

****I have reduced page margin size in order to make room for code/scripts**

Q1.1. Please report the contents of your job script [1 point].

Shown below:

```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=5:00:00
#SBATCH --mem=32GB
#SBATCH --job-name=indexfile
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=bl2477@nyu.edu

module load samtools/intel/1.14
module load bwa/intel/0.7.17

samtools faidx Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa
bwa index -a bwtsw Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa
```

Q1.2. Upon job completion, please execute `ls -al` in your hg38 directory and report the output [1 point].

Output:

```
total 8355586
drwxrwsr-x. 2 bl2477 bl2477      4096 Feb 14 01:23 .
drwxrwsr-x. 3 bl2477 bl2477      4096 Feb 14 00:11 ..
-rw-rwx---. 1 bl2477 bl2477 3130750435 Feb 14 00:12 Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa
-rw-rw-r--. 1 bl2477 bl2477    18172 Feb 14 01:02 Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa.amb
-rw-rw-r--. 1 bl2477 bl2477     7418 Feb 14 01:02 Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa.ann
-rw-rw-r--. 1 bl2477 bl2477 3099750792 Feb 14 01:01 Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa.bwt
-rw-rw-r--. 1 bl2477 bl2477     6793 Feb 14 00:19 Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa.fai
-rw-rw-r--. 1 bl2477 bl2477 774937681 Feb 14 01:02 Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa.pac
-rw-rw-r--. 1 bl2477 bl2477 1549875408 Feb 14 01:23 Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa.sa
-rw-rwx---. 1 bl2477 bl2477      433 Feb 14 00:19 indexfile_slurm.sh
-rw-rw-r--. 1 bl2477 bl2477     6108 Feb 14 01:23 slurm-14893857.out
```

Q2.1 Now either take a screen shot showing your squeue command and the output (or copy the output to your homework report) [1 point].

Output:

14900910_1	cs	bwamem_a	bl2477	R	1:03	1	cs059
14900910_2	cs	bwamem_a	bl2477	R	1:03	1	cs071
14900910_3	cs	bwamem_a	bl2477	R	1:03	1	cs073
14900910_4	cs	bwamem_a	bl2477	R	1:03	1	cs075
14900910_5	cs	bwamem_a	bl2477	R	1:03	1	cs122
14900910_6	cs	bwamem_a	bl2477	R	1:03	1	cs142
14900910_7	cs	bwamem_a	bl2477	R	1:03	1	cs144
14900910_8	cs	bwamem_a	bl2477	R	1:03	1	cs144
14900910_9	cs	bwamem_a	bl2477	R	1:03	1	cs144
14900910_10	cs	bwamem_a	bl2477	R	1:03	1	cs193
14900910_11	cs	bwamem_a	bl2477	R	1:03	1	cs193
14900910_12	cs	bwamem_a	bl2477	R	1:03	1	cs193
14900910_13	cs	bwamem_a	bl2477	R	1:03	1	cs196
14900910_14	cs	bwamem_a	bl2477	R	1:03	1	cs228
14900910_15	cs	bwamem_a	bl2477	R	1:03	1	cs228
14900910_16	cs	bwamem_a	bl2477	R	1:03	1	cs228
14900910_17	cs	bwamem_a	bl2477	R	1:03	1	cs228
14900910_18	cs	bwamem_a	bl2477	R	1:03	1	cs240
14900910_19	cs	bwamem_a	bl2477	R	1:03	1	cs258
14900910_20	cs	bwamem_a	bl2477	R	1:03	1	cs281
14900910_21	cs	bwamem_a	bl2477	R	1:03	1	cs301
14900910_22	cs	bwamem_a	bl2477	R	1:03	1	cs301
14900910_23	cs	bwamem_a	bl2477	R	1:03	1	cs308
14900910_24	cs	bwamem_a	bl2477	R	1:03	1	cs308
14900910_25	cs	bwamem_a	bl2477	R	1:03	1	cs334
14900910_26	cs	bwamem_a	bl2477	R	1:03	1	cs334
14900910_27	cs	bwamem_a	bl2477	R	1:03	1	cs334
14900910_28	cs	bwamem_a	bl2477	R	1:03	1	cs361
14900910_29	cs	bwamem_a	bl2477	R	1:03	1	cs368
14900910_30	cs	bwamem_a	bl2477	R	1:03	1	cs368

Q2.2. Please report your grep command and find commands and there outputs in your report. How many .sam files were produced? What do the exit statuses of the 30 subjobs indicate? [1 point].

grep ESTATUS slurm-*.out

```
slurm-14900910_10.out:_ESTATUS_ [ bwa mem for HG00149 ]: 0
slurm-14900910_11.out:_ESTATUS_ [ bwa mem for HG00260 ]: 0
slurm-14900910_12.out:_ESTATUS_ [ bwa mem for NA18907 ]: 0
slurm-14900910_13.out:_ESTATUS_ [ bwa mem for NA19137 ]: 0
slurm-14900910_14.out:_ESTATUS_ [ bwa mem for NA19093 ]: 0
slurm-14900910_15.out:_ESTATUS_ [ bwa mem for NA19256 ]: 0
slurm-14900910_16.out:_ESTATUS_ [ bwa mem for NA19098 ]: 0
slurm-14900910_17.out:_ESTATUS_ [ bwa mem for NA18870 ]: 0
slurm-14900910_18.out:_ESTATUS_ [ bwa mem for NA18909 ]: 0
slurm-14900910_19.out:_ESTATUS_ [ bwa mem for NA19138 ]: 0
slurm-14900910_1.out:_ESTATUS_ [ bwa mem for NA18757 ]: 0
slurm-14900910_20.out:_ESTATUS_ [ bwa mem for HG00151 ]: 0
slurm-14900910_21.out:_ESTATUS_ [ bwa mem for HG00106 ]: 0
slurm-14900910_22.out:_ESTATUS_ [ bwa mem for HG01914 ]: 0
slurm-14900910_23.out:_ESTATUS_ [ bwa mem for HG01985 ]: 0
slurm-14900910_24.out:_ESTATUS_ [ bwa mem for HG01986 ]: 0
slurm-14900910_25.out:_ESTATUS_ [ bwa mem for HG02013 ]: 0
slurm-14900910_26.out:_ESTATUS_ [ bwa mem for HG02051 ]: 0
slurm-14900910_27.out:_ESTATUS_ [ bwa mem for HG01879 ]: 0
slurm-14900910_28.out:_ESTATUS_ [ bwa mem for HG01880 ]: 0
slurm-14900910_29.out:_ESTATUS_ [ bwa mem for HG01896 ]: 0
slurm-14900910_2.out:_ESTATUS_ [ bwa mem for NA18627 ]: 0
slurm-14900910_30.out:_ESTATUS_ [ bwa mem for HG01915 ]: 0
slurm-14900910_3.out:_ESTATUS_ [ bwa mem for NA18591 ]: 0
slurm-14900910_4.out:_ESTATUS_ [ bwa mem for NA18566 ]: 0
slurm-14900910_5.out:_ESTATUS_ [ bwa mem for NA18644 ]: 0
slurm-14900910_6.out:_ESTATUS_ [ bwa mem for NA18545 ]: 0
slurm-14900910_7.out:_ESTATUS_ [ bwa mem for HG00113 ]: 0
slurm-14900910_8.out:_ESTATUS_ [ bwa mem for HG00243 ]: 0
slurm-14900910_9.out:_ESTATUS_ [ bwa mem for HG00132 ]: 0
```

find . -name *sam

```
./NA18591/NA18591.sam
./HG00260/HG00260.sam
./NA18909/NA18909.sam
./HG01985/HG01985.sam
./HG01879/HG01879.sam
./HG01915/HG01915.sam
./NA18627/NA18627.sam
./HG00106/HG00106.sam
./NA18644/NA18644.sam
./NA19093/NA19093.sam
./HG01880/HG01880.sam
./HG01986/HG01986.sam
./NA18907/NA18907.sam
./HG02013/HG02013.sam
./HG00149/HG00149.sam
./NA18757/NA18757.sam
./NA19098/NA19098.sam
./NA18870/NA18870.sam
./HG02051/HG02051.sam
./HG00243/HG00243.sam
./NA19137/NA19137.sam
./HG00151/HG00151.sam
./HG00132/HG00132.sam
./NA18545/NA18545.sam
./HG00113/HG00113.sam
./HG01896/HG01896.sam
./NA19256/NA19256.sam
./NA19138/NA19138.sam
./HG01914/HG01914.sam
./NA18566/NA18566.sam
```

So looking at the results, 30 sam files were produced and the exit statuses of the 30 subjobs all seem to be 0 which means they ran successfully

Q3.1. Review the samtools view documentation. Then, use this program to extract only the header from the bam file above and answer the following [1 point].

Full output shown on next 3 pages below:

```

@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 M5:23dcdd106897542ad87d2765d28a19a1 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 M5:96f514a9929e410c6651697bde59aec 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
@SQ SN:10 LN:135534747 M5:988c28e000e84c26d552359af1ea2e1d 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:11 LN:135006516 M5:98c59049a2df285c76ffb1c6db8f8b96 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:12 LN:133851895 M5:51851ac0e1a115847ad36449b0015864 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:13 LN:115169878 M5:283f8d7892baa81b510a015719ca7b0b 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:14 LN:107349540 M5:98f3cae32b2a2e9524bc19813927542e 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:15 LN:102531392 M5:e5645a794a8238215b2cd77acb95a078 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:16 LN:90354753 M5:fc9b1a7b42b97a864f56b348b06095e6 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:17 LN:81195210 M5:351f64d4f4f9ddd45b35336ad97aa6de UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:18 LN:78077248 M5:b15d4b2d29dde9d3e4f93d1d0f2cbc9c UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:19 LN:59128983 M5:1aacd71f30db8e561810913e0b72636d UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:20 LN:63025520 M5:0dec9660ec1efaa33281c0d5ea2560f UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:21 LN:48129895 M5:2979a6085bfe28e3ad6f552f361ed74d UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:22 LN:51304566 M5:a718acaa6135fdca8357d5bfe94211dd UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:MT LN:16569 M5:c68f52674c9fb33aef52dcf399755519 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000207.1 LN:4262 M5:f3814841f1939d3ca19072d9e89f3fd7 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:GL000226.1 LN:15008 M5:1c1b2cd1fccbc0a99b6a447fa24d1504 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000229.1 LN:19913 M5:d0f40ec87de311d8e715b52e4c7062e1 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000231.1 LN:27386 M5:ba8882ce3a1efa2080e5d29b956568a4 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000210.1 LN:27682 M5:851106a74238044126131ce2a8e5847c UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000239.1 LN:33824 M5:99795f15702caec4fa1c4e15f8a29c07 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000235.1 LN:34474 M5:118a25ca210cfbcdfb6c2ebb249f9680 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000201.1 LN:36148 M5:dfb7e7ec60ffdc85cb359ea28454ee9 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000247.1 LN:36422 M5:7de00226bb7df1c57276ca6baabaf15 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human

```



```

@SQ SN:GL000245.1 LN:36651 M5:89bc61960f37d94abf0df2d481ada0ec UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000197.1 LN:37175 M5:6f5efdd36643a9b8c8ccad6f2f1edc7b UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000203.1 LN:37498 M5:96358c325fe0e70bee73436e8bb14dbd UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000246.1 LN:38154 M5:e4afcd31912af9d9c2546acf1cb23af2 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000249.1 LN:38502 M5:1d78abec37c15fe29a275eb08d5af236 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000196.1 LN:38914 M5:d92206d1bb4c3b4019c43c0875c06dc0 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000248.1 LN:39786 M5:5a8e43bec9be36c7b49c84d585107776 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000244.1 LN:39929 M5:0996b4475f353ca98bacb756ac479140 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000238.1 LN:39939 M5:131b1efc3270cc838686b54e7c34b17b UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000202.1 LN:40103 M5:06cbf126247d89664a4faebad130fe9c UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000234.1 LN:40531 M5:93f998536b61a56fd0ff47322a911d4b UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000232.1 LN:40652 M5:3e06b6741061ad93a8587531307057d8 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000206.1 LN:41001 M5:43f69e423533e948bfcae5ce1d45bd3f1 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000240.1 LN:41933 M5:445a86173da9f237d7bcf41c6cb8cc62 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000236.1 LN:41934 M5:fdcd739913efa1fdc64b6c0cd7016779 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000241.1 LN:42152 M5:ef4258cdc5a45c206cea8fc3e1d858cf UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000243.1 LN:43341 M5:cc34279a7e353136741c9fce79bc4396 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000242.1 LN:43523 M5:2f8694fc47576bc81b5fe9e7de0ba49e UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000230.1 LN:43691 M5:b4eb71ee878d3706246b7c1dbef69299 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000237.1 LN:45867 M5:e0c82e7751df73f4f6d0ed30cdc853c0 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000233.1 LN:45941 M5:7fed60298a8d62ff808b74b6ce820001 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000204.1 LN:81310 M5:efc49c871536fa8d79cb0a06fa739722 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human

@SQ SN:GL000198.1 LN:90085 M5:868e7784040da90d900d2d1b667a1383 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000208.1 LN:92689 M5:aa81be49bf3fe63a79bdc6a6f279abf6 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000191.1 LN:106433 M5:d75b436f50a8214ee9c2a51d30b2c2cc UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000227.1 LN:128374 M5:a4aead23f8053f2655e468bcc6ecdceb UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000228.1 LN:129120 M5:c5a17c97e2c1a0b6a9cc5a6b064b714f UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000214.1 LN:137718 M5:46c2032c37f2ed899eb41c0473319a69 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000221.1 LN:155397 M5:3238fb74ea87ae857f9c7508d315babb UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000209.1 LN:159169 M5:f40598e2a5a6b26e84a3775e0d1e2c81 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000218.1 LN:161147 M5:1d708b54644c26c7e01c2dad5426d38c UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000220.1 LN:161802 M5:fc35de963c57bf7648429e6454f1c9db UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human

```

```

@SQ SN:GL000213.1 LN:164239 M5:9d424dccc9886650b58f004080a992a UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000211.1 LN:166566 M5:7daaa45c66b288847b9b32b964e623d3 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000199.1 LN:169874 M5:569af3b73522fab4b40995ae4944e78e UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000217.1 LN:172149 M5:6d243e18dea1945fb7f2517615b8f52e UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000216.1 LN:172294 M5:642a232d91c486ac339263820aef7fe0 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000215.1 LN:172545 M5:5eb3b418480ae67a997957c909375a73 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000205.1 LN:174588 M5:d22441398d99caf673e9afb9a1908ec5 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000219.1 LN:179198 M5:f977edd13bac459cb2ed4a5457dba1b3 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000224.1 LN:179693 M5:d5b2fc04f6b41b212a4198a07f450e20 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000223.1 LN:180455 M5:399dfa03bf32022ab52a846f7ca35b30 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000195.1 LN:182896 M5:5d9ec007868d517e73543b005ba48535 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000212.1 LN:186858 M5:563531689f3dbd691331fd6c5730a88b UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000222.1 LN:186861 M5:6fe9abac455169f50470f5a6b01d0f59 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000200.1 LN:187035 M5:75e4c8d17cd4addf3917d1703cacaf25 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000193.1 LN:189789 M5:dbb6e8ece0b5de29da56601613007c2a UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000194.1 LN:191469 M5:6ac8f815bf8e845bb3031b73f812c012 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000225.1 LN:211173 M5:63945c3e6962f28ffd469719a747e73c UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:GL000192.1 LN:547496 M5:325ba9e808f669dfcee210fdd7b470ac UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:NC_007605 LN:171823 M5:6743bd63b3ff2b5b8985d8933c53290a UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/re
ference/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human
@SQ SN:hs37d5 LN:35477943 M5:5b6a4b3a81a2d3c134b7d14bf6ad39f1 UR:ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/refere
nce/phase2_reference_assembly_sequence/hs37d5.fa.gz AS:NCBI37 SP:Human

@RG ID:SRR062634 LB:2845856850 SM:HG00096 PI:206 CN:WUGSC PL:ILLUMINA DS:SRP001294
@RG ID:SRR062635 LB:2845856850 SM:HG00096 PI:206 CN:WUGSC PL:ILLUMINA DS:SRP001294
@RG ID:SRR062641 LB:2845856850 SM:HG00096 PI:206 CN:WUGSC PL:ILLUMINA DS:SRP001294
@PG ID:bwa_index PN:bwa VN:0.5.9-r16 CL:bwa index -a bwtsv $reference_fasta
@PG ID:bwaaln_fastq PN:bwa PP:bwa_index VN:0.5.9-r16 CL:bwa aln -q 15 -f $sai_file $reference_fasta $fastq_file
@PG ID:bwasam PN:bwa PP:bwaaln_fastq VN:0.5.9-r16 CL:bwa sampe -a 618 -r $rg_line -f $sam_file $reference_fasta $s
ai_file(s) $fastq_file(s)
@PG ID:sam_to_fixed_bam PN:samtools PP:bwasam VN:0.1.17 (r973:277) CL:samtools view -bSu $sam_file | samtools sort -n -
o - samtools_nsort_tmp | samtools fixmate /dev/stdin /dev/stdout | samtools sort -o - samtools_csort_tmp | samtools fillmd -
u - $reference_fasta > $fixed_bam_file
@PG ID:gatk_target_interval_creator PN:GenomeAnalysisTK PP:sam_to_fixed_bam VN:1.2-29-g0acaf2d CL:java $jvm_args -jar Genom
eAnalysisTK.jar -T RealignerTargetCreator -R $reference_fasta -o $intervals_file -known $known_indels_file(s)
@PG ID:bam_realignment_around_known_indels PN:GenomeAnalysisTK PP:gatk_target_interval_creator VN:1.2-29-g0acaf2d CL:java
$jvm_args -jar GenomeAnalysisTK.jar -T IndexRealigner -R $reference_fasta -I $bam_file -o $realigned_bam_file -targetInterva
ls $intervals_file -known $known_indels_file(s) -LOO 0.4 -model KNOWNS_ONLY -compress 0 --disable_bam_indexing
@PG ID:bam_count_covariates PN:GenomeAnalysisTK PP:bam_realignment_around_known_indels VN:1.2-29-g0acaf2d CL:java $jvm_arg
s -jar GenomeAnalysisTK.jar -T CountCovariates -R $reference_fasta -I $bam_file -recalFile $bam_file.recal_data.csv -knownSi
tes $known_sites_file(s) -l INFO -L '1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;X;Y;MT' -cov ReadGroupCovariat
e -cov QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate
@PG ID:bam_recalibrate_quality_scores PN:GenomeAnalysisTK PP:bam_count_covariates VN:1.2-29-g0acaf2d CL:java $jvm_args -j
ar GenomeAnalysisTK.jar -T TableRecalibration -R $reference_fasta -recalFile $bam_file.recal_data.csv -I $bam_file -o $recal
ibrated_bam_file -l INFO -compress 0 --disable_bam_indexing
@PG ID:bam_calculate_bq PN:samtools PP:bam_recalibrate_quality_scores VN:0.1.17 (r973:277) CL:samtools calmd -Erb $bam_
file $reference_fasta > $bq_bam_file
@PG ID:bam_merge PN:picard PP:bam_calculate_bq VN:1.53 CL:java $jvm_args -jar MergeSamFiles.jar INPUT=$bam_file(s) OUTP
UT=$merged_bam VALIDATION_STRINGENCY=SILENT
@PG ID:bam_mark_duplicates PN:picard PP:bam_merge VN:1.53 CL:java $jvm_args -jar MarkDuplicates.jar INPUT=$bam_file OU
TPUT=$markdup_bam_file ASSUME_SORTED=TRUE METRICS_FILE=/dev/null VALIDATION_STRINGENCY=SILENT
@PG ID:bam_merge.1 PN:picard PP:bam_mark_duplicates VN:1.53 CL:java $jvm_args -jar MergeSamFiles.jar INPUT=$bam_file(s)
OUTPUT=$merged_bam VALIDATION_STRINGENCY=SILENT
@PG ID:samtools PN:samtools PP:bam_merge.1 VN:1.14 CL:samtools view -H /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom2
0.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
@CO $known_indels_file(s) = ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.in
dels_mills_devine_hg19_leftAligned_collapsed_double_hit.indels.sites.vcf.gz
@CO $known_indels_file(s) .= ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.1
ow_coverage_vqsr.20101123.indels.sites.vcf.gz
@CO $known_sites_file(s) = ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.dbs
np.build135.snps.sites.vcf.gz

```

Q3.1a. Report your command line.

```
module load samtools/intel/1.14
samtools view -H /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
```

Q3.1b. Report the @HD header tag line. What does the information in this line indicate?

```
@HD VN:1.0 SO:coordinate
```

This header tag essentially indicated whether or not the file is sorted by coordinates, or in ascending order. Since there is a SO tag with a coordinate next to it, this shows that it is indeed sorted.

Q3.2. Use samtools view to answer the following. Review samtools view options -c, -f, and -F.

Please answer the following questions including (1) your command line you used to obtain the answer and (2) the output written to your terminal [1 point].

output:

```
samtools view -c -f 4 /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
```

Q3.2a how many unmapped reads are there in the BAM (hint: use appropriate bitwise flag(s) described in SAM/BAM lecture and documentation?

There seems to be 7247 unmapped reads

Q3.2b How many mapped reads are there in the BAM?

2924253 mapped reads

Q3.2c What is the percentage mapping rate (total mapped reads / total reads in the alignment) for this sample?

```
samtools view -c -F 4 /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
2924253
# This is the total number of mapped reads

samtools view -c /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
2931500
# This is the total number of reads in the alignment
```

I calculated there to be 99.752789% mapping rate

Q3.3. A hypothetical SAM file has alignment records with the bitwise flag values that include 4, 147, 113, 99 on the decimal scale. What are the binary and hexadecimal representations of each of the these values? [1 point].

#	Decimal	Hexadecimal	Binary
#	4	4	100
#	147	93	10010011
#	113	71	1110001
#	99	63	1100011

For Q3.5, use samtools view with appropriate -c, -f, -F options to count the following in the BAM file at /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam. For each answer, provide the number of reads and the command line you used.

Q3.5a How many alignments are primary?

```
samtools view -c -F 256 /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
2931500
```

Q3.5b How many alignments are secondary?

```
samtools view -c -f 256 /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
0
```

Q3.5c How many alignments are supplementary?

```
samtools view -c -f 2048 /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
0
```

Q3.5d What is the number of reads excluding unmapped reads, supplementary reads, secondary reads and PCR duplicates?

```
samtools view -c -F 3332 /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
2885340
```

Q3.6. A common task is to subset a SAM/BAM to include a subset of positions on a chromosome. Use samtools view to subset the BAM from Q3.5 from chromosome 20 position 1 to 2000000 (i.e., 2 Mb), while also retaining the header. Note that to perform this type of operation, the BAM must be coordinate-sorted, (which it is).

Now count the reads in the subsetted BAM.

```
samtools view -h /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam 20:1-2000000
00
#Command line to subset the BAM including the header

samtools view -c /scratch/work/courses/BI7653/hw3.2023/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam 20:1-2000000
00
#Command line to count the reads in the subsetted BAM
95338
#Reads in the subsetted BAM
```