February 24, 2021

Server Scince Samples 2.0

What if you have no space to output data?

by Maria Shumilova

01

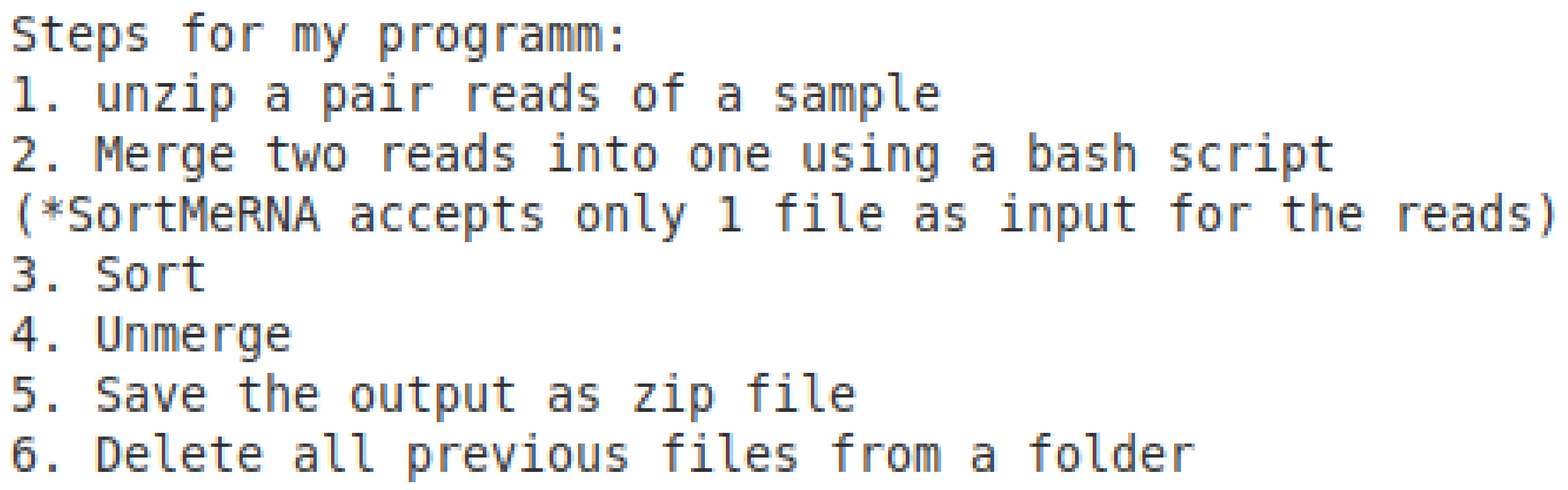
SortmeRNA is a program tool for metatranscriptomic and metagenomic data.
The programm has several neccesary steps and accept ONLY UNZIP files as input.

I have a lot of big size samples and have no space to save the output data on server

My idea

a script which will be consistently get as input a sample, unzip it, pass through the necessary steps of the program, save the output as a zip file on server and remove intermediate files.

PLAN



```
Steps for my programm:
1. unzip a pair reads of a sample
2. Merge two reads into one using a bash script
(*SortMeRNA accepts only 1 file as input for the reads)
3. Sort
4. Unmerge
5. Save the output as zip file
6. Delete all previous files from a folder
```

04

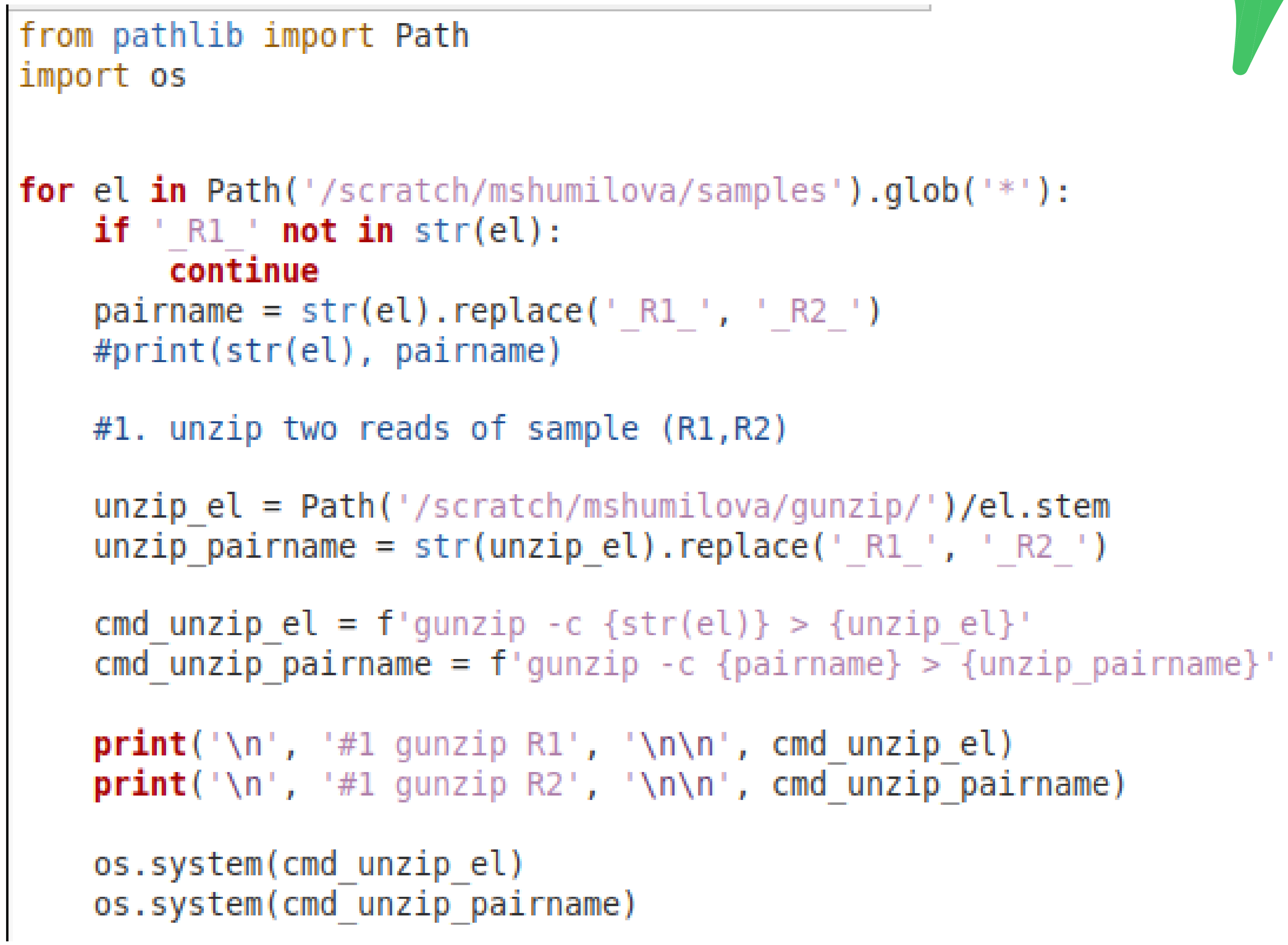
Step by step

Let's Try!

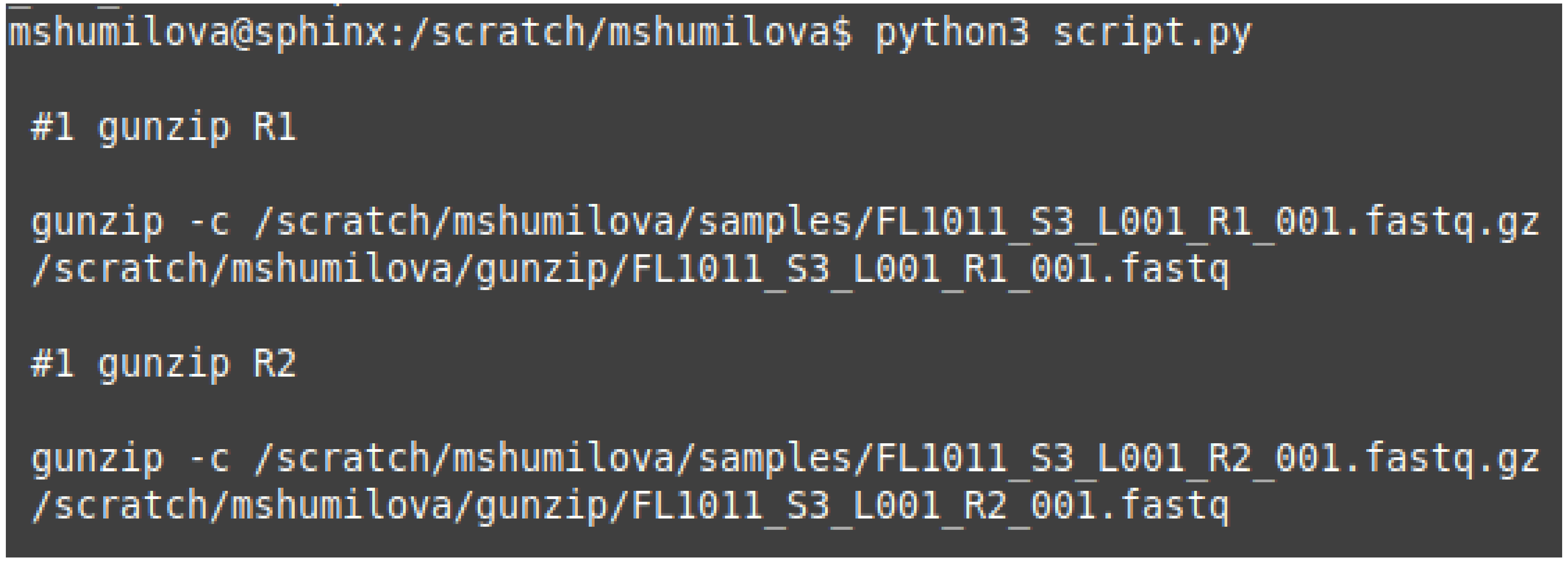
05

1. Unzip

was trying to use the shutil, but it was too difficult for .gz format, so I decided to use bash

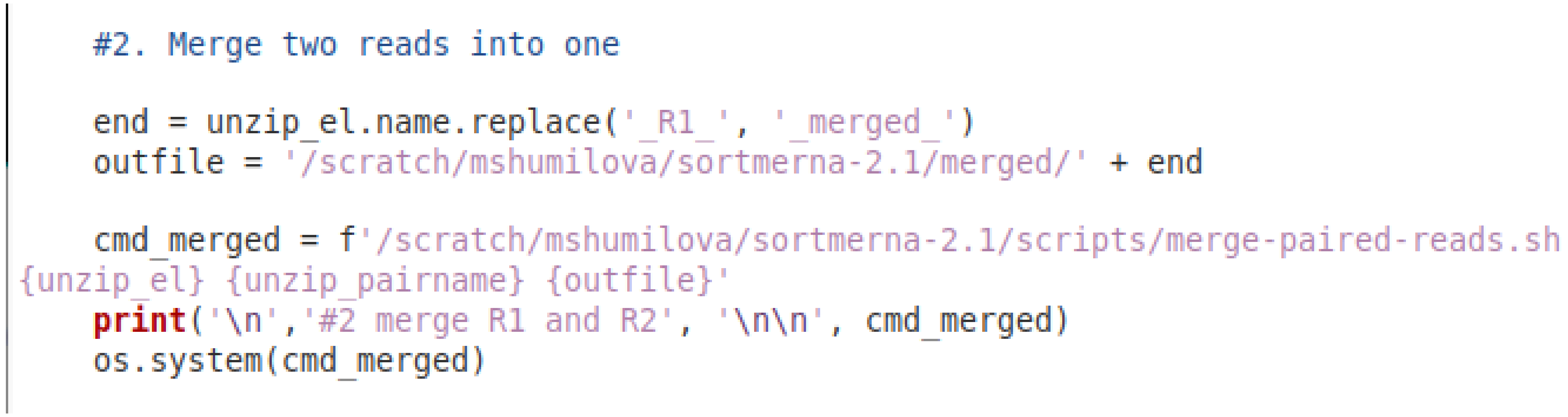


```
from pathlib import Path
import os


for el in Path('/scratch/mshumilova/samples').glob('*'):
    if '_R1_' not in str(el):
        continue
    pairname = str(el).replace('_R1_', '_R2_')
    #print(str(el), pairname)

    #1. unzip two reads of sample (R1,R2)

    unzip_el = Path('/scratch/mshumilova/gunzip/')/el.stem
    unzip_pairname = str(unzip_el).replace('_R1_', '_R2_')

    cmd_unzip_el = f'gunzip -c {str(el)} > {unzip_el}'
    cmd_unzip_pairname = f'gunzip -c {pairname} > {unzip_pairname}'

    print('\n', '#1 gunzip R1', '\n\n', cmd_unzip_el)
    print('\n', '#1 gunzip R2', '\n\n', cmd_unzip_pairname)

    os.system(cmd_unzip_el)
    os.system(cmd_unzip_pairname)
```

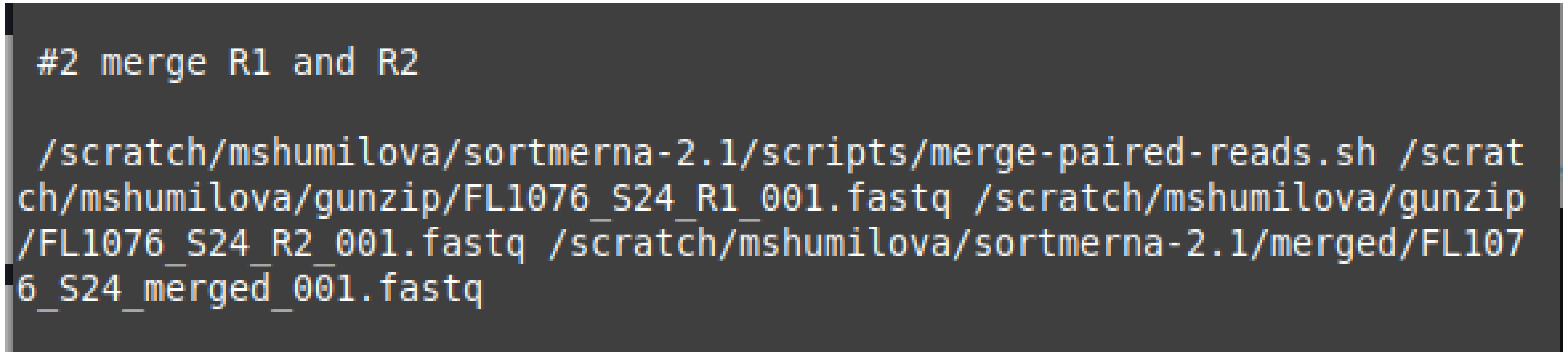


```
mshumilova@sphinx:/scratch/mshumilova$ python3 script.py

 #1 gunzip R1

 gunzip -c /scratch/mshumilova/samples/FL1011_S3_L001_R1_001.fastq.gz
 /scratch/mshumilova/gunzip/FL1011_S3_L001_R1_001.fastq

 #1 gunzip R2

 gunzip -c /scratch/mshumilova/samples/FL1011_S3_L001_R2_001.fastq.gz
 /scratch/mshumilova/gunzip/FL1011_S3_L001_R2_001.fastq
```

2. Merge



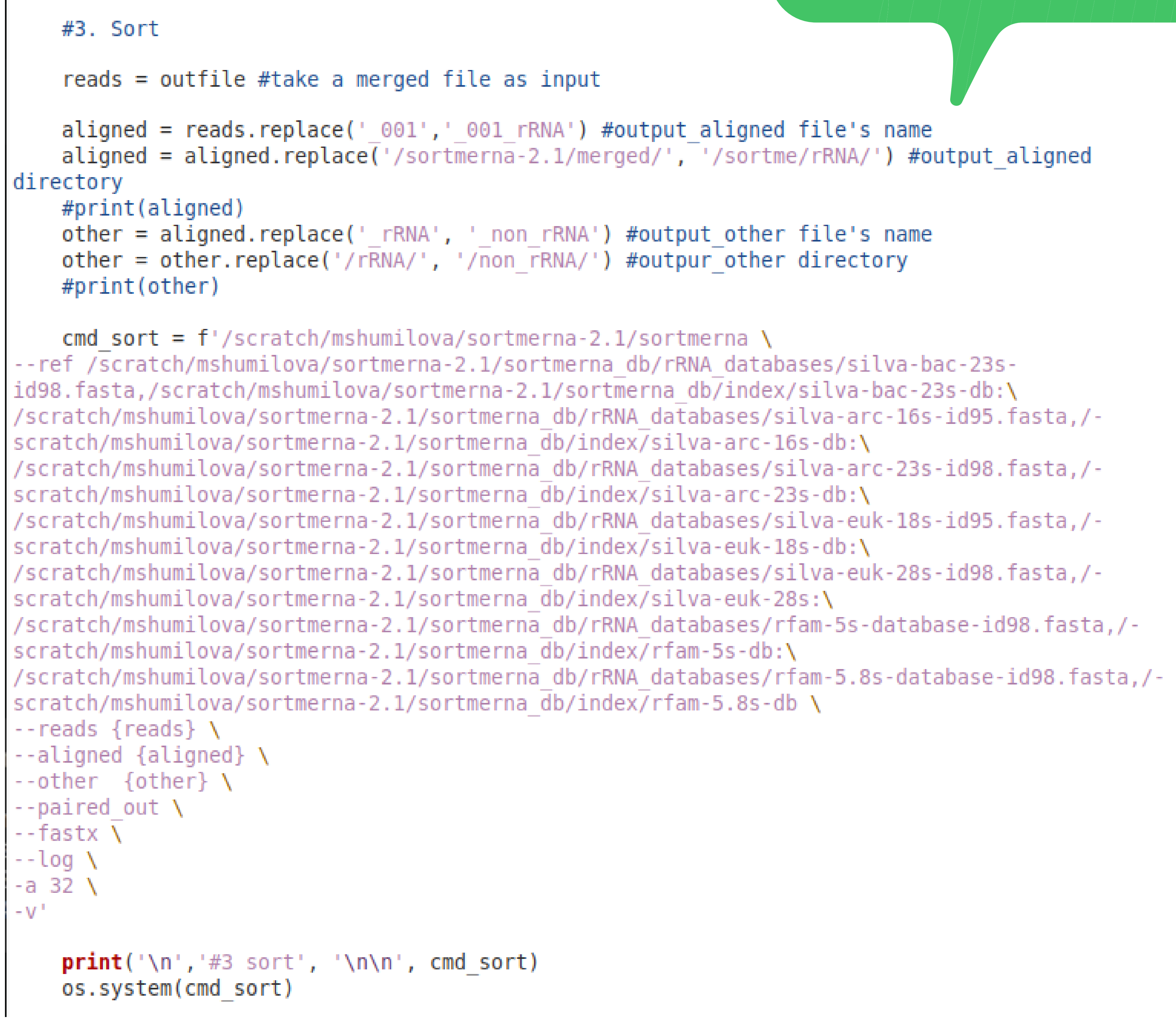
```
    #2. Merge two reads into one

    end = unzip_el.name.replace('_R1_', '_merged_')
    outfile = '/scratch/mshumilova/sortmerna-2.1/merged/' + end

    cmd_merged = f'/scratch/mshumilova/sortmerna-2.1/scripts/merge-paired-reads.sh
{unzip_el} {unzip_pairname} {outfile}'
    print('\n','#2 merge R1 and R2', '\n\n', cmd_merged)
    os.system(cmd_merged)
```



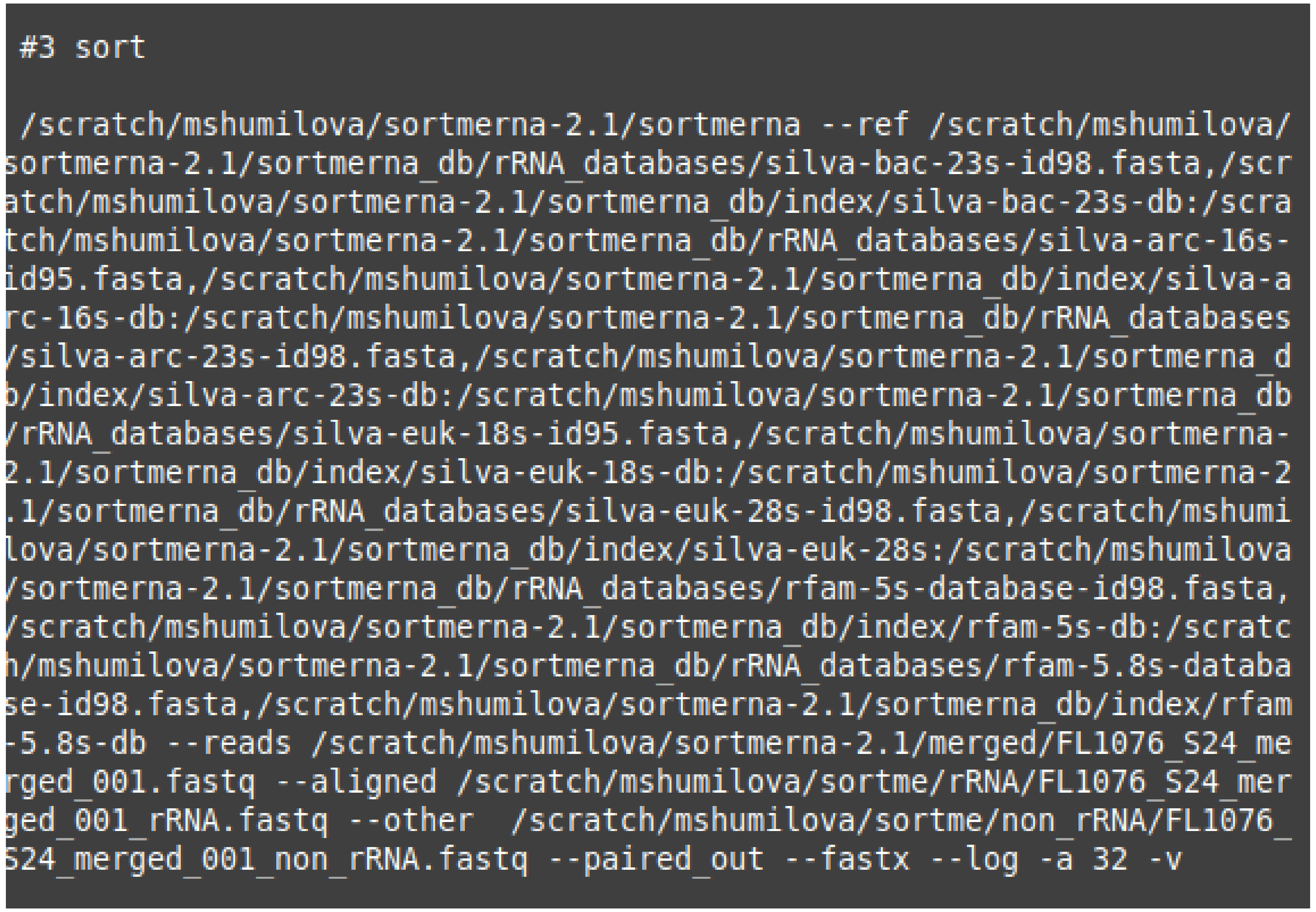
```
 #2 merge R1 and R2

 /scratch/mshumilova/sortmerna-2.1/scripts/merge-paired-reads.sh /scrat
ch/mshumilova/gunzip/FL1076_S24_R1_001.fastq /scratch/mshumilova/gunzip
/FL1076_S24_R2_001.fastq /scratch/mshumilova/sortmerna-2.1/merged/FL107
6_S24_merged_001.fastq
```

3. Sort

the code is long, but that's how this step of SortmeRNA works

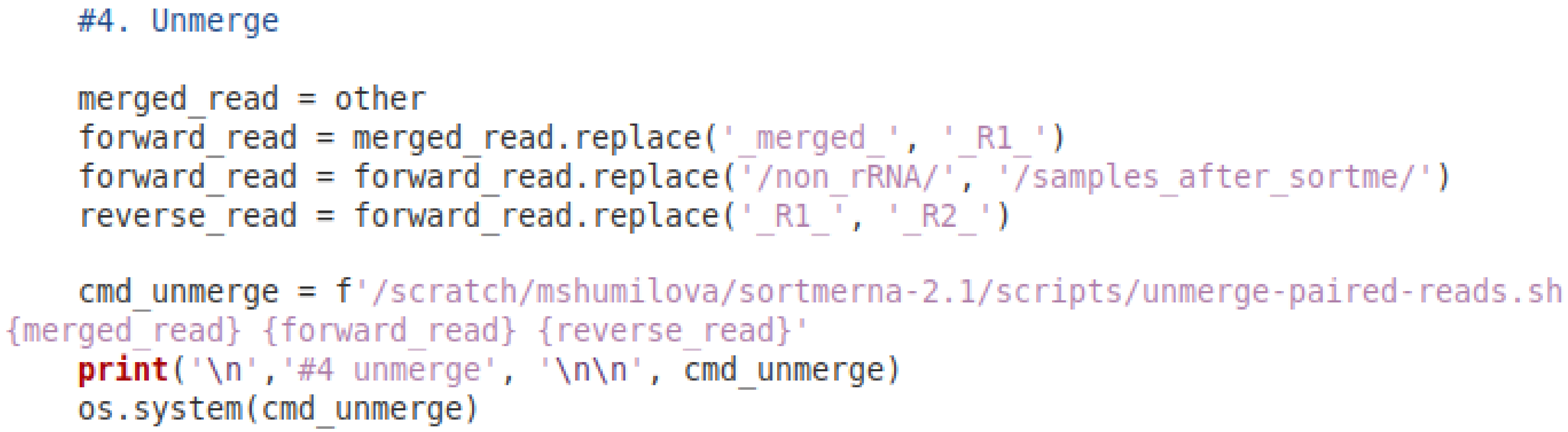


```
    #3. Sort

    reads = outfile #take a merged file as input

    aligned = reads.replace('_001','_001_rRNA') #output_aligned file's name
    aligned = aligned.replace('/sortmerna-2.1/merged/', '/sortme/rRNA/') #output_aligned
directory
    #print(aligned)
    other = aligned.replace('_rRNA', '_non_rRNA') #output_other file's name
    other = other.replace('/rRNA/', '/non_rRNA/') #outpur_other directory
    #print(other)

    cmd_sort = f'/scratch/mshumilova/sortmerna-2.1/sortmerna \
--ref /scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/silva-bac-23s-
id98.fasta,/scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/silva-bac-23s-db:\
/scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/silva-arc-16s-id95.fasta,/-
scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/silva-arc-16s-db:\
/scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/silva-arc-23s-id98.fasta,/-
scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/silva-arc-23s-db:\
/scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/silva-euk-18s-id95.fasta,/-
scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/silva-euk-18s-db:\
/scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/silva-euk-28s-id98.fasta,/-
scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/silva-euk-28s:\
/scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/rfam-5s-database-id98.fasta,/-
scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/rfam-5s-db:\
/scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/rfam-5.8s-database-id98.fasta,/-
scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/rfam-5.8s-db \
--reads {reads} \
--aligned {aligned} \
--other  {other} \
--paired_out \
--fastx \
--log \
-a 32 \
-v'

    print('\n','#3 sort', '\n\n', cmd_sort)
    os.system(cmd_sort)
```

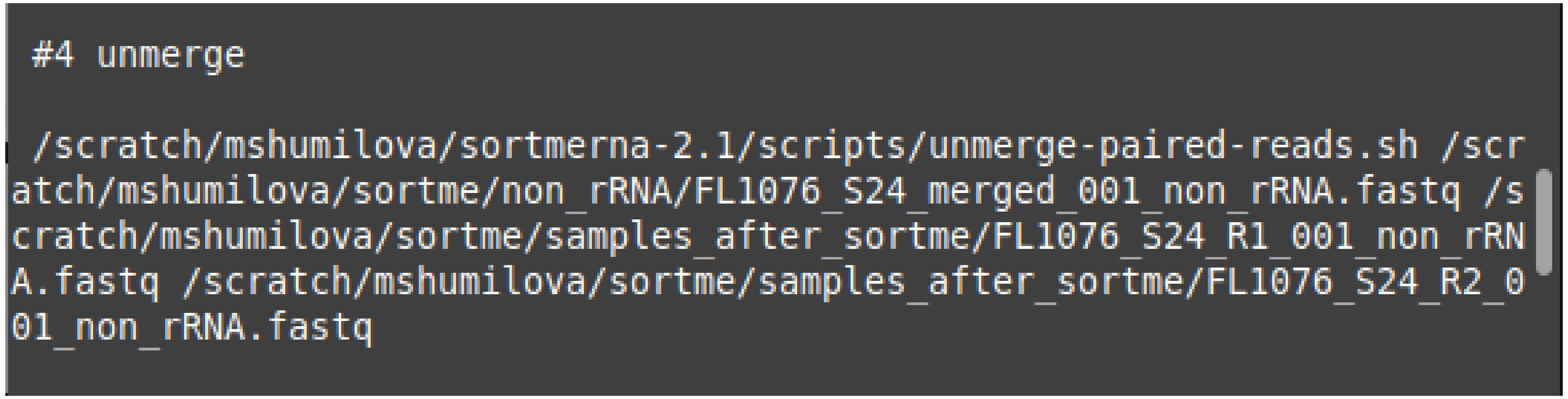


```
#3 sort

 /scratch/mshumilova/sortmerna-2.1/sortmerna --ref /scratch/mshumilova/
sortmerna-2.1/sortmerna_db/rRNA_databases/silva-bac-23s-id98.fasta,/scr
atch/mshumilova/sortmerna-2.1/sortmerna_db/index/silva-bac-23s-db:/scra
tch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/silva-arc-16s-
id95.fasta,/scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/silva-a
rc-16s-db:/scratch/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases
/silva-arc-23s-id98.fasta,/scratch/mshumilova/sortmerna-2.1/sortmerna_d
b/index/silva-arc-23s-db:/scratch/mshumilova/sortmerna-2.1/sortmerna_db
/rRNA_databases/silva-euk-18s-id95.fasta,/scratch/mshumilova/sortmerna-
2.1/sortmerna_db/index/silva-euk-18s-db:/scratch/mshumilova/sortmerna-2
.1/sortmerna_db/rRNA_databases/silva-euk-28s-id98.fasta,/scratch/mshumi
lova/sortmerna-2.1/sortmerna_db/index/silva-euk-28s:/scratch/mshumilova
/sortmerna-2.1/sortmerna_db/rRNA_databases/rfam-5s-database-id98.fasta,
/scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/rfam-5s-db:/scratc
h/mshumilova/sortmerna-2.1/sortmerna_db/rRNA_databases/rfam-5.8s-databa
se-id98.fasta,/scratch/mshumilova/sortmerna-2.1/sortmerna_db/index/rfam
-5.8s-db --reads /scratch/mshumilova/sortmerna-2.1/merged/FL1076_S24_me
rged_001.fastq --aligned /scratch/mshumilova/sortme/rRNA/FL1076_S24_mer
ged_001_rRNA.fastq --other  /scratch/mshumilova/sortme/non_rRNA/FL1076_
S24_merged_001_non_rRNA.fastq --paired_out --fastx --log -a 32 -v
```

4. Unmerge

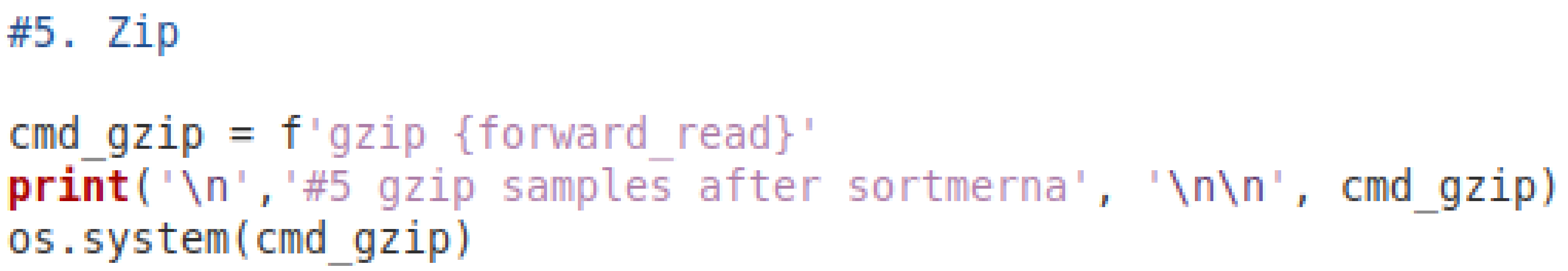


```
    #4. Unmerge

    merged_read = other
    forward_read = merged_read.replace('_merged_', '_R1_')
    forward_read = forward_read.replace('/non_rRNA/', '/samples_after_sortme/')
    reverse_read = forward_read.replace('_R1_', '_R2_')

    cmd_unmerge = f'/scratch/mshumilova/sortmerna-2.1/scripts/unmerge-paired-reads.sh
{merged_read} {forward_read} {reverse_read}'
    print('\n','#4 unmerge', '\n\n', cmd_unmerge)
    os.system(cmd_unmerge)
```

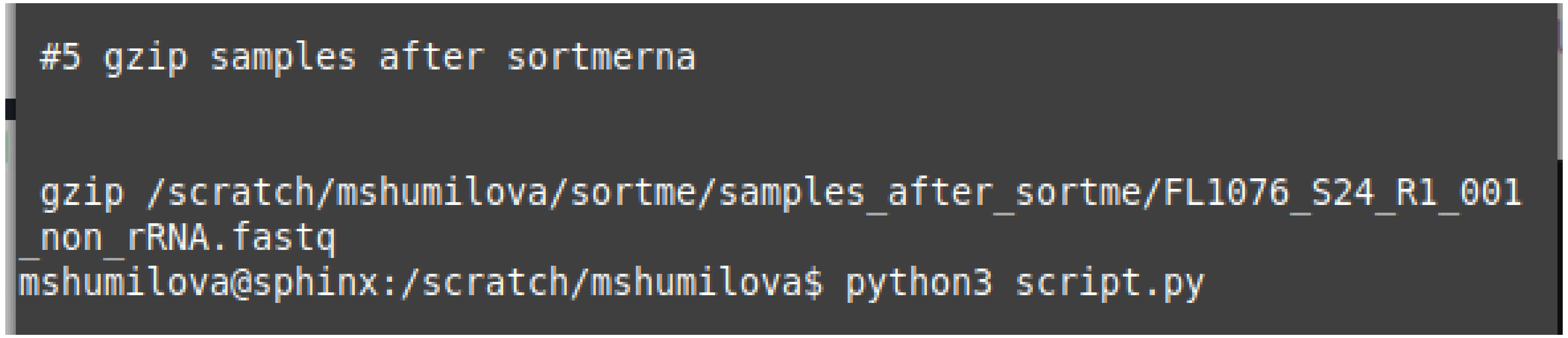


```
 #4 unmerge

 /scratch/mshumilova/sortmerna-2.1/scripts/unmerge-paired-reads.sh /scr
atch/mshumilova/sortme/non_rRNA/FL1076_S24_merged_001_non_rRNA.fastq /s
cratch/mshumilova/sortme/samples_after_sortme/FL1076_S24_R1_001_non_rRN
A.fastq /scratch/mshumilova/sortme/samples_after_sortme/FL1076_S24_R2_0
01_non_rRNA.fastq
```

5. Zip

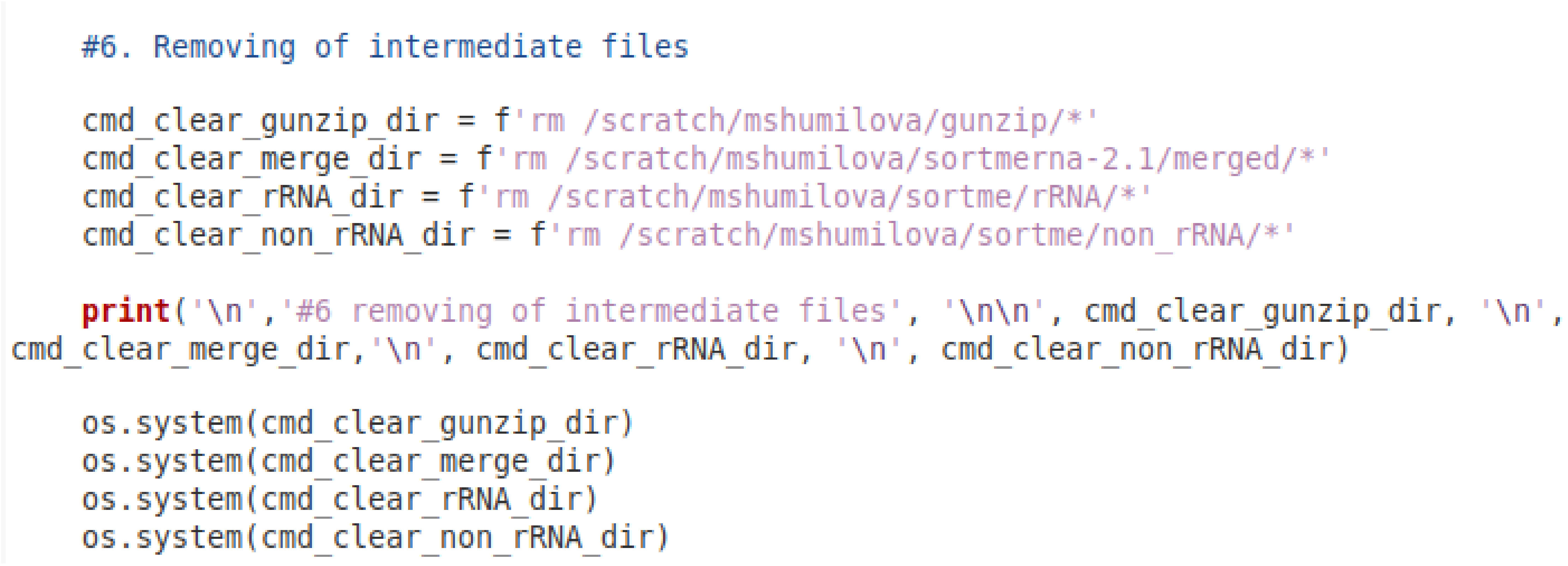


```
#5. Zip

cmd_gzip = f'gzip {forward_read}'
print('\n','#5 gzip samples after sortmerna', '\n\n', cmd_gzip)
os.system(cmd_gzip)
```

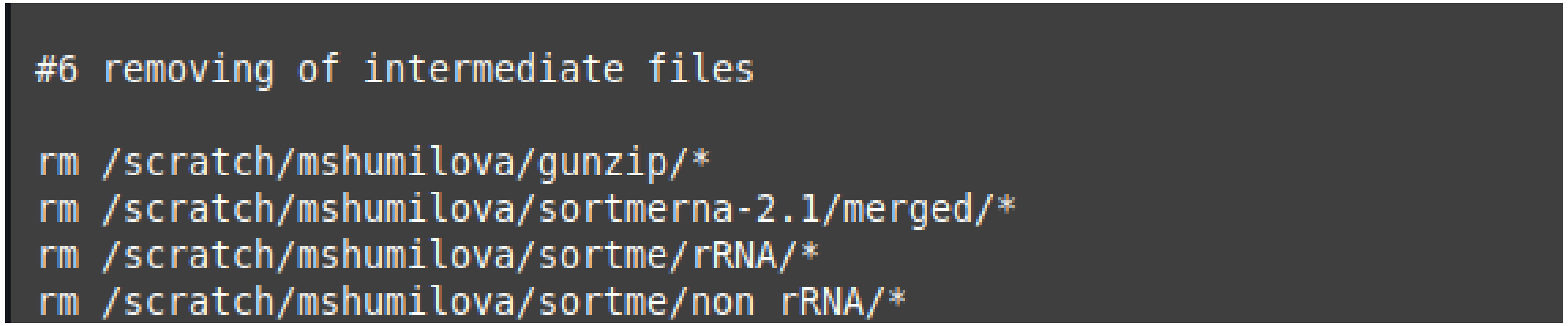


```
 #5 gzip samples after sortmerna


 gzip /scratch/mshumilova/sortme/samples_after_sortme/FL1076_S24_R1_001
_non_rRNA.fastq
mshumilova@sphinx:/scratch/mshumilova$ python3 script.py
```

6. Removing of intermediate files



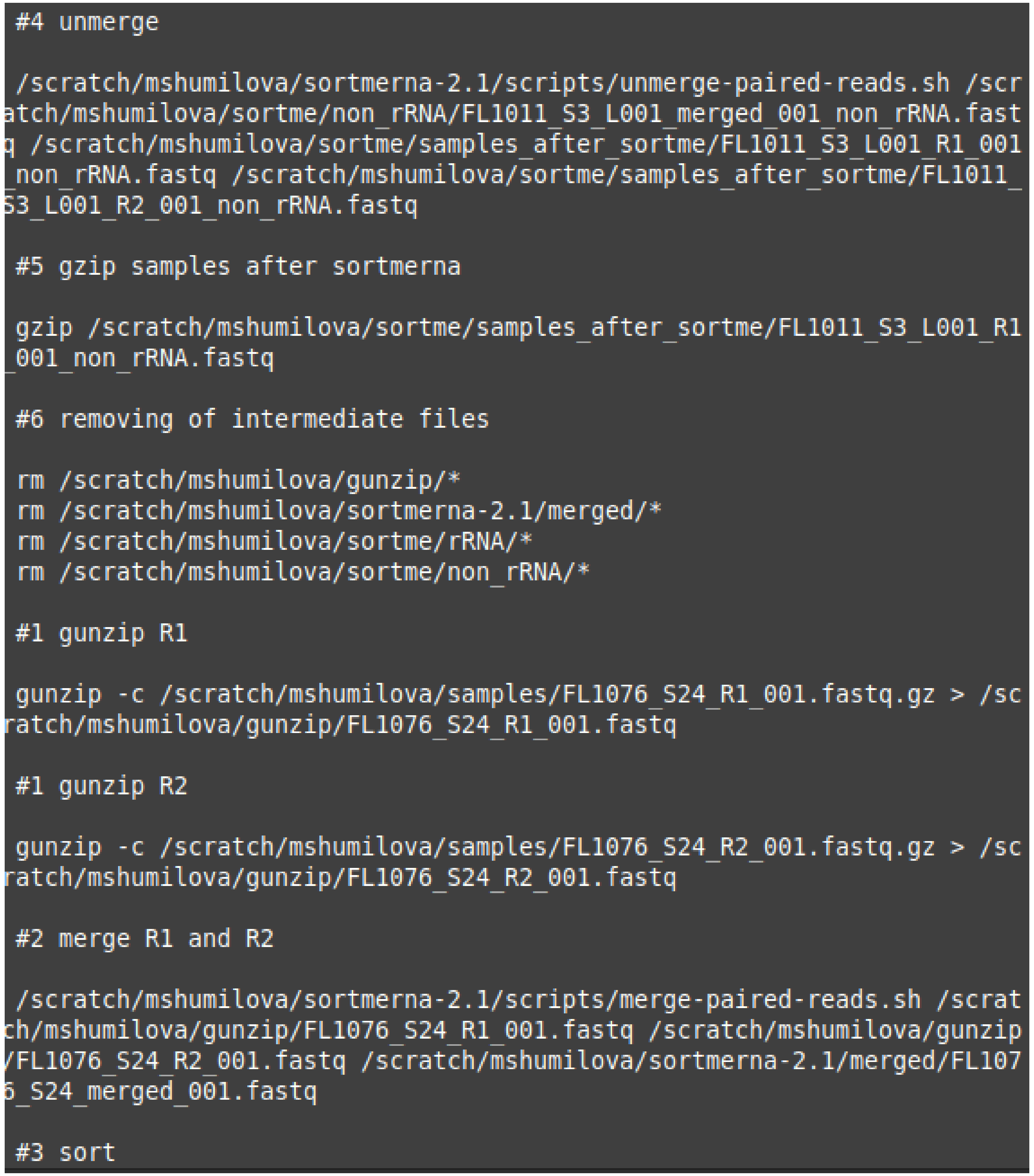
```
    #6. Removing of intermediate files

    cmd_clear_gunzip_dir = f'rm /scratch/mshumilova/gunzip/*'
    cmd_clear_merge_dir = f'rm /scratch/mshumilova/sortmerna-2.1/merged/*'
    cmd_clear_rRNA_dir = f'rm /scratch/mshumilova/sortme/rRNA/*'
    cmd_clear_non_rRNA_dir = f'rm /scratch/mshumilova/sortme/non_rRNA/*'

    print('\n','#6 removing of intermediate files', '\n\n', cmd_clear_gunzip_dir, '\n',
cmd_clear_merge_dir,'\n', cmd_clear_rRNA_dir, '\n', cmd_clear_non_rRNA_dir)

    os.system(cmd_clear_gunzip_dir)
    os.system(cmd_clear_merge_dir)
    os.system(cmd_clear_rRNA_dir)
    os.system(cmd_clear_non_rRNA_dir)
```



```
#6 removing of intermediate files

rm /scratch/mshumilova/gunzip/*
rm /scratch/mshumilova/sortmerna-2.1/merged/*
rm /scratch/mshumilova/sortme/rRNA/*
rm /scratch/mshumilova/sortme/non_rRNA/*
```

Check

Does it work consistently?
YES



```
 #4 unmerge

 /scratch/mshumilova/sortmerna-2.1/scripts/unmerge-paired-reads.sh /scr
atch/mshumilova/sortme/non_rRNA/FL1011_S3_L001_merged_001_non_rRNA.fast
q /scratch/mshumilova/sortme/samples_after_sortme/FL1011_S3_L001_R1_001
_non_rRNA.fastq /scratch/mshumilova/sortme/samples_after_sortme/FL1011_
S3_L001_R2_001_non_rRNA.fastq

 #5 gzip samples after sortmerna

 gzip /scratch/mshumilova/sortme/samples_after_sortme/FL1011_S3_L001_R1
_001_non_rRNA.fastq

 #6 removing of intermediate files

 rm /scratch/mshumilova/gunzip/*
 rm /scratch/mshumilova/sortmerna-2.1/merged/*
 rm /scratch/mshumilova/sortme/rRNA/*
 rm /scratch/mshumilova/sortme/non_rRNA/*

 #1 gunzip R1

 gunzip -c /scratch/mshumilova/samples/FL1076_S24_R1_001.fastq.gz > /sc
ratch/mshumilova/gunzip/FL1076_S24_R1_001.fastq

 #1 gunzip R2

 gunzip -c /scratch/mshumilova/samples/FL1076_S24_R2_001.fastq.gz > /sc
ratch/mshumilova/gunzip/FL1076_S24_R2_001.fastq

 #2 merge R1 and R2

 /scratch/mshumilova/sortmerna-2.1/scripts/merge-paired-reads.sh /scrat
ch/mshumilova/gunzip/FL1076_S24_R1_001.fastq /scratch/mshumilova/gunzip
/FL1076_S24_R2_001.fastq /scratch/mshumilova/sortmerna-2.1/merged/FL107
6_S24_merged_001.fastq

 #3 sort
```

Thank you for your attention. Have a great day ahead!

I've solved my "space" problem)