Khan Inan
Transcriptomics 7653
Homework #1

Q1.1a. Which directory should you write outputs that require large amounts of disk space (e.g., BAM files from whole genome sequencing) and will be the primary location where you will write large NGS files generated on compute nodes in this course? Select the single best answer.
(a) /archive
(b) /home
(c) /scratch

Scratch seems to have the most terabytes of storage, so that would be the best option for storing large NGS files

Q1.1b. Which of these directories is backed up and can be recovered should the data be lost? Select all that apply.
(a) /archive
(b) /home
(c) /scratch

Archive is the directory that is backup for recovery

Q1.1c. Which of these directories is flushed every 60 days? Indicate all that apply.
(a) /scratch
(b) /archive
(c) /home

It seems that Scratch is flushed every 60 days because the default choice for it is No compared to the other directories

Q1.1d Execute the "myquota" command to determine how much disk space you have available in each directory. Note that if you are working from a compute node prompt, the /archive directory will not appear because /archive is not mounted on compute nodes. How much space do you have remaining on each of your /home and /scratch and directories?

I seem to have 50.0 GB on /home and 5.0TB on /scratch

Q2.1 Review the SAM file above and answer the following [1 point].

Q2.1a What is the first word of the second alignment record?

The first word of the second alignment record is SQ^ISN

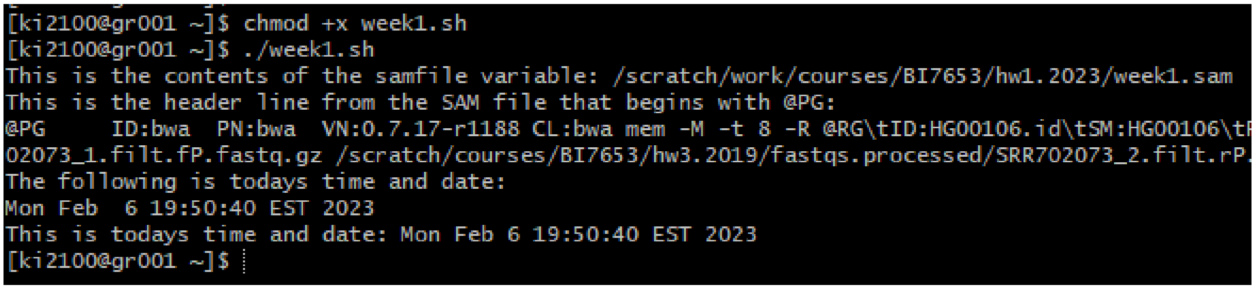
Q2.1b What is the delimiter between columns of an alignment record (row) (hint: your answer should not be ^I, You may need to use online resources to answer the question)?

From what I understand the “tab” key is being used to separate the columns

Q2.1c The $ at the end of each line is also a hidden character. What does $ represent?

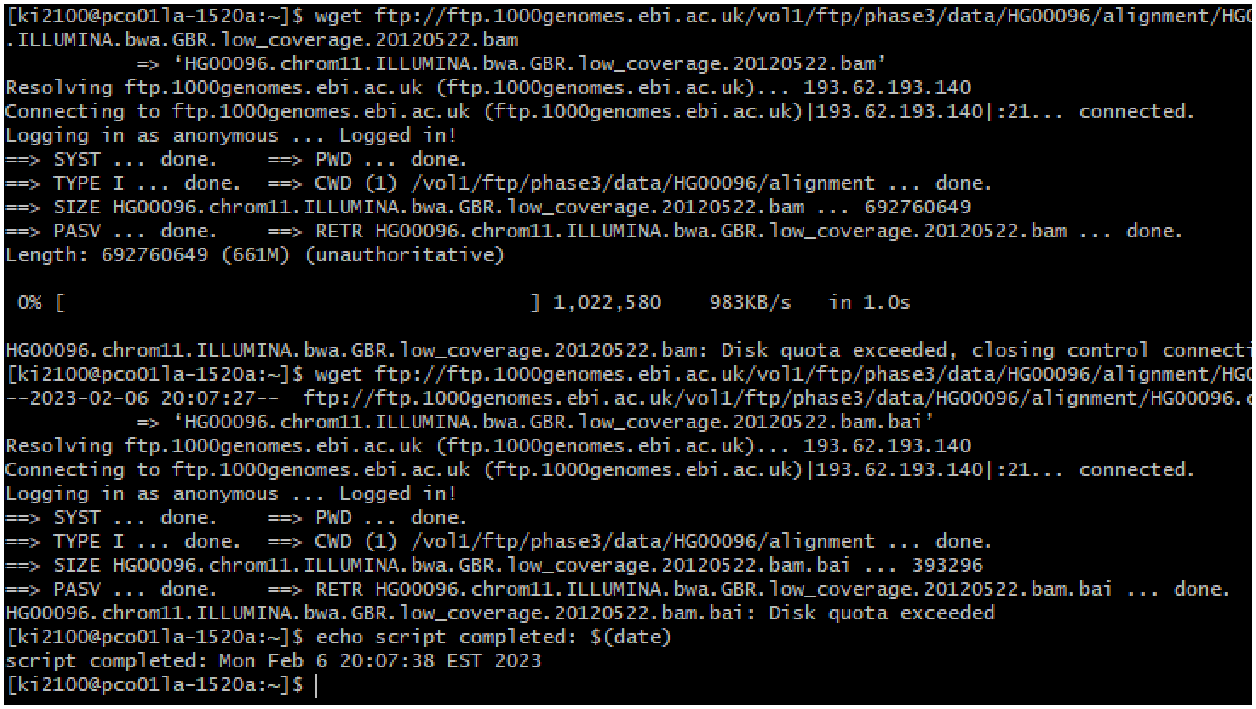
The $ is simply a key to exit the previous command

Q2.2 Execute week1.sh using your preferred method and copy both the command and output into you answers file [1 point].



```
[ki2100@gr001 ~]$ chmod +x week1.sh
[ki2100@gr001 ~]$ ./week1.sh
This is the contents of the samfile variable: /scratch/work/courses/BI7653/hw1.2023/week1.sam
This is the header line from the SAM file that begins with @PG:
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem -M -t 8 -R @RG\tID:HG00106.id\tSM:HG00106\t
02073_1.filt.fP.fastq.gz /scratch/courses/BI7653/hw3.2019/fastqs.processed/SRR702073_2.filt.rP.
The following is todays time and date:
Mon Feb  6 19:50:40 EST 2023
This is todays time and date: Mon Feb 6 19:50:40 EST 2023
[ki2100@gr001 ~]$
```

Q3.1 Now you will draft and execute a slurm job [1 point].

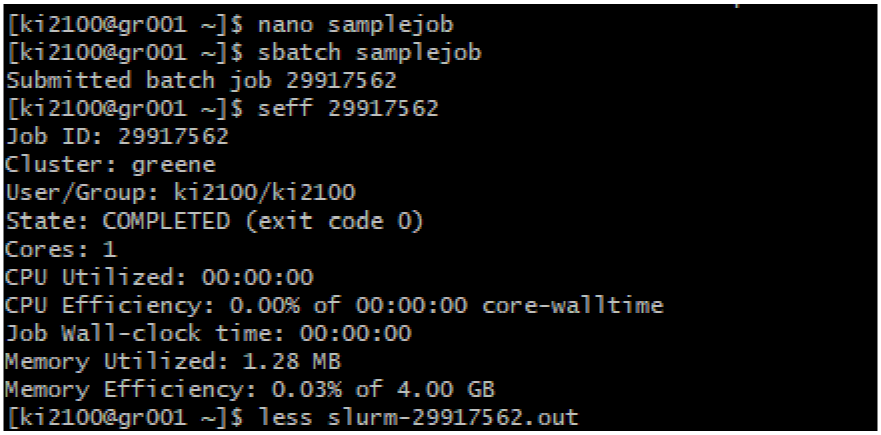


```
[ki2100@pco01la-1520a:~]$ wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00096/alignment/HG0
.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
           => ‘HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam’
Resolving ftp.1000genomes.ebi.ac.uk (ftp.1000genomes.ebi.ac.uk)... 193.62.193.140
Connecting to ftp.1000genomes.ebi.ac.uk (ftp.1000genomes.ebi.ac.uk)|193.62.193.140|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /vol1/ftp/phase3/data/HG00096/alignment ... done.
==> SIZE HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam ... 692760649
==> PASV ... done.    ==> RETR HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam ... done.
Length: 692760649 (661M) (unauthoritative)

 0% [                                       ] 1,022,580    983KB/s   in 1.0s

HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam: Disk quota exceeded, closing control connecti
[ki2100@pco01la-1520a:~]$ wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00096/alignment/HG0
--2023-02-06 20:07:27--  ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00096/alignment/HG00096.c
           => ‘HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam.bai’
Resolving ftp.1000genomes.ebi.ac.uk (ftp.1000genomes.ebi.ac.uk)... 193.62.193.140
Connecting to ftp.1000genomes.ebi.ac.uk (ftp.1000genomes.ebi.ac.uk)|193.62.193.140|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /vol1/ftp/phase3/data/HG00096/alignment ... done.
==> SIZE HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam.bai ... 393296
==> PASV ... done.    ==> RETR HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam.bai ... done.
HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam.bai: Disk quota exceeded
[ki2100@pco01la-1520a:~]$ echo script completed: $(date)
script completed: Mon Feb 6 20:07:38 EST 2023
[ki2100@pco01la-1520a:~]$ |
```

Q3.1a For your answer, report the commands you used for steps 1-5.

1. <u>**I used cp**</u>
2. <u>**I did `/scratch/work/courses/BI7653/hw1.2023/slurm_template.sh`**</u>
3. <u>**I used mv**</u>
4. <u>**I used the given code**</u>
5. <u>**I used the given script**</u>

Q3.1b Report the contents of your job submission script.



```
[ki2100@gr001 ~]$ nano samplejob
[ki2100@gr001 ~]$ sbatch samplejob
Submitted batch job 29917562
[ki2100@gr001 ~]$ seff 29917562
Job ID: 29917562
Cluster: greene
User/Group: ki2100/ki2100
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:00:00
CPU Efficiency: 0.00% of 00:00:00 core-walltime
Job Wall-clock time: 00:00:00
Memory Utilized: 1.28 MB
Memory Efficiency: 0.03% of 4.00 GB
[ki2100@gr001 ~]$ less slurm-29917562.out
```

Q3.3 Now answer the following questions [1 point].

Q3.3a What is the job id of your job?

My job ID is 29917562

Q3.3b What are the names of ALL the files in the directory where you launched the job after the job has completed?

The names of all the files are

HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam

HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam.bai

Slurm-29917562.out

slurmjob_templates.sh

Q3.3c What is the exit status of your job?

The exit status was 0

Q3.3d How much memory (RAM) was utilized? To see execute seff <job id>

1.28 mb of memory was utilized

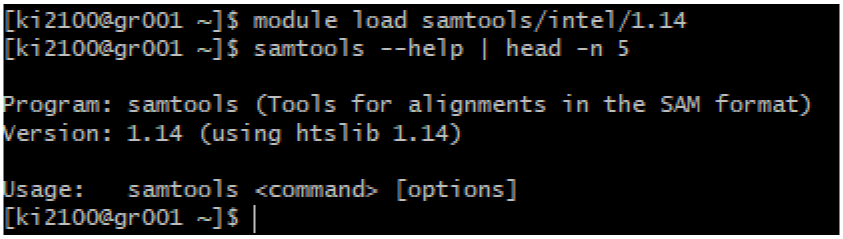
Q3.4 Answer the following [1 point]. Q3.4a What is the name of the file(s) with the STDERR and STDOUT for your job?

slurm-29917562.out

Q3.4b What is the output of the "date" command at the time of completion of your job? Check the the STDERR/STDOUT file for your job.

Mon Feb 6 20:24:09 EST 2023

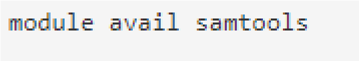
Q4.1. Perform the following steps and save commands and output for your answer using the pre-recorded video (and powerpoint) for help



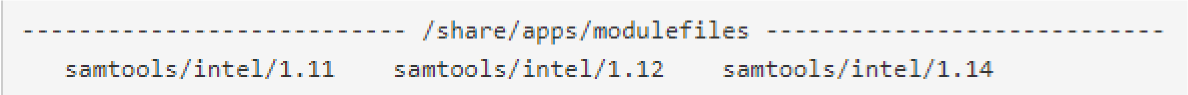
```
[ki2100@gr001 ~]$ module load samtools/intel/1.14
[ki2100@gr001 ~]$ samtools --help | head -n 5

Program: samtools (Tools for alignments in the SAM format)
Version: 1.14 (using htslib 1.14)

Usage:   samtools <command> [options]
[ki2100@gr001 ~]$ |
```

Report for your answer all command lines and output returned to the terminal from Q4.1 [1 point].

Part 1

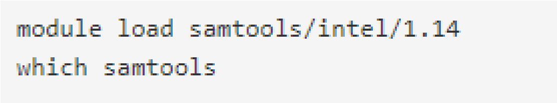


```
module avail samtools
```

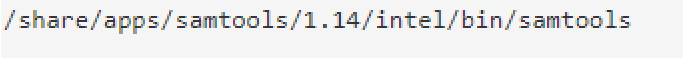


```
--------------------------- /share/apps/modulefiles ----------------------------
   samtools/intel/1.11    samtools/intel/1.12    samtools/intel/1.14
```

Part 2

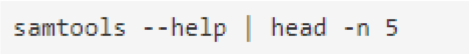


```
module load samtools/intel/1.14
which samtools
```

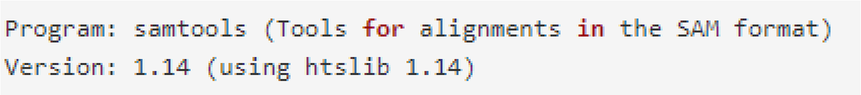


```
/share/apps/samtools/1.14/intel/bin/samtools
```

Part 3

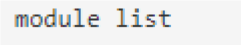


```
samtools --help | head -n 5
```

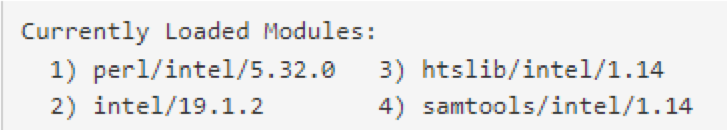


```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.14 (using htslib 1.14)
```

Part 4

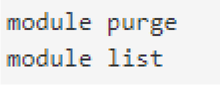


```
module list
```

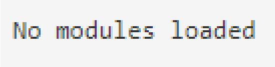


```
Currently Loaded Modules:
  1) perl/intel/5.32.0   3) htslib/intel/1.14
  2) intel/19.1.2        4) samtools/intel/1.14
```

Part 5

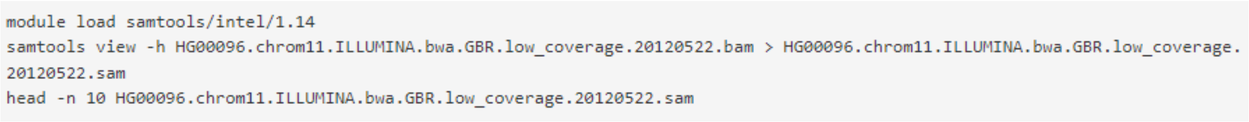


```
module purge
module list
```

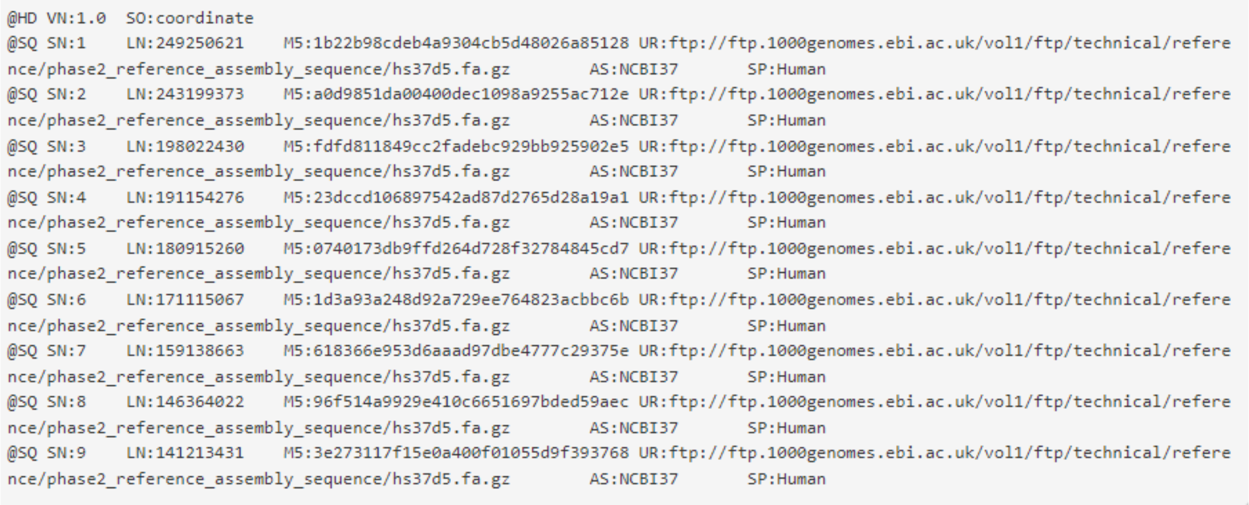


```
No modules loaded
```

For your Q4.2 answer, report the first 10 lines of the sam file (e.g., head -n 10). If you submit via a markdown document (e.g., Rmarkdown), please include such text in a code block for readability [1 point]



```
module load samtools/intel/1.14
samtools view -h HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.bam > HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.
20120522.sam
head -n 10 HG00096.chrom11.ILLUMINA.bwa.GBR.low_coverage.20120522.sam
```



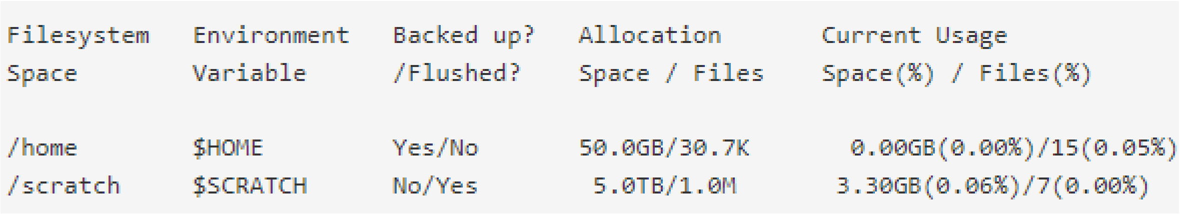
```
@HD VN:1.0  SO:coordinate
@SQ SN:1    LN:249250621    M5:1b22b98cdeb4a9304cb5d48026a85128 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:2    LN:243199373    M5:a0d9851da00400dec1098a9255ac712e UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:3    LN:198022430    M5:fdfd811849cc2fadebc929bb925902e5 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:4    LN:191154276    M5:23dccd106897542ad87d2765d28a19a1 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:5    LN:180915260    M5:0740173db9ffd264d728f32784845cd7 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:6    LN:171115067    M5:1d3a93a248d92a729ee764823acbbc6b UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:7    LN:159138663    M5:618366e953d6aaad97dbe4777c29375e UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:8    LN:146364022    M5:96f514a9929e410c6651697bded59aec UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
@SQ SN:9    LN:141213431    M5:3e273117f15e0a400f01055d9f393768 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz        AS:NCBI37       SP:Human
```

Q5.1. What is the size of the SAM file in human readable bytes? [1 point].

It is about 2.7 gb

Q5.2 How did your /scratch quota change relative to your myquota command from Task 1? [1 point].

The memory usage increased compared to task 1 since it has now become 3.30gb which is more than before



```
Filesystem    Environment    Backed up?    Allocation        Current Usage
Space         Variable       /Flushed?     Space / Files     Space(%) / Files(%)

/home         $HOME          Yes/No        50.0GB/30.7K        0.00GB(0.00%)/15(0.05%)
/scratch      $SCRATCH       No/Yes         5.0TB/1.0M        3.30GB(0.06%)/7(0.00%)
```