

# NGS Bioinformatics

**Module topic:** Pathogen variant calling

**Contact session title:** reference mapping and variant calling

**Trainer:** Narender Kumar and Jon Ambler

**Participant:** <write your name here>

**Date:** <write today's date here>

## Reference mapping and variant calling

### Introduction

In the assignment, you will be mapping the reads to the reference genome. This will be followed by variant calling and filtering using various parameters. The last part of this assignment would be to annotate the variants identified and identify specific variants for resistance.

### Tools used in this session

bwa, samtools, picard, bcftools, snpeff

### 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.

### **Session1: Reference mapping**

**Q1.1:** What is the proportion of read pairs assigned as duplicates?

Ans: **300973/2524511 = ~12%**

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

Ans: **~13.7% or 0.137**

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

Ans: **5280578**

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

Ans: **289914**

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

Ans: **5049022**

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

Ans: **350**

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

Ans: **89.7**

## **Session2: Variant Calling and annotation**

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

Ans: **1849**

---

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

Ans: The number of high-quality ref-forward, ref-reverse, alt-forward, alt-reverse bases

---

**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
```

Ans: 0

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

$10^{-(23/10)} = 0.005$

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

cat MD0001\_SNPs\_filtered\_snpEff.ann.vcf | grep HIGH | wc -l  
11

**Q2.6:** What was the TS TV ratio? (Look in the MD001\_snpEff.csv summary file)

1.568528

**Q2.7:** Are there any other mutations in resistance related genes? (Still to edit)

isolate	gene	Drug	Mutation	position	Genotype (R/S)
MD001	<i>rpoB</i>	RIFAMPICIN	Ser450X (S450X), Asp435X (D435X)	761110	R (D435X)
MD001	<i>rpsL</i>	STREPTOMYCIN	Lys43Arg (K43R)	781687	R (K43R)
MD001	<i>gyrA</i>	FLUOROQUINO LONE	Asp94X (D94X), Ala90Val (A90V)	-	S

MD001	<i>katG</i>	ISONIAZID	Ser315Thr	-	S
-------	-------------	-----------	-----------	---	---