Efficient recovery of complete gut viral genomes by combined short- and long-read sequencing V.1

chuqingsun[1]

[1]Huazhong University of Science and Techonology

chuqingsun

VERSION 1

AUG 14, 2023

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.36wgq339klk5/v1

Protocol Citation: chuqingsun 2023. Efficient recovery of complete gut viral genomes by combined short- and long-read sequencing. **protocols.io** https://dx.doi.org/10.17504/protocols.io.36wgq339klk5/v1

Protocol status: Other

We attempted this protocol but could not get it to work in our workspace

ABSTRACT

Current metagenome-assembled human gut phage catalogs contained mostly fragmented genomes. Here, we developed a vigorous gut virome detection procedure involving viral-like particle enrichment from increased amount of feces (~500g) and combined sequencing of short- and long-reads. Applied to 135 fecal samples, we assembled a Chinese Gut Virome Catalog (CHGV) consisting of 21,646 non-redundant phage genomes that were significantly longer than those obtained by short-read sequencing and contained ~35% complete ones, which was ~nine times more than those in the Gut Virome Database (~4%). Interestingly, majority (~60%) of the CHGV genomes were obtained by either long-read or hybrid assemblies, which overlapped little with those assembled from only the short-reads, indicating the necessity of combined sequencing in gut virome discovery. With this dataset, we elucidated the vast diversity of the gut virome from several aspects, including the identification of 32% novel genomes as compared with public gut virome databases, dozens of phages that were more prevalent than the crAssphages and/or Gubaphages, and several viral clades that are more diverse than the two. In the end, we also characterized the functional capacities of the CHGV encoded proteins and constructed a viral-host interaction network to facilitate future research and applications of the gut viruses.

Created: Aug 14, 2023

Last Modified: Aug 14, 2023

PROTOCOL integer ID: 86442

Assembly

1 *NGS Assembly*

```
#!/bin/bash
#SBATCH --cpus-per-task=16
#SBATCH -o slurm.%N.%j.out        # STDOUT
#SBATCH -e slurm.%N.%j.err        # STDERR

infile=$1
NGS_PATH=$2
export
LD_LIBRARY_PATH=$LD_LIBRARY_PATH:~/local/lib:/mnt/raid6/sunchuqing/S
oftwares/MCR/v94/runtime/glnxa64:/mnt/raid6/sunchuqing/Softwares/MCR
/v94/sys/os/glnxa64:/mnt/raid6/sunchuqing/Softwares/MCR/v94/extern/g
lnxa64：/mnt/raid3/wchen/miniconda2/pkgs/libgcc-7.2.0-h69d50b8_2/lib
cd NGS
mkdir -p 02_trimmed 03_bac_cpn60 03_human_hg38 04_Stats
mkdir -p 05_Removed 06_Assembly 07_CD-HIT

TRIMMO_JAR_FILE='/mnt/raid1/tools/ngs_tools/Trimmomatic-
0.38/trimmomatic-0.38.jar'
TRIMMO_ADAPTOR_FILE_PE='/mnt/raid1/tools/ngs_tools/Trimmomatic-
0.38/adapters/TruSeq3-PE.fa'
R1=`ls ${NGS_PATH}/*${infile}*R1*|head -n 1 `
R2=`ls ${NGS_PATH}/*${infile}*R2*|head -n 1 `

if [ ! -s 02_trimmed/${infile}_clean.1.fq ];then
    java -jar $TRIMMO_JAR_FILE PE -threads 16 ${R1} ${R2}
02_trimmed/${infile}_clean.1.fq
02_trimmed/${infile}_clean_unpaired.1.fq
02_trimmed/${infile}_clean.2.fq
02_trimmed/${infile}_clean_unpaired.2.fq
ILLUMINACLIP:$TRIMMO_ADAPTOR_FILE_PE:2:15:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:15:30 MINLEN:50
fi
#rm human
export
PATH=/home/sunchuqing/bin:/mnt/raid6/sunchuqing/Softwares/miniconda3
/condabin:/mnt/raid6/sunchuqing/Softwares/miniconda3/bin:/mnt/raid8/
```

```
sunchuqing/Softwares/bin:/mnt/raid1/puzi/software/metaphlan2:/mnt/ra
id1/puzi/software/metaphlan2/utils:/usr/bin:/usr/local/ncbi/sra-
tools/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/b
in:/usr/games:/usr/local/games:/snap/bin:/mnt/raid6/sunchuqing/Softw
ares/miniconda3/bin:/mnt/raid1/sunchuqing/bc/bc/h/parallel-
meta/bin:/mnt/raid6/sunchuqing/Softwares/miniconda3/bin://mnt/raid1/
data/Software/prokka/bin:/mnt/raid6/sunchuqing/Softwares/miniconda3/
bin:/mnt/raid7/sunchuqing/Softwares/bin

bowtie2 -p 16 --un-conc 05_Removed/${infile}_%.fastq --no-unal -k 20
-x /mnt/raid5/sunchuqing/Human_Gut_Phage/ref/hg38_ref -1
02_trimmed/${infile}_clean.1.fq -2 02_trimmed/${infile}_clean.2.fq
>log
bowtie2 -p 16 --al-conc 05_Removed/${infile}_%_prophage.fastq --end-
to-end -x /mnt/raid7/sunchuqing/Human_Gut_Phage/db/HGYG_prophage -1
05_Removed/${infile}_1.fastq -2 05_Removed/${infile}_2.fastq -S
../04_Abundance/${infile}_NGS_HGYG_prophage.sam   >log
bowtie2 -p 16 --un-conc 05_Removed/${infile}_%_phage.fastq --end-to-
end -x /mnt/raid7/sunchuqing/Human_Gut_Phage/db/HGYG -1
05_Removed/${infile}_1.fastq -2 05_Removed/${infile}_2.fastq -S
../04_Abundance/${infile}_NGS_HGYG.sam  >log
cat 05_Removed/${infile}_1_prophage.fastq
05_Removed/${infile}_1_phage.fastq >
05_Removed/${infile}_virome_bft_1.fastq
cat 05_Removed/${infile}_2_prophage.fastq
05_Removed/${infile}_2_phage.fastq >
05_Removed/${infile}_virome_bft_2.fastq

R1=`ls 05_Removed/${infile}_virome_bft_1.fastq`
R2=`ls 05_Removed/${infile}_virome_bft_2.fastq`
java -jar $TRIMMO_JAR_FILE PE -threads 16 ${R1} ${R2}
05_Removed/${infile}_virome_1.fastq
02_trimmed/${infile}_clean_unpaired.vir.1.fq
05_Removed/${infile}_virome_2.fastq
02_trimmed/${infile}_clean_unpaired.vir.2.fq
ILLUMINACLIP:$TRIMMO_ADAPTOR_FILE_PE:2:15:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:15:30 MINLEN:50

export
PYTHONPATH=/mnt/raid6/sunchuqing/Softwares/miniconda3/lib/python3.7/
site-packages:/mnt/raid8/sunchuqing/Softwares/lib/site-packages
# casper 02_trimmed/${infile}_clean.1.fq
 02_trimmed/${infile}_clean.2.fq  -o
02_trimmed/${infile}_clean.merged -t 16
pandaseq -f 02_trimmed/${infile}_clean.1.fq -r
02_trimmed/${infile}_clean.2.fq -F -w
02_trimmed/${infile}_clean.merged.fastq -T 16
export PATH=$PATH:/mnt/raid6/sunchuqing/Softwares/miniconda3/bin
```

```
#16s
/mnt/raid6/sunchuqing/Softwares/ViromeQC/viromeqc/viromeQC.py  \
    -i 02_trimmed/${infile}_clean.merged.fastq \
    -o 04_Stats/${infile}.viromeqc \
    --bowtie2_threads 16\
    --diamond_threads 16
#bac=`tail -n 1 04_Stats/${infile}.viromeqc | awk 'BEGIN {FS="\t"}
{print $4}'`
bacpct=`tail -n 1 04_Stats/${infile}.viromeqc | awk 'BEGIN {FS="\t"}
{print $4}'`
# /mnt/raid5/sunchuqing/Softwares/SortMeRNA/sortmerna-2.1/sortmerna
--ref /mnt/raid5/sunchuqing/Softwares/SortMeRNA/sortmerna-
2.1/rRNA_databases/silva-bac-16s-
id90.fasta,/mnt/raid5/sunchuqing/Softwares/SortMeRNA/sortmerna-
2.1/index/silva-bac-16s-db --reads
02_trimmed/${infile}_clean.merged.fastq --aligned 03_bac/${infile} -
-blast 1
#bowtie2 -x /mnt/raid6/sunchuqing/Database/Bacteria_bowtie/bac -1
02_trimmed/${infile}_clean.1.fq -2 02_trimmed/${infile}_clean.2.fq
-S 03_bac/${infile}.sam
#bac=`/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools
flagstat 03_bac/${infile}.sam | grep 'mapped' | awk 'BEGIN {FS=" "}
{print $1}' | head -n 1`


#cpn60
bowtie2 -x /mnt/raid5/sunchuqing/Human_Gut_Phage/ref/cpn60_ref -1
02_trimmed/${infile}_clean.1.fq -2 02_trimmed/${infile}_clean.2.fq
-S 03_bac_cpn60/${infile}.sam
cpn=`/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools
flagstat 03_bac_cpn60/${infile}.sam | grep 'mapped' | awk 'BEGIN
{FS=" "} {print $1}' | head -n 1`
reads=`/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools
flagstat 03_bac_cpn60/${infile}.sam | grep 'total' | awk 'BEGIN
{FS=" "} {print $1}' | head -n 1`

#hg38
bowtie2 -x /mnt/raid5/sunchuqing/Human_Gut_Phage/ref/hg38_ref -1
02_trimmed/${infile}_clean.1.fq -2 02_trimmed/${infile}_clean.2.fq
-S 03_human_hg38/${infile}.sam
#/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools flagstat
03_human_hg38/${infile}.sam
human=`/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools
flagstat 03_human_hg38/${infile}.sam | grep 'mapped' | awk 'BEGIN
{FS=" "} {print $1}' | head -n 1`

#bacpct=`awk 'BEGIN {printf "%.10f\n",
("'"${bac}"'"*100/"'"$reads"'")}'`
```

```
#bacpct=${bac}"%"
humanpct=`awk 'BEGIN {printf "%.10f\n",
("'"${human}"'"*100/"'"$reads"'")}'`
cpnpct=`awk 'BEGIN {printf "%.10f\n",
("'"${cpn}"'"*100/"'"$reads"'")}'`
echo
"File_name,reads_num,16spct,cpn60_reads,cpn60pct,human_reads,humanpc
t">04_Stats/${infile}.csv
echo
"${infile},${reads},${bacpct}%,${cpn},${cpnpct}%,${human},${humanpct
}%" >>04_Stats/${infile}.csv

#IDBA
/mnt/raid5/sunchuqing/Softwares/idba/bin/fq2fa --merge --filter
05_Removed/${infile}_virome_1.fastq
05_Removed/${infile}_virome_2.fastq 05_Removed/${infile}.fa
/mnt/raid5/sunchuqing/Softwares/idba/bin/idba_ud -r
05_Removed/${infile}.fa --maxk 120 --step 10 -o
06_Assembly/${infile} --num_threads 16 --min_contig 1000

cat 06_Assembly/${infile}/contig.fa >06_Assembly/${infile}.fasta
/mnt/raid6/sunchuqing/Softwares/cdhit-master/cd-hit-est -i
06_Assembly/${infile}.fasta -o 07_CD-HIT/${infile}.fa -c 0.95 -n 5
-T 15 -M 16000 >07_CD-HIT/${infile}.log

cd ..
```

2 ***# TGS assembly***

```
#!/bin/bash
#SBATCH --cpus-per-task=16
#SBATCH -o slurm.%N.%j.out        # STDOUT
#SBATCH -e slurm.%N.%j.err        # STDERR

infile=$1
NGS_PATH=$2
G3_PATH=$3
Software='/mnt/raid6/sunchuqing/Softwares'
export
LD_LIBRARY_PATH=$LD_LIBRARY_PATH:~/local/lib:/mnt/raid6/sunchuqing/S
oftwares/MCR/v94/runtime/glnxa64:/mnt/raid6/sunchuqing/Softwares/MCR
/v94/sys/os/glnxa64:/mnt/raid6/sunchuqing/Softwares/MCR/v94/extern/g
lnxa64：/mnt/raid3/wchen/miniconda2/pkgs/libgcc-7.2.0-h69d50b8_2/lib
cd G3
mkdir -p 01_ccs 02_Removed/${infile}
#Run CCS corrction
```

```
CCS=`ls ${G3_PATH}/*${infile}*.subreads.bam `
if [ ! -s 02_Removed/${infile}/${infile}.virome.fastq ];then
  if [ ! -s ${G3_PATH}/CCS/${infile}.subreads.bam ];then
    ${Software}/miniconda3/bin/ccs ${CCS} 01_ccs/${infile}.ccs.fastq
-j 16
  else
    CCS=`ls ${G3_PATH}/CCS/${infile}.subreads.bam`
    bedtools bamtofastq -i ${CCS} -fq 01_ccs/${infile}.ccs.fastq
  fi
  #Remove human genome
  bowtie2 -p 16 --un 02_Removed/${infile}/${infile}.fastq  -x
/mnt/raid5/sunchuqing/Human_Gut_Phage/ref/hg38_ref -U
01_ccs/${infile}.ccs.fastq  > log
  bowtie2 --end-to-end -x
/mnt/raid7/sunchuqing/Human_Gut_Phage/db/HGYG_prophage -U
02_Removed/${infile}/${infile}.fastq -S
../04_Abundance/${infile}_G3_HGYG_prophage.sam -p 16 --al
02_Removed/${infile}/${infile}.prophage.fastq --quiet
  bowtie2 --end-to-end -x
/mnt/raid7/sunchuqing/Human_Gut_Phage/db/HGYG -U
02_Removed/${infile}/${infile}.fastq -S
../04_Abundance/${infile}_G3_HGYG.sam -p 16 --un
02_Removed/${infile}/${infile}.phage.fastq --quiet
  cat 02_Removed/${infile}/${infile}.prophage.fastq
02_Removed/${infile}/${infile}.phage.fastq >
02_Removed/${infile}/${infile}.virome.fastq
  seqtk seq -a 02_Removed/${infile}/${infile}.virome.fastq  |
/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/seqkit rmdup -s -o
 02_Removed/${infile}/${infile}.fa
  rm 02_Removed/${infile}/${infile}.*hage.fastq
02_Removed/${infile}/${infile}.fastq
fi


#Assembly with Canu
rm -rf 03_canu_Assembly/${infile}
mkdir -p 03_canu_Assembly/${infile}
G3=`pwd`
cd 03_canu_Assembly/${infile}
host=`hostname`
if [ "$host" = "bork" ];then
    /mnt/raid7/sunchuqing/Softwares/OPERAMS-bork/canu/build/bin/canu
 \
      -p ${infile} \
      -d ./ genomeSize=20k corOutCoverage=1 \
      -corrected \
      -pacbio ../../02_Removed/${infile}/${infile}.fa  \
      useGrid=false
```

```
else
    /mnt/raid6/sunchuqing/Softwares/canu/Linux-amd64/bin/canu  \
      -p ${infile} \
      -d ./ genomeSize=20k corOutCoverage=1 \
      -corrected \
      -pacbio ../../02_Removed/${infile}/${infile}.fa  \
      useGrid=false
fi

cd ${G3}

#Assembly with Flye
mkdir -p 03_Flye_Assembly/${infile}
mkdir -p 05_Assembly/${infile}
zcat 03_canu_Assembly/$infile/${infile}.trimmedReads.fasta.gz
>03_canu_Assembly/$infile/${infile}.trimmedReads.fasta
flye --pacbio-corr  02_Removed/${infile}/${infile}.fa \
  --meta --genome-size 20k \
  --out-dir 03_Flye_Assembly/${infile}/ \
  --threads 16  --min-overlap 1000
cat 03_canu_Assembly/${infile}/${infile}.contigs.fasta
03_Flye_Assembly/${infile}/assembly.fasta >
05_Assembly/${infile}_fc.fa

#Binning with MetaBAT
if [ -s 03_canu_Assembly/${infile}/${infile}.unitigs.fasta ];then
  mkdir -p 04_MetaBAT_Assembly/${infile}
  cd 04_MetaBAT_Assembly/${infile}
  mkdir -p db
  bowtie2-build
../../03_canu_Assembly/${infile}/${infile}.unitigs.fasta
db/${infile}
  bowtie2 -x db/${infile} -U ../../01_ccs/${infile}.ccs.fastq -S
${infile}.sam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools view -bS
${infile}.sam -o ${infile}.bam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools sort
${infile}.bam > ${infile}.sort.bam
  ${Software}/berkeleylab-
metabat*/bin/jgi_summarize_bam_contig_depths --outputDepth
depth_var.txt ${infile}.sort.bam
  ${Software}/berkeleylab-metabat*/bin/metabat -i
../../03_canu_Assembly/${infile}/${infile}.unitigs.fasta -a
depth_var.txt  -o metabat -v

  for file in ./metabat.*.fa
    do
      num=${file//[!0-9]/}
```

```
      #echo $num
      sed -e "/^>/ s/$/ ${num}/" metabat.$num.fa  >>
metabat_binned.concat.fasta
  done
  grep '>' metabat_binned.concat.fasta | sed 's/>//g' >
metabat_binned.info


  cd ../../05_Assembly/${infile}
  cut -f1 ../../03_canu_Assembly/${infile}/${infile}.unitigs.bed |
sort|uniq -c > contig.list
  while read -r contig
  do
    num=`echo ${contig} | awk 'BEGIN {FS=" "} {print $1}'`
    tig=`echo ${contig} | awk 'BEGIN {FS="ctg"} {print $2}'`
    if [ $num -eq 1 ];then
      awk '/^>/ {printf("\n%s\t",$0);next;} {printf("%s",$0);} END
{printf("\n");}' <
../../03_canu_Assembly/${infile}/${infile}.contigs.fasta |egrep -v
'^$'|tr "\t" "\n"
>../../03_canu_Assembly/${infile}/${infile}.n1.contigs.fasta
      grep "$tig"
../../03_canu_Assembly/${infile}/${infile}.n1.contigs.fasta -A 1|sed
's/>tig/>ctg/g' >> ../${infile}.fasta
      echo $tig >> infa.contig.list
      grep "ctg${tig}"
../../03_canu_Assembly/${infile}/${infile}.unitigs.bed |cut -f4
>>infa.unitig.list
    else
      grep "ctg${tig}"
../../03_canu_Assembly/${infile}/${infile}.unitigs.bed |cut -f4
>unitig.list
      sed 's/utg/tig/g' unitig.list > uni.list
      binnum=`grep -Ff uni.list
../../04_MetaBAT_Assembly/${infile}/metabat_binned.info | awk 'BEGIN
{FS=" "} {print $2}' |sort |uniq |wc -l`
      inbin=`grep -Ff uni.list
../../04_MetaBAT_Assembly/${infile}/metabat_binned.info|wc -l`
      uninum=`cat unitig.list | wc -l `
      #echo "$binnum,$uninum,${inbin}"
      if [[ $binnum == 1 && $uninum == $inbin ]];then

        grep "$tig"
../../03_canu_Assembly/${infile}/${infile}.contigs.fasta -A 100| awk
-v RS='>' 'NR>1{i++}i==1{print ">"$0}' >> ../${infile}.fasta
        echo $tig >> infa.contig.list
        cat unitig.list >> infa.unitig.list
      fi
```

```
    fi

  done <"contig.list"
  cut -f4 ../../03_canu_Assembly/${infile}/${infile}.unitigs.bed
>unitig.list
  awk '/^>/ {printf("\n%s\t",$0);next;} {printf("%s",$0);} END
{printf("\n");}' <
../../03_canu_Assembly/${infile}/${infile}.unitigs.fasta |egrep -v
'^$'|tr "\t" "\n"
>../../03_canu_Assembly/${infile}/${infile}.n1.unitigs.fasta
  for unitig in `grep -v -Ff infa.unitig.list unitig.list `
  do
    tig=`echo ${unitig} | awk 'BEGIN {FS="utg"} {print $2}'`
    #echo $tig
    grep "$tig"
../../03_canu_Assembly/${infile}/${infile}.n1.unitigs.fasta -A 1 |
sed 's/class=contig/class=unitig/g'|sed 's/>tig/>utg/g' >>
../${infile}.fasta
  done
  cat ../../03_Flye_Assembly/${infile}/assembly.fasta
../${infile}.fasta > ../${infile}_fc.fa
  cd ../..
fi
mkdir -p 06_CD-HIT/${infile}
/mnt/raid6/sunchuqing/Softwares/cdhit-master/cd-hit-est -i
05_Assembly/${infile}_fc.fa -o 06_CD-HIT/${infile}/${infile}.fa -c
0.95 -n 5 -T 16 -M 16000


Software='/mnt/raid6/sunchuqing/Softwares'
PATH="/mnt/raid6/sunchuqing/Softwares/miniconda3/bin":$PATH
if [ -s ${NGS_PATH}/*${infile}*R1* ];then
  if [ ! -s ../NGS/02_trimmed/*${infile}*clean.1.fq ];then
    TRIMMO_JAR_FILE='/mnt/raid1/tools/ngs_tools/Trimmomatic-
0.38/trimmomatic-0.38.jar'
    TRIMMO_ADAPTOR_FILE_PE='/mnt/raid1/tools/ngs_tools/Trimmomatic-
0.38/adapters/TruSeq3-PE.fa'
    R1=`ls ${NGS_PATH}/*${infile}*R1*`
    R2=`ls ${NGS_PATH}/*${infile}*R2*`
    mkdir -p ../NGS/02_trimmed
    java -jar $TRIMMO_JAR_FILE PE -threads 4 $R1 $R2
../NGS/02_trimmed/${infile}_clean.1.fq
../NGS/02_trimmed/${infile}_clean_unpaired.1.fq
../NGS/02_trimmed/${infile}_clean.2.fq
../NGS/02_trimmed/${infile}_clean_unpaired.2.fq
ILLUMINACLIP:$TRIMMO_ADAPTOR_FILE_PE:2:15:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:15:30 MINLEN:50
  fi
```

```
  mkdir -p 07_Pilon/${infile}
  cd 07_Pilon/${infile}
  mkdir -p index
  bwa index -p index/draft ../../06_CD-HIT/${infile}/${infile}.fa
  R1=`ls ../../../NGS/02_trimmed/*${infile}*clean.1.fq`
  R2=`ls ../../../NGS/02_trimmed/*${infile}*clean.2.fq`
  bwa mem -t 16 index/draft ${R1} ${R2} |
/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools sort -@ 10 -
O bam -o align.bam
 /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools index -@ 10
align.bam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools sort -n
align.bam > align.sort.bam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools fixmate -m
align.sort.bam fixmate.bam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools sort -o
align.som.bam fixmate.bam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools markdup
align.som.bam align_markdup.bam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools view -@ 10
-q 30 -b align_markdup.bam > align_filter.bam
  /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/samtools index -@
10 align_filter.bam
  java -Xmx5G -jar ${Software}/pilon-1.23.jar --genome ../../06_CD-
HIT/${infile}/${infile}.fa --frags align_filter.bam \
      --fix snps,indels \
      --output ${infile}.pilon
  cd ../..
fi
cd ..
```

3 *Hybrid Assembly*

```
#!/bin/bash
#SBATCH --cpus-per-task=16
#SBATCH -o slurm.%N.%j.out        # STDOUT
#SBATCH -e slurm.%N.%j.err        # STDERR

cd G2G3
infile=$1
NGS_PATH=
G3_PATH=
Software=
```

```
R1=`ls ../NGS/05_Removed/${infile}_virome_1.fastq`
R2=`ls ../NGS/05_Removed/${infile}_virome_2.fastq`
CCS="../G3/02_Removed/${infile}"
#rm -rf 01_OPERA_MS_assembly/${infile}
mkdir -p 01_OPERA_MS_assembly/${infile}
chmod 777 01_OPERA_MS_assembly/${infile}

perl OPERA-MS.pl \
    --short-read1 $R1 \
    --short-read2 $R2 \
    --long-read ${CCS}/${infile}.virome.fastq \
    --out-dir 01_OPERA_MS_assembly/${infile} \
    --contig-len-thr 1000 \
    --num-processors 64 \
    --polishing --no-strain-clustering --no-ref-clustering


#metaSPAdes
mkdir -p 02_metaSPAdes/${infile}

/mnt/raid6/sunchuqing/Softwares/SPAdes-3.13.1-
Linux/bin/metaspades.py \
    --pacbio ${CCS}/${infile}.virome.fastq \
    -1 $R1 \
    -2 $R2 \
    -o 02_metaSPAdes/${infile} \
    -t 16 \
    -m 750

mkdir -p 02_CD-HIT/${infile}
cat 02_metaSPAdes/${infile}/contigs.fasta
01_OPERA_MS_assembly/${infile}/contigs.fasta > 02_CD-
HIT/${infile}/${infile}.all.fa
/mnt/raid6/sunchuqing/Softwares/cdhit-master/cd-hit-est -i 02_CD-
HIT/${infile}/${infile}.all.fa -o 02_CD-HIT/${infile}/${infile}.fa
-c 0.95 -n 8 -T 16 -M 16000
cd ..
```

Additional analysis

4 *Viral annotation*

```
#!/usr/bin/bash
```

```
infile=$1
export PATH=/mnt/raid8/sunchuqing/Softwares/blast2.2.26/ncbi-blast-
2.2.26+/bin:/mnt/raid7/sunchuqing/Softwares/bin:$PATH:/home/sunchuqi
ng/.local/bin:/mnt/raid6/sunchuqing/Softwares/VirSorter/VirSorter/bi
n:/mnt/raid6/sunchuqing/Softwares/VirSorter/VirSorter:/usr/bin:/mnt/
raid6/sunchuqing/Softwares/miniconda3/envs/virsorter/bin:/mnt/raid6/
sunchuqing/Softwares/PPR-
Meta:/mnt/raid6/sunchuqing/Softwares/miniconda3/envs/virsorter/bin:/
usr/local/bin
export
LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/mnt/raid6/sunchuqing/Softwares/MCR
/v94/bin/glnxa64:/mnt/raid6/sunchuqing/Softwares/MCR/v94/runtime/gln
xa64:/mnt/raid6/sunchuqing/Softwares/MCR/v94/sys/os/glnxa64:/mnt/rai
d6/sunchuqing/Softwares/MCR/v94/extern/glnxa64:/mnt/raid3/wchen/mini
conda2/pkgs/libgcc-7.2.0-h69d50b8_2/lib

echo "Dealing with $infile"

export
LD_LIBRARY_PATH=/mnt/raid8/sunchuqing/Softwares/lib/perllib:$LD_LIBR
ARY_PATH
export
PERL5LIB=/mnt/raid8/sunchuqing/Softwares/lib/perllib:$PERL5LIB
fafile="00_CAT/${infile}_cdhit.len1.5k.fa"
seq="00_CAT/${infile}_cdhit.len1.5k.fa"
if [ ! -s $seq ];then
    awk '/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0" ":$0 }'
 00_CAT/${infile}_cdhit.fa |awk 'length($NF)>=1500 {print
$1"\n"$NF}' > 00_CAT/${infile}_cdhit.len1.5k.fa
fi

n1=`ls 01_Glimmer/${infile}/*.predict|wc -l`
n2=`grep -c ">" 00_CAT/${infile}_cdhit.len1.5k.fa `

if [ $n1 -ne $n2 ];then
    #awk '/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0" ":$0 }'
 00_CAT/${infile}_cdhit.fa |awk 'length($NF)>=1500 {print
$1"\n"$NF}'| > 00_CAT/${infile}_cdhit.len1.5k.fa
    #awk '/^>/{s=++num}{print >
"01_Glimmer/"file"/"file"_"s"";close("01_Glimmer/"file"/"file"_"s"")
}' file="${infile}" $fafile
    rm -rf 01_Glimmer/${infile}
    mkdir -p 01_Glimmer/${infile}
    #awk '/^>/{s=++num}{print > "01_Glimmer/"file"/"file"_"s""}'
file="${infile}" $fafile
    #awk '/^>/{s=++num}{print >>
"01_Glimmer/"file"/"file"_"s"";close("01_Glimmer/"file"/"file"_"s"")
}' file="${infile}" $fafile
```

```
    awk '/^>/{s=++num}{print > "01_Glimmer/"file"/"file"_"s"";
("01_Glimmer/"file"/"file"_"s"")}' file="${infile}" $fafile
    cd 01_Glimmer/${infile}
    ls | grep -v "\." | grep "." |awk 'BEGIN {FS="."} {print
$1}'|sort|uniq > ../${infile}.list
    while read -r line
    do
        /mnt/raid5/sunchuqing/Softwares/glimmer3.02/bin/long-orfs -
n -t 1.15 ${line} ${line}.longorfs
        /mnt/raid5/sunchuqing/Softwares/glimmer3.02/bin/extract -t
${line} ${line}.longorfs > ${line}.train
        /mnt/raid5/sunchuqing/Softwares/glimmer3.02/bin/build-icm -
r ${line}.icm < ${line}.train
        /mnt/raid5/sunchuqing/Softwares/glimmer3.02/bin/glimmer3 -
o50 -g100 -t30 ${line} ${line}.icm ${line}
        /mnt/raid5/sunchuqing/Softwares/glimmer3.02/bin/extract -t
${line} ${line}.predict > ${line}_predict.fasta
    done < "../${infile}.list"
    cd ../..
fi
if [ ! -s "01_Glimmer/${infile}/${infile}.faa " ];then
    rm 01_Glimmer/${infile}/${infile}.faa
    while read -r line
    do
            declare -i len=0
            declare -i aorf=0
            declare -i orflen=0
            len=`grep '>' 01_Glimmer/${infile}/${line}.predict |awk
'BEGIN {FS=" "} {print $2}' |awk 'BEGIN {FS="_"} {print $2}'`
            orflen=`grep  ">"
01_Glimmer/${infile}/${line}_predict.fasta | awk 'BEGIN {FS="="}
{print $2}'| awk '{sum+=$1}END{print sum}'`
            ((aorf=$orflen*10000))
            if [ $len -gt $aorf ]; then
                echo ${line} >>
 02_Annotate/${infile}/${infile}.no_assignmnt_under10k
                continue
            fi
            sed "s/^>/>${line}_/"
01_Glimmer/${infile}/${line}_predict.fasta >>
 01_Glimmer/${infile}/${infile}.faa
    done < "01_Glimmer/${infile}.list"
fi
#VirSorter
mkdir -p 04_is_phage
cd 04_is_phage
```

```
if [ ! -s "04_Positive/${infile}.VirSorter.genome" ];then
    echo "--Start virSorter--"
    #rm -rf ./01_virsorter_result/${infile}
    mkdir -p 01_virsorter_result

__conda_setup="$('/mnt/raid6/sunchuqing/Softwares/miniconda/bin/cond
a' 'shell.bash' 'hook' 2> /dev/null)"
    if [ $? -eq 0 ]; then
        eval "$__conda_setup"
    else
        if [ -f
"/mnt/raid6/sunchuqing/Softwares/miniconda/etc/profile.d/conda.sh"
]; then
            .
"/mnt/raid6/sunchuqing/Softwares/miniconda/etc/profile.d/conda.sh"
        else
            export
PATH="/mnt/raid6/sunchuqing/Softwares/miniconda/bin:$PATH"
        fi
    fi
    unset __conda_setup
    # # <<< conda initialize <<<
    # source activate virsorter

    conda activate
/mnt/raid6/sunchuqing/Softwares/miniconda3/envs/vs2
    virsorter run \
        -w ./01_virsorter_result/${infile} \
        -i ../00_CAT/${infile}_cdhit.len1.5k.fa \
        --db-dir /mnt/raid8/sunchuqing/Softwares/Virsorterdb\
        -j 16 --rm-tmpdir\
        --tmpdir ./01_virsorter_result/${infile}/temp --min-score
0.7
    conda deactivate
    # mkdir -p 02_vir_positive
    # cat
./01_virsorter_result/${infile}/Predicted_viral_sequences/VIRSorter_
cat-1.fasta
./01_virsorter_result/${infile}/Predicted_viral_sequences/VIRSorter_
cat-2.fasta > 02_vir_positive/${infile}.vir.fasta
    # #VirSprter Positive 基因组 02_vir_positive/${infile}.vir.fasta
    mkdir -p 04_Positive
    #grep '>' 02_vir_positive/${infile}.vir.fasta |sed
's/>VIRSorter_//g'|sed 's/-cat_1//g'|sed 's/-cat_2//g'|sed 's/-
circular//g' > 04_Positive/${infile}.VirSorter.genome
    cat ./01_virsorter_result/${infile}/final-viral-score.tsv | awk
'BEGIN {FS="|"} {print $1}'|sort|uniq  >
04_Positive/${infile}.VirSorter.genome
```

```
    echo "--End virSorter--"
fi

#Complete
if [ ! -s "04_Positive/${infile}.cir.genome" ];then
    mkdir -p 03_circle/db 03_circle/${infile}
    /mnt/raid8/sunchuqing/Softwares/blast2.2.26/ncbi-blast-
2.2.26+/bin/makeblastdb -in ../00_CAT/${infile}_cdhit.len1.5k.fa -
out 03_circle/db/${infile} -dbtype nucl
    blastall -p blastn \
        -i ../00_CAT/${infile}_cdhit.len1.5k.fa  \
        -d 03_circle/db/${infile} \
        -o 03_circle/${infile}/${infile}.cir.tab \
        -m 8 -e 1e-5 -a 16
    awk '/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0" ":$0 }'
../00_CAT/${infile}_cdhit.len1.5k.fa |awk '{print
$1","length($NF)}'|sed 's/>//g' > 03_circle/${infile}/${infile}.len
    awk 'BEGIN {FS="\t"} $1==$2 && $7==1 && $3==100.00 {print $0}'
03_circle/${infile}/${infile}.cir.tab >
03_circle/${infile}/${infile}.cir711.tab
    rm 03_circle/${infile}/${infile}.cir.len.tab
    while read -r line
    do
        contig=`echo $line |awk 'BEGIN {FS=","} {print $1}'`
        sed "s/^$contig\t/$line\t/g"
03_circle/${infile}/${infile}.cir711.tab |grep "$line" >>
03_circle/${infile}/${infile}.cir.len.tab
    done <"03_circle/${infile}/${infile}.len"
    awk 'BEGIN {FS="[,\t]"} ( $2==$10 || $2==$11 ) && $2!=$9 {print
$0}' 03_circle/${infile}/${infile}.cir.len.tab >
03_circle/${infile}/${infile}.circle.tab
    awk 'BEGIN {FS=","} {print $1}'
03_circle/${infile}/${infile}.circle.tab |sed 's/^/>/g'  >
03_circle/${infile}/${infile}.complete
    awk '/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0"\n":$0 }'
../00_CAT/${infile}_cdhit.len1.5k.fa >
03_circle/${infile}/${infile}.fna
    #grep -w -A 1 -Ff 03_circle/${infile}/${infile}.complete
03_circle/${infile}/${infile}.fna |grep -v "-" >
03_circle/${infile}/${infile}.complete.fa
    #mkdir -p 04_complete
    #cp 03_circle/${infile}/${infile}.complete.fa 04_complete/
    sed 's/>//g' 03_circle/${infile}/${infile}.complete >
04_Positive/${infile}.cir.genome
fi
#成环基因组 03_circle/${infile}/${infile}.complete

#02_vir_positive/${infile}.vir.fasta
```

```
#pip install numpy
#pip install h5py
#pip install tensorflow==1.4.1  #CPU version
#pip install keras==2.0.8
if [ ! -s "04_Positive/${infile}.ppr.genome" ];then
    pathh=`pwd`
    #PPR_Meta
    # export PATH=/usr/bin:$PATH
    # >>> conda initialize >>>
    # !! Contents within this block are managed by 'conda init' !!

__conda_setup="$('/mnt/raid6/sunchuqing/Softwares/miniconda/bin/cond
a' 'shell.bash' 'hook' 2> /dev/null)"
    if [ $? -eq 0 ]; then
        eval "$__conda_setup"
    else
        if [ -f
"/mnt/raid6/sunchuqing/Softwares/miniconda/etc/profile.d/conda.sh"
]; then
            .
"/mnt/raid6/sunchuqing/Softwares/miniconda/etc/profile.d/conda.sh"
        else
            export
PATH="/mnt/raid6/sunchuqing/Softwares/miniconda/bin:$PATH"
        fi
    fi
    unset __conda_setup

    # <<< conda initialize <<<
    conda activate
/mnt/raid6/sunchuqing/Softwares/miniconda3/envs/tensorflow

    pathh=`pwd`
    mkdir -p 03_PPR_META
    cd /mnt/raid6/sunchuqing/Softwares/PPR-Meta
    ./PPR_Meta $pathh/../00_CAT/${infile}_cdhit.len1.5k.fa
$pathh/03_PPR_META/${infile}.csv
    cd $pathh
    awk 'BEGIN {FS=","} $3>0.7 {print $1}'
 03_PPR_META/${infile}.csv |awk 'BEGIN {FS=" "} {print $1}' >
04_Positive/${infile}.ppr.genome
    conda deactivate
fi

#VirFinder
#Bork
if [ ! -s "04_Positive/${infile}.virfiner.genome" ];then
    mkdir -p 03_VirFinder
```

```
    /usr/bin/Rscript  /mnt/raid5/sunchuqing/Buffalo_gut/VirFinder.R
../00_CAT/${infile}_cdhit.len1.5k.fa 03_VirFinder/${infile}.csv
    #VirFinder输出文件
    awk 'BEGIN {FS=","} $3>0.6 {print $1}'
03_VirFinder/${infile}.csv|awk 'BEGIN {FS=" "} {print $1}' >
04_Positive/${infile}.virfiner.genome
fi
#blast Virus ref
if [ ! -s "04_Positive/${infile}.ref.genome" ];then
    mkdir -p 03_Blast_m8/${infile}
    blastall -p blastn \
        -i ../00_CAT/${infile}_cdhit.len1.5k.fa \
        -d /mnt/raid6/sunchuqing/Database/Virus/virus \
        -o 03_Blast_m8/${infile}/${infile}.m8.tab \
        -m 8 -e 1e-10 -a 16
    grep -v -wFf /mnt/raid6/sunchuqing/Database/Virus/not.list
03_Blast_m8/${infile}/${infile}.m8.tab >tmp && mv tmp
03_Blast_m8/${infile}/${infile}.m8.tab
    #echo "chrom start   end" > 03_Blast_m8/${infile}/${infile}.bed
    awk 'BEGIN {FS="\t"} $3>50 {print $1 "\t" $7 "\t" $8 }'
03_Blast_m8/${infile}/${infile}.m8.tab |sort -k 1 -t $'\t' |sed
's/[[:space:]]*$//'|tr -d " "|sort -k1,1 -k2,2n>
03_Blast_m8/${infile}/${infile}.bed.1
    /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/bedtools merge -i
03_Blast_m8/${infile}/${infile}.bed.1
>03_Blast_m8/${infile}/${infile}.bed
    sed 's/,/\t/g' 03_circle/${infile}/${infile}.len >
03_Blast_m8/${infile}/${infile}.len
    cut -f 1 03_Blast_m8/${infile}/${infile}.len >
03_Blast_m8/${infile}/${infile}.list
    /mnt/raid6/sunchuqing/Softwares/miniconda3/bin/bedtools
genomecov -i 03_Blast_m8/${infile}/${infile}.bed -g
03_Blast_m8/${infile}/${infile}.len |awk 'BEGIN {FS="\t"} $2>=1
{print $0}'|grep -v "genome" >
03_Blast_m8/${infile}/${infile}.bedtools
    rm 03_Blast_m8/${infile}/${infile}.genomecov.csv

    # while read -r line
    # do
    #     grep "$line" 03_Blast_m8/${infile}/${infile}.bedtools |
awk 'BEGIN {FS="\t"} {sum += $NF} {print $1 "," sum*100}'|tail -n 1
>> 03_Blast_m8/${infile}/${infile}.genomecov.csv
    #     name=`grep "$line"
03_Blast_m8/${infile}/${infile}.bedtools |wc -l`
    #     if [ $name -eq 0 ];then
    #         echo "$line,0" >>
03_Blast_m8/${infile}/${infile}.genomecov.csv
    #     fi
```

```
    # done <"03_Blast_m8/${infile}/${infile}.list"
    awk 'BEGIN {FS=","} $NF>0.9 {print $1}'
03_Blast_m8/${infile}/${infile}.genomecov.csv >
04_Positive/${infile}.ref.genome
    # awk 'BEGIN {FS=","} $2>90  {print $0}'
03_Blast_m8/${infile}/${infile}.genomecov.csv >
03_Blast_m8/${infile}/${infile}.genomecov.over90.csv
    # awk 'BEGIN {FS=","} {print $1}'
03_Blast_m8/${infile}/${infile}.genomecov.over90.csv >
04_Positive/${infile}.ref.genome
fi
#覆盖度超过90的基因组
03_Blast_m8/${infile}/${infile}.genomecov.over90.csv
pathh=`pwd`
#blast pVOGs
if [ ! -s "04_Positive/${infile}.pvog.genome" ];then
    mkdir -p 03_Blast_pVOG/${infile}
    rm 03_Blast_pVOG/${infile}/${infile}_cds_num.csv
    cd ../01_Glimmer/${infile}
    ls | grep -v "\." | grep "." |awk 'BEGIN {FS="."} {print
$1}'|sort|uniq > ../${infile}.list
    cd $pathh
    blastall -p blastx -i ../01_Glimmer/${infile}/${infile}.faa -d
/mnt/raid6/sunchuqing/Database/Virus/blastdb/POGseqs \
            -o 03_Blast_pVOG/${infile}/${infile}.m8.tab \
            -m 8 -e 1e-10 -a 16
    while read -r line
    do
        file="../01_Glimmer/${infile}/${line}"
        filename=`echo "$file" | awk 'BEGIN {FS="/"} {print $NF}'`
        CDSnum=`grep '>' ${file}_predict.fasta |wc -l `
        name=`grep '>' ../01_Glimmer/${infile}/${line} |head -n
1|sed 's/>//g' `
        length=`grep -w "$name"
03_Blast_m8/${infile}/${infile}.len`
        hitnum=`grep "${filename}_"
03_Blast_pVOG/${infile}/${infile}.m8.tab |awk 'BEGIN {FS="\t"} $3>50
{print $1} ' |sort |uniq  |wc -l`
        echo "$length,$CDSnum,$hitnum"|sed 's/\t/,/g'|awk 'BEGIN
{FS=","} $3>3 && $2/5000<$3 && $2/5000<$4 {print $0}' >>
03_Blast_pVOG/${infile}/${infile}_cds_num.csv
    done <"../01_Glimmer/${infile}.list"
    awk 'BEGIN {FS=","} {print $1}'
 03_Blast_pVOG/${infile}/${infile}_cds_num.csv >
04_Positive/${infile}.pvog.genome
fi
mkdir -p 05_Phage_Positive
```

```
awk 'BEGIN {FS=","} {print $1}'
 03_Blast_pVOG/${infile}/${infile}_cds_num.csv >
04_Positive/${infile}.pvog.genome
awk 'BEGIN {FS=","} {print $1}'
03_Blast_m8/${infile}/${infile}.genomecov.over90.csv >
04_Positive/${infile}.ref.genome
sed 's/>//g' 03_circle/${infile}/${infile}.complete >
04_Positive/${infile}.cir.genome
awk 'BEGIN {FS=","} $3>0.6 {print $1}'
03_VirFinder/${infile}.csv|awk 'BEGIN {FS=" "} {print $1}' >
04_Positive/${infile}.virfiner.genome
awk 'BEGIN {FS=","} $3>0.7 {print $1}'  03_PPR_META/${infile}.csv
|awk 'BEGIN {FS=" "} {print $1}' > 04_Positive/${infile}.ppr.genome
#grep '>' 02_vir_positive/${infile}.vir.fasta |sed
's/>VIRSorter_//g'|sed 's/-cat_1//g'|sed 's/-cat_2//g'|sed 's/-
circular//g' > 04_Positive/${infile}.VirSorter.genome

#cat 04_Positive/${infile}*|sort |uniq -c|sort -n|awk 'BEGIN {FS="
"} $1>=1 {print $2}'|sed 's/_length/ length/g'|awk 'BEGIN {FS=" "}
{print $1}'  >05_Phage_Positive/${infile}.phage.genome
cat 04_Positive/${infile}*|sort |uniq -c|sort -n|awk 'BEGIN {FS=" "}
$1>=2 {print $2}'|sed 's/_length/ length/g'|awk 'BEGIN {FS=" "}
{print $1}'  >05_Phage_Positive/${infile}.phage.genome2
cat 04_Positive/${infile}.cir.genome >>
05_Phage_Positive/${infile}.phage.genome2
sed 's/_length/ length/g' 03_circle/${infile}/${infile}.fna >
03_circle/${infile}/${infile}.fna2
#grep -w -A 1 -Ff 05_Phage_Positive/${infile}.phage.genome
03_circle/${infile}/${infile}.fna2 |grep -v -e "--" >
05_Phage_Positive/${infile}.phage.genome.fa
grep -w -A 1 -Ff 05_Phage_Positive/${infile}.phage.genome2
03_circle/${infile}/${infile}.fna2 |grep -v -e "--" |awk
'/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0" ":$0 }' |awk '{print
$1","length($NF)}'|sed 's/>//g' |awk 'BEGIN {FS=","} $2>1500 {print
$1}'|sort|uniq > 05_Phage_Positive/${infile}_fullfill2
grep -w -A 1 -Ff 05_Phage_Positive/${infile}.phage.genome2
03_circle/${infile}/${infile}.fna2 |grep -v -e "--" |awk
'/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0" ":$0 }' |awk '{print
$1","length($NF)}'|sed 's/>//g' |awk 'BEGIN {FS=","} $2>1500 {print
$0}'|sort|uniq > 05_Phage_Positive/${infile}_fullfill2.length.csv

grep -w -A 1 -Ff 05_Phage_Positive/${infile}_fullfill2
03_circle/${infile}/${infile}.fna2 |grep -v -e "--" >
05_Phage_Positive/${infile}.phage.1.5k.genome.fa
```

5 *UHGG filteration*

```
#UHGG filter----
    seq=../CHGV.filtered.fa
    infile=CHGV
    blastall -p blastn \
        -i $seq \
        -d /mnt/raid8/sunchuqing/Database/UHGG/uhgg\
        -o ./${infile}.uhgg.m8.tab \
        -m 8 -e 1e-10 -a 20

    blastall -p blastn \
        -i $seq \
        -d
/mnt/raid7/sunchuqing/Human_Gut_Phage/HGYG/db/HGYG_prophage \
        -o ./${infile}.uhgg.pro.m8.tab \
        -m 8 -e 1e-10 -a 20
    awk 'BEGIN {FS="\t"} $3>90 {print $1 "\t" $7 "\t" $8 }'
./${infile}.uhgg.m8.tab |sort -k 1 -t $'\t' |sed
's/[[:space:]]*$//'|tr -d " "|sort -k1,1 -k2,2n>
${infile}.uhgg.bed.1
    bedtools merge -i ${infile}.uhgg.bed.1 >${infile}.uhgg.bed
    seqtk comp $seq|cut -f 1,2 > ${infile}.uhgg.len
    bedtools genomecov -i ${infile}.uhgg.bed -g ${infile}.uhgg.len
|awk 'BEGIN {FS="\t"} $2>=1 {print $0}'|grep -v "genome" >
${infile}.uhgg.bedtools

    awk 'BEGIN {FS="\t"} $3>90 {print $1 "\t" $7 "\t" $8 }'
./${infile}.uhgg.pro.m8.tab |sort -k 1 -t $'\t' |sed
's/[[:space:]]*$//'|tr -d " "|sort -k1,1 -k2,2n>
${infile}.uhgg.pro.bed.1
    bedtools merge -i ${infile}.uhgg.pro.bed.1
>${infile}.uhgg.pro.bed
    bedtools genomecov -i ${infile}.uhgg.pro.bed -g
${infile}.uhgg.len |awk 'BEGIN {FS="\t"} $2>=1 {print $0}'|grep -v
"genome" > ${infile}.uhgg.pro.bedtools


    while read -r lenf
    do
        name=`echo $lenf |awk 'BEGIN {FS=" "} {print $1}'`
        len=`echo $lenf |awk 'BEGIN {FS=" "} {print $2}'`
        line=`grep -w "$name" $infile.uhgg.bedtools`
        pct=0
        if [ `grep -w "$name" $infile.uhgg.bedtools|wc -l` == 1
];then
            pct=`echo $line |awk 'BEGIN {FS=" "} {print $NF}'|awk
```

```
'{printf("%f",$0)} '`
        fi
        pct2=0
        if [ `grep "$name" ${infile}.uhgg.pro.bedtools|wc -l` == 1
];then
            pct2=`grep "$name" ${infile}.uhgg.pro.bedtools |awk
'BEGIN {FS=" "}  {print $NF}'|awk '{printf("%f",$0)} '`
        fi
        pct3=`echo "$pct-$pct2"|bc`
        if [ $(echo "$pct3 < 0"|bc) -eq 1 ];then
            pct3=0
        fi
        echo "$name,$pct3,$len"
    done < "${infile}.uhgg.len" > ${infile}.np90.csv
    awk 'BEGIN {FS=","wc} $2>0.5' CHGV.np90.csv|wc
    awk 'BEGIN {FS=","} $2>0.5 {print $1}'
UHGG.filtered/CHGV.np90.csv > UHGG.filtered/CHGV.uhgg.txt
    seqkit grep -v -f UHGG.filtered/CHGV.uhgg.txt CHGV.filtered.fa
> CHGV.filtered.uhgg.fa
```

6 *checkv*

```
infile=$1
export CHECKVDB=/mnt/raid6/sunchuqing/Softwares/checkv/checkv-db-
v0.6
export PATH=/mnt/raid6/sunchuqing/Softwares/miniconda3/bin:$PATH
awk '/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0" ":$0 }'
 00_CAT/${infile}_cdhit.fa |awk 'length($NF)>=1500 {print
$1"\n"$NF}' > 00_CAT/${infile}_cdhit.len1.5k.fa

mkdir -p 05_checkV/${infile}_1.5k
/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/checkv contamination
00_CAT/${infile}_cdhit.len1.5k.fa 05_checkV/${infile}_1.5k -t 16
/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/checkv completeness
00_CAT/${infile}_cdhit.len1.5k.fa 05_checkV/${infile}_1.5k -t 16
/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/checkv
complete_genomes  00_CAT/${infile}_cdhit.len1.5k.fa
05_checkV/${infile}_1.5k
/mnt/raid6/sunchuqing/Softwares/miniconda3/bin/checkv
quality_summary 00_CAT/${infile}_cdhit.len1.5k.fa
05_checkV/${infile}_1.5k
```

7 *RPKM calculation*

```
#prevalence.RPKM
db=CHGV.filtered
seq=CHGV.filtered.fa
# bowtie2-build $seq db/$db
while read -r infile
do
    path="$NGS/"
    R1=`ls $path/$infile*R1*|head -n1`
    R2=`ls $path/$infile*R2*|head -n1`
    if [ -s 04_mapping/${infile}_NGS.bam ];then
        echo "Mapped $infile"
        continue
    fi
    $Software/miniconda3/bin/bowtie2 -x db/$db \
        -1 $R1\
        -2 $R2 \
        -S 04_mapping/${infile}_NGS.sam -p 8
    samtools view -bS 04_mapping/${infile}_NGS.sam >
04_mapping/${infile}_NGS.bam
    rm 04_mapping/${infile}_NGS.sam
done < "sam.list"


while read -r infile
do
    if [ ! -s 04_mapping/${infile}_NGS.bam ];then
        echo "Unmapped $infile"
        continue
    fi
    if [ ! -s 06_bamst_cov/${infile}/chromosomes.report ];then
        echo "Bamdst $infile"
        mkdir -p 06_bamst_cov/${infile}
        output=06_bamst_cov/${infile}
        samtools sort -@4 04_mapping/${infile}_NGS.bam -o
04_mapping/${infile}_NGS.sort.bam
        cd $output
        $Software/bamdst/bamdst -p ../../CHGV.bed -o ./
../../04_mapping/${infile}_NGS.sort.bam
        cut -f 1,6 chromosomes.report|sed "s/^/${infile}\t/"|awk
'BEGIN {FS="\t"} $3>=50' >>  ../Sample.4x.50.cov.tbl
        cut -f 1,6 chromosomes.report|sed "s/^/${infile}\t/"|awk
'BEGIN {FS="\t"} $3>=50'|cut -f 2 > ./$infile.4x.50.cov.list
        grep -wFf ./$infile.4x.50.cov.list ../../CHGV.bed >
```

```
./$infile.4x.50.cov.bed
        cd ../..
    fi
    if [ ! -s 07_bam2rpkm/${infile}.rpkm.txt ];then
        echo "Bam2rpkm $infile"
            mkdir -p 07_bam2rpkm/
        rm -r 07_bam2rpkm/$infile.rpkm.txt
        bam=04_mapping/${infile}_NGS.sort.bam
        samtools index -@4 $bam
        bed=06_bamst_cov/${infile}/${infile}.4x.50.cov.bed
        echo $infile
        export total_reads=$(samtools idxstats  $bam|awk -F '\t'
'{s+=$3}END{print s}')
        echo The number of reads is $total_reads
        bedtools multicov -bams  $bam -bed $bed |\
            perl -alne '{$len=$F[2]-$F[1];if($len <1 ){print
"$.\t$F[3]\t0" }else{$rpkm=(1000000000*$F[3]/($len*
$ENV{total_reads}));print "$F[0]\t$F[3]\t$rpkm"}}' |\
            sed "s/^/${infile}\t/" >07_bam2rpkm/${infile}.rpkm.txt
    fi
done < "sam.list"
```