# COMP 3353 Assignment 3

## Name: Chan Tsai Hor

## UID: 3035376355

## Email: [howardchanth@gmail.com](mailto:howardchanth@gmail.com)

## Question 1

**Command:**

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf

**Output:**

# This file was produced by bcftools stats (1.11+htslib-1.11) and can be plotted using plot-vcfstats.

# The command line was: bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf

#

# Definition of sets:

# ID [2]id [3]tab-separated file names

ID 0 NA12878\_chr21\_GRCh38\_assignment3.vcf

# SN, Summary numbers:

# number of records .. number of data rows in the VCF

# number of no-ALTs .. reference-only sites, ALT is either "." or identical to REF

# number of SNPs .. number of rows with a SNP

# number of MNPs .. number of rows with a MNP, such as CC>TT

# number of indels .. number of rows with an indel

# number of others .. number of rows with other type, for example a symbolic allele or

# a complex substitution, such as ACT>TCGA

# number of multiallelic sites .. number of rows with multiple alternate alleles

# number of multiallelic SNP sites .. number of rows with multiple alternate alleles, all SNPs

#

# Note that rows containing multiple types will be counted multiple times, in each

# counter. For example, a row with a SNP and an indel increments both the SNP and

# the indel counter.

#

# SN [2]id [3]key [4]value

SN 0 number of samples: 1

SN 0 number of records: 12688

SN 0 number of no-ALTs: 0

SN 0 number of SNPs: 12669

SN 0 number of MNPs: 0

SN 0 number of indels: 19

SN 0 number of others: 0

SN 0 number of multiallelic sites: 0

SN 0 number of multiallelic SNP sites: 0

…….

So the total number of variants is 12669 + 19 = **12688**

## Question 2

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf | grep "QUAL" | awk '$3>40 {print}' | awk -F '\t' '{sum += $4} END {print sum}'

5482

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf | grep "QUAL" | awk '$3>40 {print}' | awk -F '\t' '{sum += $7} END {print sum}'

8

There are **5490** variants with QUAL greater than 40

## Question 3

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf | grep "QUAL" | awk -F '\t' '{sum += $5} END {print sum}'

7474

There are **7474** transitions

## Question 4

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf | grep "QUAL" | awk -F '\t' '{sum += $5} END {print sum}'

7474

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf | grep "QUAL" | awk -F '\t' '{sum += $6} END {print sum}'

5194

The number of transition is **7474** and number of transversions is **5194**. The ratio of transitions and transversions is 7474/5194 = 1.44, which is a bit different from the **expected 2.05** for humans.

## Question 5

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf | grep "IDD"

# IDD, InDel distribution:

# IDD [2]id [3]length (deletions negative) [4]number of sites [5]number of genotypes [6]mean VAF

IDD 0 -5 **2**  0 .

IDD 0 -3  **2** 0 .

IDD 0 -2 **1** 0 .

IDD 0 1 **12**  0 .

IDD 0 2  **2** 0 .

Number of deletions = 2 + 2 + 1 = 5

Number of insertion = 1 + 12 + 2 = 14

## Question 6

bcftools stats NA12878\_chr21\_GRCh38\_assignment3.vcf | grep "DP" | awk '$3>=20 {print}' | awk -F '\t' '{sum += $6} END {print sum}'

744

There are **744** SNPs having depth greater than or equal to 20.

## Question 7

Please see the Task7 directory and Task7/Q7.py for more details

## Question 8

Please see the Task8 directory and Task8/Q8.py for more details

## Question 9