Metagenomics analysis pipeline for the study "Mapping human microbiome drug metabolism by gut bacteria and their genes"

This project is part of the work by Michael Zimmermann, Maria Zimmermann-Kogadeeva, Rebekka Wegmann and Andrew L. Goodman.

In this workflow, metagenomics samples of 28 healthy human gut communities are analyzed with BioBakery tools and QIIME to obtain a list of OTUs, alpha-diversity scores, and quantify a subset of proteins of interest.

Prerequisites: Software

- BioBakery KneadData: a tool designed to perform quality control on metagenomic and metatranscriptomic sequencing data, especially data from microbiome experiments. https://bitbucket.org/biobakery/kneaddata
- BioBakery MetaPhlan2: a computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species-level. https://bitbucket.org/biobakery/metaphlan2
- BioBakery ShortBRED: a pipeline to take a set of protein sequences, reduce them to a set of unique identifying strings ("markers"), and then search for these markers in metagenomic data and determine the presence and abundance of the protein families of interest. https://bitbucket.org/biobakery/shortbred
- Biom convert: The convert command in the biom-format project can be used to convert between biom and tab-delimited table formats (to convert OTU table obtained from MetaPhlan2 to biom format). http://biom-format.org/documentation/biom_conversion.html
- The QIIME alpha_diversity script: this script calculates alpha diversity, or within-sample diversity, using an OTU table (in biom format). http://qiime.org/scripts/alpha_diversity.html

Prerequisites: Data

- Fastq files of the 28 communities available online at ENA repository (Accession number PRJEB31790)
- Fasta file of proteins sequences of interest (example provided in infile_drugmet_proteins.fasta)
- Human genome as a reference for sample preprocessing (remove human reads) Can be downloaded through KneadData BioBakery tool:

```
> kneaddata_database --download human_genome bowtie2 $DIR
```

- UniRef90 reference database to create protein markers for identification of the specified proteins in the samples (download from https://www.uniprot.org/downloads)

KneadData workflow

1. Preprocess all samples with KneadData to remove sequencing barcodes and contaminating human sequences.

Create an array of identifiers to preprocess all 28 samples in a loop using ARRAY elements as part of the name.

```
> ARRAY=(01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28)
> for i in {1..28}
do
kneaddata --input MV${ARRAY[$i]}*R1* --input MV${ARRAY[$i]}*R2* --
trimmomatic Trimmomatic-0.38 --trimmomatic-
options="ILLUMINACLIP:TruSeq3-PE.fa:2:30:10" --trimmomatic-
options="LEADING:3" --trimmomatic-options="TRAILING:3" --trimmomatic-
options="SLIDINGWINDOW:4:20" --trimmomatic-options="MINLEN:36" -db
human -t 16 --output communityHumanKnead
done
```

2. Concatenate paired-end files for further analysis (as recommended in BioBakery tutorial).

```
> for i in {1..28}
do
cat communityHumanKnead/MV${ARRAY[$i]}*paired_1.fastq
communityHumanKnead/MV${ARRAY[$i]}d*_paired_2.fastq >
commmunityHumanKnead/MV${ARRAY[$i]}_knead_paired_cat.fastq
done
```

MetaPhlan2 workflow

3. Analyze species abundances in the samples with MetaPhlan2.
   –nproc 16 parameter can be used to parallelize the workflow in case multiple processors are available.

```
> ARRAY=(01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21
22 23 24 25 26 27 28)
> for i in {1..28}
do
metaphlan2.py MV${ARRAY[$i]}*paired_1*,MV${ARRAY[$i]}*paired_2* --
bowtie2out MV${ARRAY[$i]}_knead_paired_dna.fastq.bowtie2.bz2 --nproc 16
--input_type fastq >
profiled_metagenome_knead_paired_MV${ARRAY[$i]}.txt
done
```

4. Merge MetaPhlan2 tables into one OTU summary table.

```
> merge_metaphlan_tables.py profiled*.txt >
profiled_metagenomes_knead_paired_combined.txt
```

ShortBRED workflow

5. With ShortBRED tools, create a list of sequence identifiers for the proteins of interest using UniRef90 as a reference.
   –threads 16 parameter can be used to parallelize the workflow in case multiple processors are available.

```
> shortbred_identify.py --goi  infile_drugmet_proteins.fasta --ref
uniref90.fasta --markers mydrugproteinmarkers.faa --tmp
drugproteins_identify --usearch usearchpath --threads 16
```

6. For each preprocessed fastq file, calculate abundances of the proteins of interest with shortbred_quantify.py.

```
> for i in {1..28}
do
shortbred_quantify.py --markers mydrugproteinmarkers.faa --wgs
commmunityHumanKnead/MV${ARRAY[$i]}_knead_paired_cat.fastq  --results
biobakery_shortbred/MV${ARRAY[$i]}_shortbred_drugprotein_results.txt --
tmp biobakery_shortbred/communityMV${ARRAY[$i]}_temp --threads 16 --
usearch usearchpath
done
```

QIIME workflow

7. Perform alpha-diversity analysis of the OTU table with QIIME utiliy alpha_diversity.
   a. Convert MetaPhlan2 output OTU tables to biom format

```
> for f in profiled*.txt
do
newname="${f%.txt}.biom"
biom convert -i $f -o $newname --table-type="OTU table" --to-json
done
```

   b. Calculate alpha-diversity on the OTU summary table.

```
> alpha_diversity.py -i profiled_metagenomes_knead_paired_combined.biom
-o qiime_alpha_chao1_ci.txt -m shannon
```