**1.1** Why are there two *fastq* files, one with the \_1.fastq extension and another with \_2.fastq? (Hint: this has to do with the way that reads are made in the NGS technique used here).

There are two reads because of the paired-end reads approach. Each read is sequenced from the right and from the left, resulting in two different fastq files.

**1.2** Phred scores above 30 are generally considered trustworthy. However, 99.9% accuracy across millions of base pair calls still results in thousands of sequencing errors. How can we accurately identify variations in our genome (compared to the refence genome) and distinguish these from simple sequencing errors? (Hint: remember the highly parallelized approach of NGS.)

This is why we want to have high coverage. Usually, we will have many reads that cover the same base pair due to the highly parallelized approach of NGS. If the same variation appears in many different reads, then it is likely that it is not just a sequencing error.

**2.1** What is the purpose of all of the ‘echo’ commands in the script? If you’re unsure, then consider the results of the command in your terminal: echo Hello World >> log.txt

Hint: We’re looking for a more complete answer than “they function as print statements, like in R”. Why would we want to print out these statements?

They function as print statements in BASH. The reason you would want these print statements is for logging purposes. For example, if something is printed, we know that the script got to and executed that part of the code. It can help for us to tell where the error is if something goes wrong.

**2.2** There are multiple **options** specified for the bwa aln commands and bwa sampe command. Check the [BWA manual](https://bio-bwa.sourceforge.net/bwa.shtml) and **describe**: (A quick Google search can often turn up more useful explanations than program documentation.)

1. What is -q option in *bwa aln* for?

* The -q is used to specify the quality threshold for filtering out low-quality reads (any base with a lower quality score than the threshold will not be used in the alignment process)

1. What is -t option in *bwa aln* for?

* The -t is used to specify the number of threads/CPU cores to use for the alignment process

1. What is -P option in *bwa sampe* for?

* The -P is used to indicate that the input files contain data from paired-end sequencing reads

**2.3** Look into SRS003660.align.log with *less* command. About how long did the alignment of each *fastq* file take? How long did the *bwa sampe* process take to complete?

The alignment of each fastq file took around 1 minute 30 seconds. The sampe process took around 16 minutes.

**2.4** List the names the files generated by this run. Which is the largest, and how large is this file? *(-lh* option of *ls* converts bytes to larger metric units).

Files generated:

SRS003660.align.log, SRS003660.bam, SRS003660.sorted.bam, SRS003660\_1.sai, SRS003660\_2.sai, slurm-26083001.out

Largest File: SRS003660.bam (1.3 GB)

**2.5** How many sequences were processed in total from both paired-end reads? How many of these reads were rejected by quality control (QC)?

17052008 total processed reads, 0 reads were rejected by QC

**2.6** What percentage of these reads mapped to the reference genome in your alignment process?

94.16% of reads were mapped to the reference genome.

**2.7** Each short read in our NGS data has a mate that is taken from the same segment of DNA but in the opposite direction. We could reasonably expect, then, that any given read and its pair will map to the same chromosome. How many paired-end reads in total (irrespective of mapQ) did map to different chromosomes? Why might this occur?

78846 reads in total did map to different chromosomes. This may occur due to errors or deletions in reference genome, contamination of DNA before sequencing, etc. Also, some of the reads might have matched with identical parts of another chromosome before matching to the correct one.

**3.1** For example, let’s say we have 33 reads that map to nucleotide position 104,204 of chromosome 6. 23 of these reads place an “A” base at this position, and 10 reads have a “T”.

**a)** Intuitively, would you conclude that the individual has a heterozygous AT genotype? Or would you conclude that this individual has a homozygous AA genotype, and the eight “T” reads are sequencing or alignment errors?

Intuitively, because there are so many fewer Ts than As, I would think that the ten Ts are probably sequencing or alignment errors. However, there would need to be much more information about the coverage, quality score, etc. to really make a definite conclusion.

**b)** How would the knowledge that the sequencing or alignment quality scores on the reads that produced the ten “T” calls are very high affect your confidence in your conclusion?

A high quality score for the ten Ts may mean that the individual actually has a heterozygous AT genotype, and it would decrease my confidence in my previous conclusion. Still, more evidence would be needed to make a definitive conclusion.

**c)** If the individual is truly AT heterozygous, how many of the reads (on average) would you expect to indicate a T base at this position?

We would expect that on average half the reads indicate a T base and the other half indicate an A base. If there are 33 reads covering a base, we would expect 16-17 to indicate a T base, and the other 16-17 to indicate A.

**3.2** Which parameter should you change to limit the genotyping to chromosome 5? Open ***mpileup\_genotyping.sh*** with nano and change this parameter accordingly (change it to chr5). Save your modified script. Hint: consult the ***mpileup*** section of the [*samtools*](https://www.htslib.org/doc/samtools.html) manual

The parameter that needs to be changed is the -r parameter (specifically changed to -r chr5)

**3.3** One crafty student suggests that to save further computational time, we could have limited our **alignment step** to align our reads solely to chromosome 5 rather than the entire genome. Why might this not be a good idea?

This may not be a good idea because it could lead to loss of information (reads originating from other chromosomes will be missed). In addition, if the variant is located outside of chromosome 5, limiting the alignment could result in reduced sensitivity in detecting it.

**3.4** Look into SRS003660.genotype.log with *less*. Which step of this process took the longest to complete?

Creating the raw vcf file took the longest to complete.

**3.5** Output the last 8 lines of your vcf file with the *tail* command. List the mutations from your data (reference nucleotide, variant nucleotide, and chromosomal coordinates) that pass this Phred quality cutoff.

In order chromosomal coordinates, reference nucleotide, variant nucleotide

181114707, C, T

181237105, C, G

181243433, T, G

181259957, C, T

181260211, C, T

181260427, CTCTTCTTCTTCTTC, CTCTTCTTCTTC

**3.6** Approximately how many variants are in our ***.vcf*** file output? Use the *wc* command with the required parameter to give the number of lines (check *man wc*) in the ***.vcf*** file. Header lines start with “#” and are not variants (you can use *grep “#” SRS003660.flt.vcf | wc* with the proper parameter to find the number of header lines).

The required parameter into wc is -l

Number of lines total: 7334

Header Lines: 3395

Number of variants: 3939

**3.7** Approximately how many high-quality variants are in your ***.hq.vcf*** output file?

Number of lines total: 6507

Header Lines: 3395

Number of high quality variants: 3112

**4.1** From this output, how many SNPs and how many indels were present in your ***hq.vcf*** file?

The output reported 2914 SNPs and 198 indels

**4.2** Does ANNOVAR report anything concerning of which we should be mindful with regards to our ***hq.vcf*** input?

The terminal prints out several warnings which we should be mindful of. For example, here is a warning that I saw: WARNING: A total of 515 sequences will be ignored due to lack of correct ORF annotation

**4.3** What are four types (choose any four) of non-exonic functional regions from which we see annotated variants in our results file (this information is in the sixth column)?

Intronic, ncRNA\_intronic, UTR5, intergenic

**4.4** How many of our annotated variants are exonic? (Remember to exclude the header line from the line count!). Why are we mainly interested in exonic variants?

615 of our annotated variants are exonic. We are mainly interested in exonic variants because they are the mutations that will actually code for functional (or non-functional) proteins.

**4.5** Construct another set of commands to create a third file that contains all nonsynonymous exonic variants. Paste the commands you used in your lab text. How many of the exonic variants are nonsynonymous amino acid substitutions? (NOTE: a query for “synonymous” would pick up all lines that contain synonymous, even if that word is nested within a larger word such as nonsynonymous. **Be careful of this fact and always verify that your *grep* search is picking up exclusively those lines you intend!**

Commands:

head -1 SRS003660.chr5.annovar.hg38\_multianno.exonic.txt > SRS003660.chr5.annovar.hg38\_multianno.exonic.nonsynonymous.txt

grep "nonsynonymous" SRS003660.chr5.annovar.hg38\_multianno.exonic.txt >> SRS003660.chr5.annovar.hg38\_multianno.exonic.nonsynonymous.txt

255 nonsynonymous exonic variants

**4.6** How many of the exonic variants are insertions? How many are deletions?

Hint: if you search for ‘insertions”, “deletions” or “frameshift” using grep and the output you get is 0 then open up the output file using Excel and look at the columns that correspond to the nucleotides. Think about **what the corresponding mutation should be when there is a gap '-' in different columns** and you can **count the number of insertions and deletions by counting the number of gaps in each column**.

Number of insertions: 0

Number of deletions: 6

**5.1** There should be a few exonic variant loci you discovered that are associated with a pathology through genome-wide association studies (GWAS) in the ClinVar database. Identify a non-synonymous mutation that is associated with a specified condition. Which gene does this variant occur in, and with what phenotypic trait does it correlate? Hint: search the variant in the [NCBI data base](https://www.ncbi.nlm.nih.gov/).

There is an exonic variant loci rs11949577 which is a non-synonymous mutation. This variant occurs in the gene RHO and is associated with the phenotypic trait Congenital stationary night blindness autosomal dominant and Retinitis pigmentosa.

**5.2** What are the relevant nucleotide and amino acid substitutions for this non-synonymous mutation? (Hint: look at columns D, E, and J. *c.A570T* means that the mutation was from Adenine to Thymine, and *p.R190S* means Arginine (R) to Serine (S). Note that many diseases have more than one mutation.)

The relevant nucleotide substitution is from G (the reference genome) to A (the mutation). Glycine at position 1513 gets substituted to Alanine.

**5.3** This allele has a ***dbSNP*** reference ID, listed in the ***avsnp150*** column. This means that this variant has been discovered previously. Navigate to the NCBI [***dbSNP***](http://www.ncbi.nlm.nih.gov/SNP/) database and search for information on this SNP.

a) What is an allele? (Google would be a good resource!)

- An allele is a variation of the same sequence of nucleotides at the same place on a long DNA molecule

b) What is the reference allele in human population at this site?

- G

c) What is the alternative allele; and what is its allele frequency listed? (Hint: click on the "frequency" tab and check the values listed for the "Global" study).

- A; the frequency is 0.125

**5.4** Navigate to the population diversity section. Are there any population groups that noticeably differ in their genotype frequencies at this position according to this data? What is the range of MAF across these populations? You can find the data in the frequency tab boxed in the screenshot below.

- In Asian people, the frequency is low, and in Africans the frequency was high. The MAF ranges from 0.006 – 0.324.

**5.5** There are many different metrics used by [dbNSFP](http://varianttools.sourceforge.net/Annotation/dbNSFP%20(This%20site%20contains%20the%20information%20on%20the%20columns%20of%20the%20Excel%20file,%20including%20the%20metrics)). In order to predict **the effects of our non-synonymous variants**, such as **SIFT, PolyPhen-2**, etc. Pick **two metrics** to research and briefly summarize what kinds of criteria are considered in these metrics’ scoring schemes. For your chosen variant, describe what these metrics predict about any functional ramifications.

SIFT is a tool that predicts whether an AA substitution in a protein will affect its function based on sequence conservation analysis. It evaluates the evolutionary conservation of AA at a particular position in a protein sequence and predicts whether a substitution is tolerated or deleterious based on the degree of conservation. SIFT scores range from 0 to 1 where 0 is deleterious and 1 is tolerated.

PolyPhen2 uses both sequence based and structure based features to predict whether a variant is likely to be deleterious, possibly damaging, or benign. The scores range from 0 to 1 where 0 is benign and 1 is deleterious. It also provides a probability score that reflects how confident the prediction is.

For chosen gene:

SIFT: 0.864

PolyPhen-2: 0.007

Both predict that the variant is benign.