# NGS Bioinformatics

# Pathogen variant calling practical assignment

**Module topic: Pathogen variant calling**
**Contact session title: Day1**
**Trainer: Narender, Jon**
**Participant:** <*write your name here*>
**Date:** <*write today's date here*>

## Reference mapping and variant calling: Pathogens

**Introduction**

*All information is found in practical material. We will use this document to answer the questions.*

**Please note**

- **Hand-in information** please upload your completed assignment to the Vula 'Assignments' tab. Take note of the final hand-in date for each assignment, which will be indicated on Vula.

### Session 1: Reference mapping

**Q1.1:** How many read pairs were assigned as duplicates?

<*start typing your answer here*>

**Q1.2:** What proportion of the mapped sequence was marked as duplicates?

<*start typing your answer here*>

**Q1.3:** What is the total number of mapped reads?

<*start typing your answer here*>

**Q1.4:** What is the total number of unmapped reads?

<*start typing your answer here*>

**Q1.5:** What is the total number of mapped and properly paired reads?

<*start typing your answer here*>

**Q1.6:** What is the average insert size?

<*start typing your answer here*>

**Q1.7:**What is the percentage of reads properly paired?

<*start typing your answer here*>

**Session 2: Variant Calling and annotation**

**Q2.1:** At what position is the first variant in the unfiltered vcf file for MD001?

<*start typing your answer here*>

---

**Q2.2:** What does the DP4 value represent?

<*start typing your answer here*>

_____

**Q2.3:** What is the read depth of the variant with an ID = rs6040355 for sample NA00002?

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF   ALT    QUAL  FILTER   INFO                              FORMAT       NA00001        NA00002          NA00003
20     14370    rs6054257  G     A      29    PASS     NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51   1/1:43:5:.,.
20     17330    .          T     A      3     q10      NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3     0/0:41:3
20     1110696  rs6040355  A     G,T    67    PASS     NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2     2/2:35:4
20     1230237  .          T     .      47    PASS     NS=3;DP=13;AA=T                   GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51   0/0:61:2
20     1234567  microsat1  GTC   G,GTCT 50    PASS     NS=3;DP=9;AA=G                    GT:GQ:DP     0/1:35:4        0/2:17:2         1/1:40:3
```

<*start typing your answer here*>

---

**Q2.4:** What is the probability that a variant with a GQ of 23 is not a true variant?

<*start typing your answer here*>

---

**Q2.5:** How many HIGH effect variants were there for sample MD001?

<*start typing your answer here*>

---

**Q2.6:** What was the TS TV ratio? (Look in the MD001_snpEff.csv summary file)

*<start typing your answer here>*

---

**Q2.7:** Are there any other mutations in resistance related genes?

Fill in your answers in the table below

| **isolate** | **gene** | **Drug** | **Mutation** | **position** | **Genotype (R/S)** |
|---|---|---|---|---|---|
| MD001 | *rpoB* | RIFAMPICIN | Ser450X (S450X), Asp435X (D435X) | 761110 | R (D435X) |
| MD001 | *rpsL* | STREPTOMYCIN | | | |
| MD001 | *gyrA* | FLUOROQUINOLONE | | | |
| MD001 | *katG* | ISONIAZID | | | |