample_pathabundance_relab.tsv -units relab -special n`
- `kneaddata -input HD48R4_subsample.fastq.gz -output main -threads 1 -output-prefix HD48R4_subsample -reference-db Homo_sapiens_demo -reference-db SILVA_demo -serial`
- `metaphlan2.py HD48R4_subsample.fastq -input_type fastq -output_file HD48R4_subsample_taxonomic_profile.tsv -samout HD48R4_subsample_bowtie2.sam -nproc 1 -no_map -tmp_dir main`
- `sample2markers.py -ifn_samples HD48R4_subsample_bowtie2.sam -input_type sam -output_dir strainphlan -nprocs 1`
- `humann2 -input HD48R4_subsample.fastq -output main -o-log HD48R4_subsample.log -threads 1 -taxonomic-profile HD48R4_subsample_taxonomic_profile.tsv`
- `humann2_regroup_table -input HD48R4_subsample_genefamilies.tsv -output HD48R4_subsample_ecs.tsv -groups uniref90_level4ec`
- `humann2_renorm_table -input HD48R4_subsample_ecs.tsv -output HD48R4_subsample_ecs_relab.tsv -units relab -special n`
- `humann2_renorm_table -input HD48R4_subsample_genefamilies.tsv -output HD48R4_subsample_genefamilies_relab.tsv -units relab -special n`
- `humann2_renorm_table -input HD48R4_subsample_pathabundance.tsv -output HD48R4_subsample_pathabundance_relab.tsv -units relab -special n`
- `kneaddata -input LV16R4_subsample.fastq.gz -output main -threads 1 -output-prefix LV16R4_subsample -reference-db Homo_sapiens_demo -reference-db SILVA_demo -serial`
- `metaphlan2.py LV16R4_subsample.fastq -input_type fastq -output_file LV16R4_subsample_taxonomic_profile.tsv -samout LV16R4_subsample_bowtie2.sam -nproc 1 -no_map -tmp_dir main`
- `sample2markers.py -ifn_samples LV16R4_subsample_bowtie2.sam -input_type sam -output_dir strainphlan -nprocs 1`
- `humann2 -input LV16R4_subsample.fastq -output main -o-log LV16R4_subsample.log -threads 1 -taxonomic-profile LV16R4_subsample_taxonomic_profile.tsv`
- `humann2_regroup_table -input LV16R4_subsample_genefamilies.tsv`

- output LV16R4_subsample_ecs.tsv -groups uniref90_level4ec
- humann2_renorm_table -input LV16R4_subsample_ecs.tsv -output LV16R4_subsample_ecs_relab.tsv -units relab -special n
- humann2_renorm_table -input LV16R4_subsample_genefamilies.tsv -output LV16R4_subsample_genefamilies_relab.tsv -units relab -special n
- humann2_renorm_table -input LV16R4_subsample_pathabundance.tsv -output LV16R4_subsample_pathabundance_relab.tsv -units relab -special n
- kneaddata -input HD32R1_subsample.fastq.gz -output main -threads 1 -output-prefix HD32R1_subsample -reference-db Homo_sapiens_demo -reference-db SILVA_demo -serial
- metaphlan2.py HD32R1_subsample.fastq -input_type fastq -output_file HD32R1_subsample_taxonomic_profile.tsv -samout HD32R1_subsample_bowtie2.sam -nproc 1 -no_map -tmp_dir main
- sample2markers.py -ifn_samples HD32R1_subsample_bowtie2.sam -input_type sam -output_dir strainphlan -nprocs 1
- humann2 -input HD32R1_subsample.fastq -output main -o-log HD32R1_subsample.log -threads 1 -taxonomic-profile HD32R1_subsample_taxonomic_profile.tsv
- humann2_regroup_table -input HD32R1_subsample_genefamilies.tsv -output HD32R1_subsample_ecs.tsv -groups uniref90_level4ec
- humann2_renorm_table -input HD32R1_subsample_ecs.tsv -output HD32R1_subsample_ecs_relab.tsv -units relab -special n
- humann2_renorm_table -input HD32R1_subsample_genefamilies.tsv -output HD32R1_subsample_genefamilies_relab.tsv -units relab -special n
- humann2_renorm_table -input HD32R1_subsample_pathabundance.tsv -output HD32R1_subsample_pathabundance_relab.tsv -units relab -special n
- kneaddata -input LV20R4_subsample.fastq.gz -output main -threads 1 -output-prefix LV20R4_subsample -reference-db Homo_sapiens_demo -reference-db SILVA_demo -serial
- kneaddata_read_count_table -input main -output kneaddata_read_count_table.tsv
- metaphlan2.py LV20R4_subsample.fastq -input_type fastq -output_file

```
LV20R4_subsample_taxonomic_profile.tsv -samout LV20R4_subsample_bowtie2.sam
-nproc 1 -no_map -tmp_dir main
```

- sample2markers.py -ifn_samples LV20R4_subsample_bowtie2.sam
-input_type sam -output_dir strainphlan -nprocs 1
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -
print_clades_only > clades_list.txt
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers
all_markers.fasta -clade g__Faecalibacterium -ofn_markers g__Faecalibacterium.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -
clades g__Faecalibacterium -nprocs_main 1 -keep_alignment_files
-marker_in_clade 0.01 -ifn_markers g__Faecalibacterium.markers.fasta
> 9_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers
all_markers.fasta -clade s__Bacteroides_caccae -ofn_markers
s__Bacteroides_caccae.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -
clades s__Bacteroides_caccae -nprocs_main 1 -keep_alignment_files
-marker_in_clade 0.01 -ifn_markers s__Bacteroides_caccae.markers.fasta
> 0_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers
all_markers.fasta -clade s__Bacteroides_cellulosilyticus -ofn_markers
s__Bacteroides_cellulosilyticus.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -clades
s__Bacteroides_cellulosilyticus -nprocs_main 1 -keep_alignment_files -
marker_in_clade 0.01 -ifn_markers s__Bacteroides_cellulosilyticus.markers.fasta
> 1_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers
all_markers.fasta -clade s__Bacteroides_dorei -ofn_markers
s__Bacteroides_dorei.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -
clades s__Bacteroides_dorei -nprocs_main 1 -keep_alignment_files
-marker_in_clade 0.01 -ifn_markers s__Bacteroides_dorei.markers.fasta
> 2_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers
all_markers.fasta -clade s__Bacteroides_finegoldii -ofn_markers

s__Bacteroides_finegoldii.markers.fasta

- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -clades s__Bacteroides_finegoldii -nprocs_main 1 -keep_alignment_files -marker_in_clade 0.01 -ifn_markers s__Bacteroides_finegoldii.markers.fasta > 3_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers all_markers.fasta -clade s__Bacteroides_fragilis -ofn_markers s__Bacteroides_fragilis.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -clades s__Bacteroides_fragilis -nprocs_main 1 -keep_alignment_files -marker_in_clade 0.01 -ifn_markers s__Bacteroides_fragilis.markers.fasta > 4_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers all_markers.fasta -clade s__Bacteroides_ovatus -ofn_markers s__Bacteroides_ovatus.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -clades s__Bacteroides_ovatus -nprocs_main 1 -keep_alignment_files -marker_in_clade 0.01 -ifn_markers s__Bacteroides_ovatus.markers.fasta > 5_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers all_markers.fasta -clade s__Bacteroides_thetaiotaomicron -ofn_markers s__Bacteroides_thetaiotaomicron.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -clades s__Bacteroides_thetaiotaomicron -nprocs_main 1 -keep_alignment_files -marker_in_clade 0.01 -ifn_markers s__Bacteroides_thetaiotaomicron.markers.fasta > 6_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers all_markers.fasta -clade s__Bacteroides_vulgatus -ofn_markers s__Bacteroides_vulgatus.markers.fasta
- strainphlan.py -ifn_samples *.markers -output_dir strainphlan -clades s__Bacteroides_vulgatus -nprocs_main 1 -keep_alignment_files -marker_in_clade 0.01 -ifn_markers s__Bacteroides_vulgatus.markers.fasta > 7_clade.log
- extract_markers.py -mpa_pkl mpa_v20_m200.pkl -ifn_markers all_markers.fasta -clade s__Bifidobacterium_longum -ofn_markers s__Bifidobacterium_longum.markers.fasta

- `strainphlan.py -ifn_samples *.markers -output_dir strainphlan -clades s__Bifidobacterium_longum -nprocs_main 1 -keep_alignment_files -marker_in_clade 0.01 -ifn_markers s__Bifidobacterium_longum.markers.fasta > 8_clade.log`
- `humann2_join_tables -input main -output metaphlan2_taxonomic_profiles.tsv -file_name taxonomic_profile`
- `count_features.py -input metaphlan2_taxonomic_profiles.tsv -output metaphlan2_species_counts_table.tsv -include s__ -filter t__ -reduce-sample-name`
- `humann2 -input LV20R4_subsample.fastq -output main -o-log LV20R4_subsample.log -threads 1 -taxonomic-profile LV20R4_subsample_taxonomic_profile.tsv`
- `humann2_join_tables -input main -output genefamilies.tsv -file_name genefamilies`
- `humann2_join_tables -input main -output pathabundance.tsv -file_name pathabundance`
- `humann2_renorm_table -input LV20R4_subsample_genefamilies.tsv -output LV20R4_subsample_genefamilies_relab.tsv -units relab -special n`
- `humann2_join_tables -input genes -output genefamilies_relab.tsv`
- `humann2_renorm_table -input LV20R4_subsample_pathabundance.tsv -output LV20R4_subsample_pathabundance_relab.tsv -units relab -special n`
- `humann2_join_tables -input pathways -output pathabundance_relab.tsv`
- `get_counts_from_humann2_logs.py -input main -output humann2_read_and_species_count_table.tsv`
- `humann2_regroup_table -input LV20R4_subsample_genefamilies.tsv -output LV20R4_subsample_ecs.tsv -groups uniref90_level4ec`
- `humann2_renorm_table -input LV20R4_subsample_ecs.tsv -output LV20R4_subsample_ecs_relab.tsv -units relab -special n`
- `humann2_join_tables -input ecs -output ecs_relab.tsv`
- `humann2_join_tables -input regrouped -output ecs.tsv -file_name ecs`

- `count_features.py -input genefamilies_relab.tsv -output humann2_genefamilies_relab_counts.tsv -reduce-sample-name -ignore-un-features -ignore-stratification`
- `count_features.py -input pathabundance_relab.tsv -output humann2_pathabundance_relab_counts.tsv -reduce-sample-name -ignore-un-features -ignore-stratification`
- `count_features.py -input ecs_relab.tsv -output humann2_ecs_relab_counts.tsv -reduce-sample-name -ignore-un-features -ignore-stratification`
- `humann2_join_tables -input counts -output humann2_feature_counts.tsv -file_name _relab_counts.tsv`