**Bioinformatic workflows of NAFLD study**

**1. Quality control for raw data**

Quality control by fastQC (0.11.8) and Trimmomatic (v 0.32)

conda create --name myenv FastQC=0.11.8 Trimmomatic=0.32

fastqc sampleID.fastq -o quality/

trimmomatic PE -phred33 -trimlog log.txt $INPUT\_PATH/SampleID\_1.fastq.gz $INPUT\_PATH/SampleID\_2.fastq.gz $OUTPUT\_PATH/paired\_SampleID\_1.fq.gz $OUTPUT\_PATH/paired\_SampleID\_2.fq.gz $OUTPUT\_PATH/paired\_SampleID\_1.fq.gz $OUTPUT\_PATH/unpaired\_SampleID\_2.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50

**2. Host genomic contaminant removal**

conda install --name myenv -c bioconda bowtie2

wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/635/GCF\_000001635.27\_GRCm39/GCF\_000001635.27\_GRCm39\_genomic.fna.gz -O genomic.fa

bowtie2-build genomic.fa genome\_index

bowtie2 -x genome\_index -1 $INPUT\_PATH/trimmed\_SampleID\_1.fq.gz -2 $INPUT\_PATH/trimmed\_SampleID\_2.fq.gz --un-conc-gz $OUTPUT\_PATH/SampleID.hostfree.fq.gz -p 60 -S $OUTPUT\_PATH/SampleID.sam 1>$OUTPUT\_PATH/SampleID.o 2>$OUTPUT\_PATH/SampleID.e

**3. Paired-end reads assembly**

conda install --name myenv -c bioconda spades

spades.py --meta -t 20 -m 500 -1 $INPUT\_PATH/SampleID.hostfree\_1.fq.gz -2 $INPUT\_PATH/SampleID.hostfree\_2.fq.gz -o $OUTPUT\_PATH/SampleID\_assembly 1>$OUTPUT\_PATH/SampleID.o 2>$OUTPUT\_PATH/SampleID.e

**4. Open reading frames (ORFs) prediction and non-redundant gene catalog construction**

**For ORFs prediction, runing on linux sever**

Download MetaGeneMark from <http://topaz.gatech.edu/genemark/license_download.cgi>

tar –xvzf MetaGeneMark\_linux\_64.tar.gz –C $PATH/MetaGeneMark\_linux\_64

$PATH/MetaGeneMark\_linux\_64/mgm/gmhmmp -a -d -f 3 -m $PATH/MetaGeneMark\_linux\_64/mgm/MetaGeneMark\_v1.mod $INPUT\_PATH/contigs.fa -A $OUTPUT\_PATH/protein.fasta -D $OUTPUT\_PATH/nuleotide.fasta

**Constrcuct non-redundant gene catalog**

Download current CD-HIT at https://github.com/weizhongli/cdhit/releases, for example cd-hit-v4.8.1-2019-0228.tar.gz

tar –xvzf cd-hit-v4.8.1-2019-0228.tar.gz –C $PATH/cd-hit-v4.8.1

$PATH/cd-hit-v4.8.1/cd-hit -i $INPUT\_PATH/protein.fasta -o nr95\_protein.fa -n 5 -g 1 -c 0.95 -G 0 -M 0 -d 0 -aS 0.9

**5. Gene abundance calculation**

conda install --name myenv -c bioconda salmon

salmon index -p 20 -t $INPUT\_PATH/nr95\_gene.fa -i $OUTPUT\_PATH/nr95\_gene\_index

salmon quant -i /data2/yjk/Caojia\_ref\_unmapped\_reads/nr95\_salmon/nr95\_gene\_index --libType IU -1 $INPUT\_PATH/SampleID.hostfree\_1.fastq.gz -2 $INPUT\_PATH/SampleID.hostfree\_2.fastq.gz -o $OUTPUT\_PATH/SampleID.quant --meta -p 30

**6. Taxonomic and functional analysis**

**Taxonmic analysis**

conda install --name myenv -c conda-forge -c bioconda metaphlan

metaphlan2.py SampleID.fasta --input\_type fasta > SampleID\_profile.txt

**Functional annotation**

conda install --name myenv -c bioconda eggnog-mapper

emapper.py -i nr95\_protein.fa -o eggNOG\_annotation --cpu 0 --evalue 0.00001