Variant Calling Workflow

Answers to questions

1. What does "SO:coordinate" in the "@HD" tag on the first line of the bam file mean?

SO stand for “sort order”

Coordinate means that the reads in the bam file are sorted in ascending order by sequence name (i.e. chromosome) and position.

2. What does "SN:2" and "LN:243199373" in the "@SQ tag mean?

SN:2 means that sequence name is “2”. We have selected chromosome 2 as reference because the data is selected on chromosome 2.

LN:243199373 means that the length of the reference sequence is 243199373 bp. This is the length of chromosome 2.

3. What is encoded in the @RG tag?

Information about read groups.

4. What is the leftmost mapping position of the first read in the bamfile?

Chromosome 2, position 3843448

5. What is the read length?

101 bp

6. How can you estimate the coverage in IGV?

Count the number of reads that align at a position.

7. Is the coverage evenly distributed across the genome?

No

8. What column of the VCF file contains genotype information for the sample HG00097?

The 10th column with header “HG00097”

9. What does *GT* in the *FORMAT* column of the data lines mean?

Genotype

10. What genotype does the sample HG00097 have at position 2: 136234279?

0/1 which means one copy of the reference allele and one copy of the alternative allele.

11. What does *AD* in the *FORMAT* column of the data lines mean?

Number of reads that match the reference allele and the alternative alleles, respectively.

12. What is the allelic depths for the reference and alternative alles for sample HG00097 at position 2: 136234279?

3 reads match the reference allele and 4 reads match the alternative allele.

13. How many genetic variants was detected in the sample? The linux command `grep -v "#" filename.txt | wc ` extracts all lines in HG00097.vcf that *don't* start with "#", and counts these lines.

206 variants