Variant Calling Workflow
Answers to questions

1. **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.

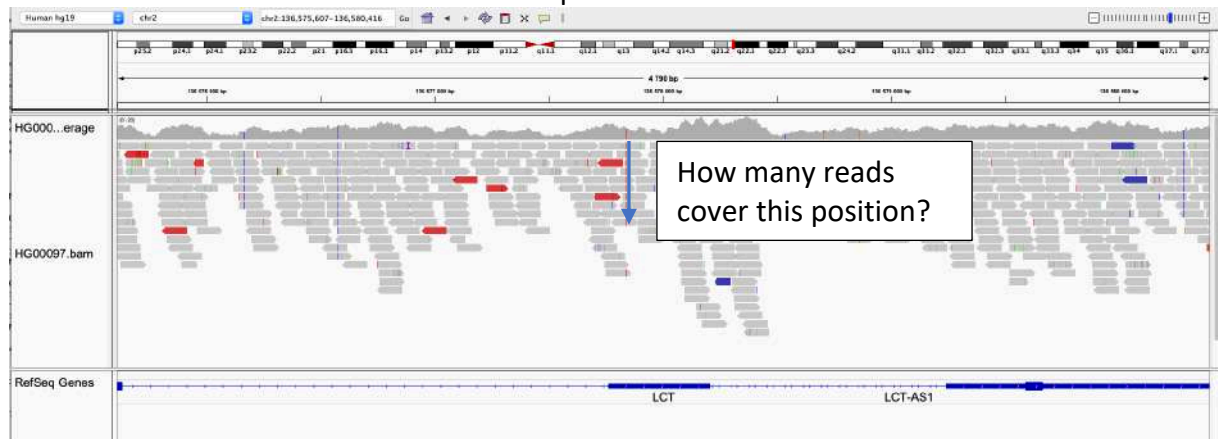2. **What is the leftmost mapping position of the first read in the bamfile?**

Chromosome 2, position 3843448

3. **What is the read length?**

101bp

4. **Approximately how many reads cover an arbitrary position in the genomic region we are looking at?**

Count the number of reads that cover the position. This varies but is often around 10.

5. **Which genes are located within the region chr2:136545000-136617000?**

LCT, LCT-AS1 (also called NR_045486) and MCM6

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

The 10th column with header "HG00097"

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
	INFO	FORMAT	HG00097			

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

Genotype

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

8. What genotype does the sample HG00097 have at position 2: 136545844?

1/1
This individual has the alternative allele on both copies of chromosome 2.

9. 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.

```
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic
depths for the ref and alt alleles in the order listed">
```

10. What are the allelic depths for the reference and alternative alleles in sample HG00097 at position 2: 136545844?

0 reads match the reference allele and 11 reads match the alternative allele.

```
2            136545844            .            C            G            427.02
             .
             AC=2;AF=1.00;AN=2;DP=11;ExcessHet=3.0103;FS=0.000;M
LEAC=2;MLEAF=1.00;MQ=60.00;QD=34.86;SOR=1.270
             GT:AD:DP:GQ:PL    1/1:0,11:11:33:441,33,0
```

11. How many genetic variants was detected in HG00097?
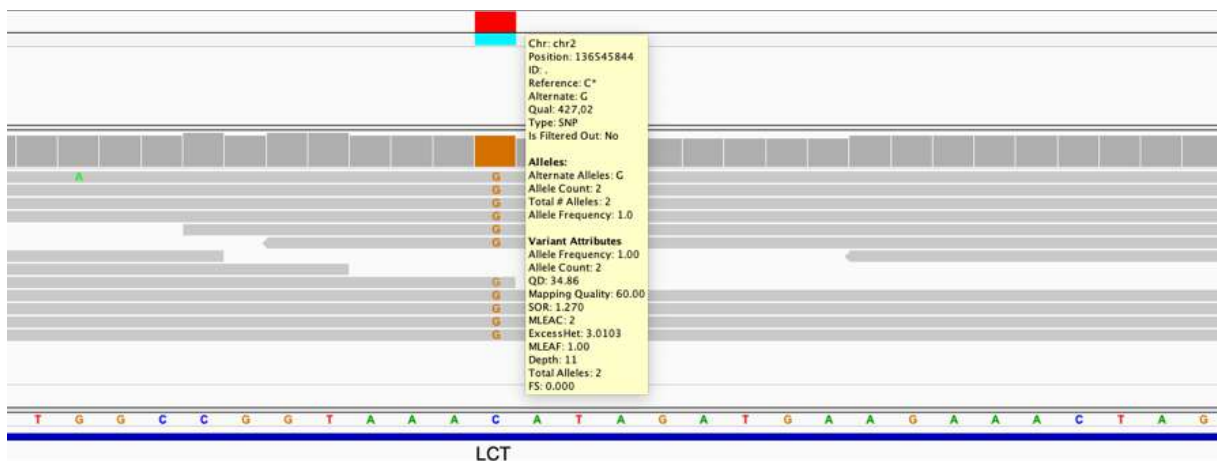The linux command
grep -v "#" HG00097.vcf | wc -l
extracts all lines in that don't start with "#", and then counts these lines.
206 variants

12. Hoover the mouse over the upper row of the vcf track. What is the reference and alternative alleles of the variant at position 2:136545844?

Referenec allele = C
Alternative allele = G

13. Hoover the mouse over the lower row of the vcf track and look under "Genotype Information". What genotype does HG00097 have at position 2:136545844? Is this the same as you found by looking directly in the vcf file in question 10?
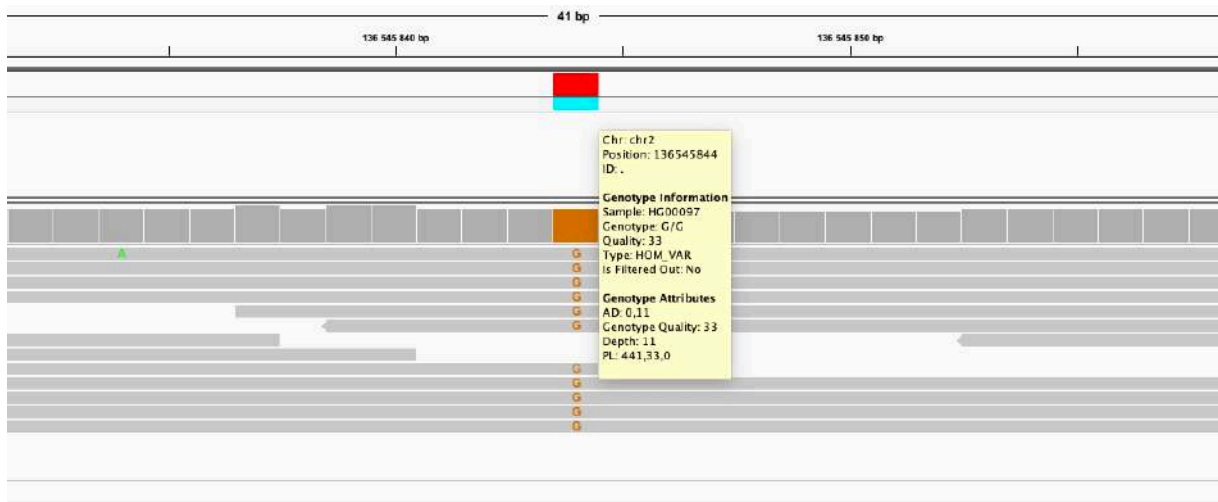
Genotype = G/G
Yes, this is the same genotype as can be seen directly in the vcf file, but in the vcf file it is encoded as 1/1 which means two copies of the alternative allele.

14. Look in the bam track and count the number of reads that have "G" and "C", respectively, at position 2:136545844. How is this information captured under "Genotype Attributes"? (Hoover the mouse over the lower row of the vcf track to find the "Genotype Attributes")

0 reads have "C" which is the reference allele, 11 reads have "G" which is the alternative allele for this variant. This information is captured as "AD=0,11" under Genotype Attributes.

15. How many data lines do the cohort.g.vcf file have? You can use the linux command `grep -v "#" cohort.g.vcf` to extract all lines in "cohort.g.vcf" that don't start with "#", then `|`, and then `wc -l` to count those lines.

```
grep -v "^#" cohort.g.vcf | wc -l
```
This returns 313376 lines (or similar)

16. How many data lines do the cohort.vcf file have?

```
grep -v "^#" cohort.vcf | wc -l
```
This returns 718 lines (or similar)

17. Explain the difference in number of data lines?
Cohort.g.vcf contains information about every position in the analyzed region (although some positions are merged into blocks), cohort.vcf contain information about sites where genetic variants was detected.

18. Look at the header line of the cohort.vcf file. What columns does it have?
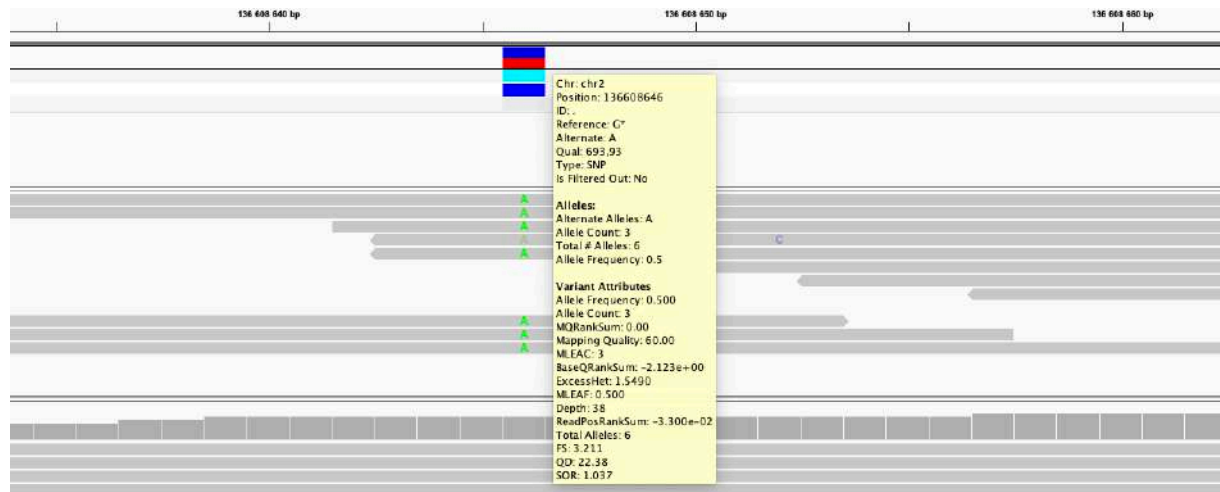
```
grep "#CHROM" cohort.vcf
```

gives:

```
#CHROM  POS ID  REF ALT QUAL     FILTER  INFO     FORMAT
    HG00097 HG00100 HG00101
```

19. What is encoded in the last three columns of the data lines?

Genotypes and genotype attributes of the samples HG00097, HG00100 and HG00101

20. What is the reference and alternative alleles at chr2:136608646?

Reference allele = G
Alternative allele = A

21. What genotype do the three samples have at chr2:136608646? Note how genotypes are color coded in IGV (it is possible to modify the color coding but lest stick with the default settings in this lab).
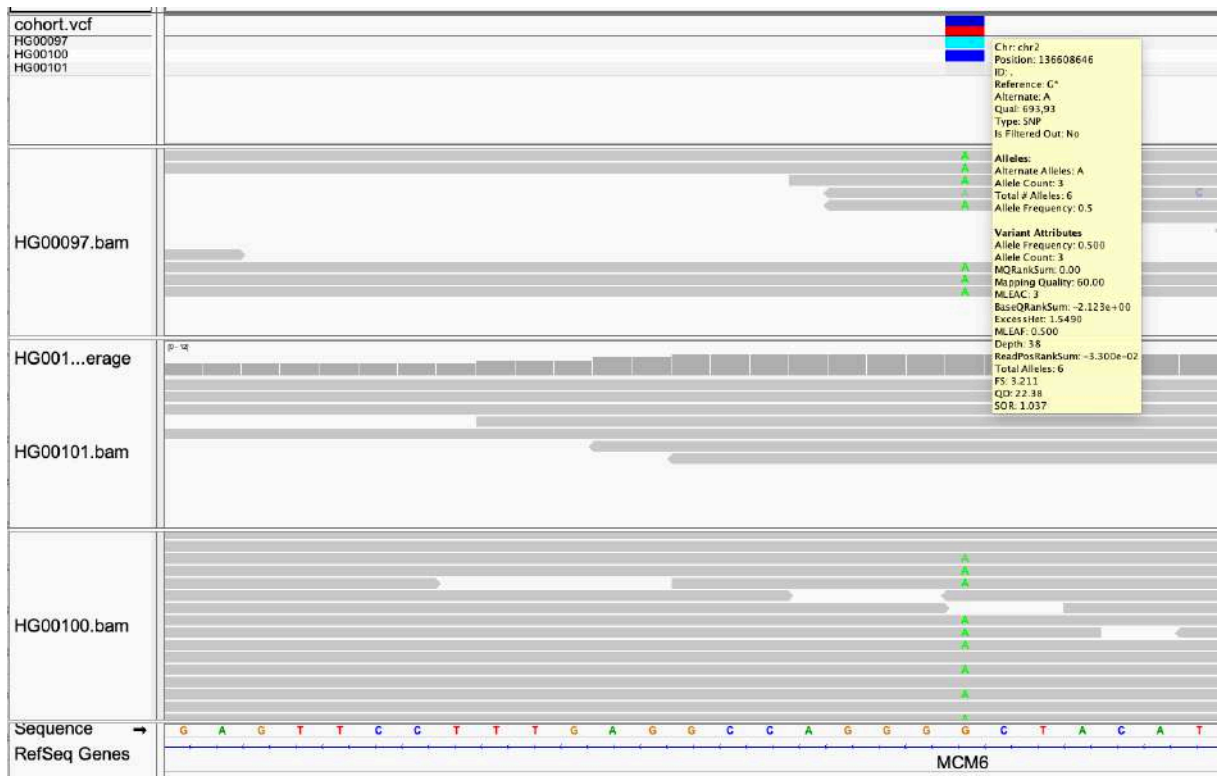
HG00097 is A/A
HG00100 is A/G
HG00101 is G/G

Homozygote genotypes for the alternative allele are colored in light blue
Heterozygote genotypes are colored in dark blue
Homozygote genotypes for the reference allele are colored in light grey.

22. Should any of the individuals avoid drinking milk?
Yes, HG00101 is homozygote for the G/G allele, and therefore do not have new transcription factor binding site that functions as an enhancer that upregulates *LCT* in adulthood.

23. Now let's compare the data shown in IGV with the data in the VCF file. Extract the row for the chr2:136608646 variant in the cohort.vcf file, for example using `grep '136608646' cohort.vcf`. What columns of the vcf file contain the information shown in the upper part of the vcf track in IGV?

```
grep '136608646' cohort.vcf
```
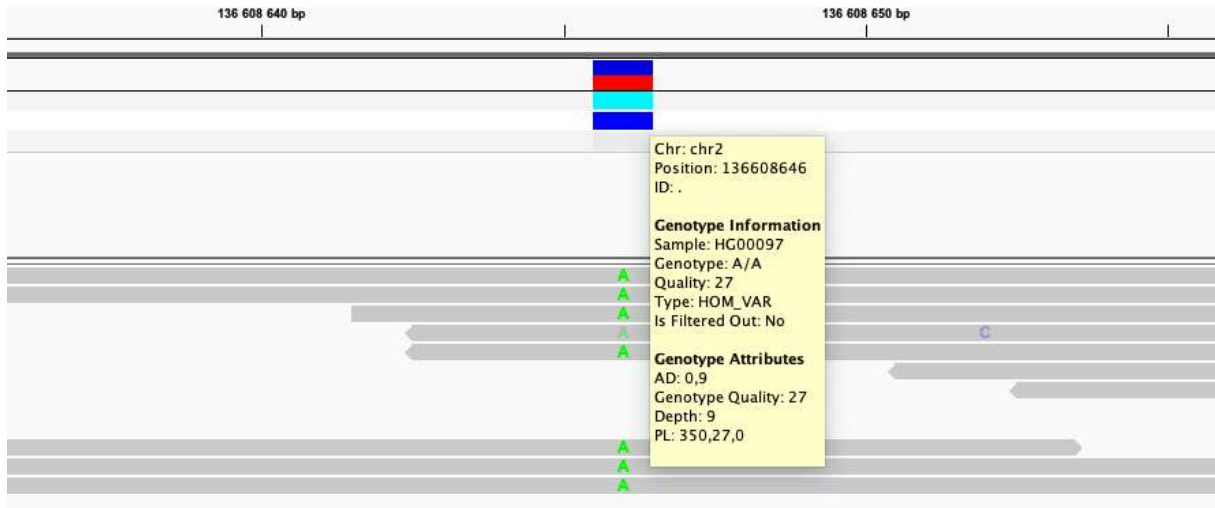
results in

```
2 136608646  .        G        A            693.93      .
 AC=3;AF=0.500;AN=6;BaseQRankSum=-
2.123e+00;DP=38;ExcessHet=1.5490;FS=3.211;MLEAC=3;MLEAF=0.500;MQ=60.00;MQ
RankSum=0.00;QD=22.38;ReadPosRankSum=-3.300e-02;SOR=1.037
 GT:AD:DP:GQ:PL        1/1:0,9:9:27:350,27,0 0/1:11,11:22:99:360,0,405
 0/0:7,0:7:21:0,21,239
```
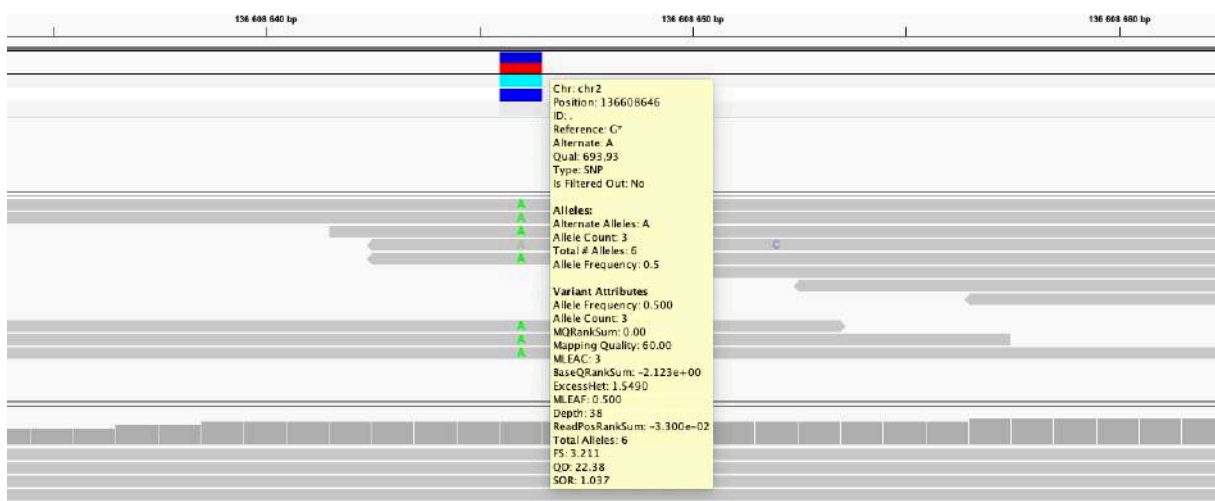
Columns 1-8 of the VCF file is shown in the upper part of the vcf track in IGV.

24. What columns of the vcf file contain the information shown in the lower part of the vcf track?

The sample columns (column 10 and forward) are shown in the lower part of the vcf track, which are called genotype tracks. There is one genotype track for each sample column in the VCF file.

25. Zoom out so that you can see the MCM6 and LCT genes. Is the variant at chr2:136608646 locate within the LCT gene?

No, it is located in an exon of MCM6 which acts as enhancer for LCT.

26. How many variants are present in the cohort.filtered.vcf file?

```
grep -v "^#" cohort.filtered.vcf | wc -l
```

Approximately 719 variants but this can differ a bit.

How many variants passed the filters?

```
grep -v '^#' cohort.filtered.vcf | grep 'PASS' | wc -l
```

The difference should be small, around 4 variants fewer than above (could probably differ a bit).

27. at the variants that did not pass the filters using `grep -v 'PASS' cohort.filtered.vcf`. Try to understand why these variants didn't pass the filter.

```
grep -v 'PASS' cohort.filtered.vcf | grep 'PASS' | wc -l
```

Example output (might not be identical for you):

```
#CHROM    POS   ID    REF   ALT   QUAL  FILTER      INFO  FORMAT          HG00097
HG00100   HG00101
2    136164548   .    C    G    129.26     MQfilter
    AC=1;AF=0.167;AN=6;BaseQRankSum=0.431;DP=32;ExcessHet=3.0103;FS=4.771;MLEAC=1;MLEA
F=0.167;MQ=34.64;MQRankSum=0.431;QD=21.54;ReadPosRankSum=-4.310e-01;SOR=3.258
    GT:AD:DP:GQ:PL    0/0:15,0:15:36:0,36,5400/0:11,0:11:33:0,33,444
    0/1:2,4:6:51:138,0,51
2    136174478   .    G    A    120.26     MQfilter
    AC=1;AF=0.167;AN=6;BaseQRankSum=-
1.981e+00;DP=33;ExcessHet=3.0103;FS=2.430;MLEAC=1;MLEAF=0.167;MQ=38.80;MQRankSum=0.524;Q
D=12.03;ReadPosRankSum=-8.160e-01;SOR=1.60GT:AD:DP:GQ:PL  0/0:12,0:12:33:0,33,495
    0/0:11,0:11:33:0,33,4200/1:6,4:10:99:129,0,205
2    136174552   .    C    T    42.41 MQfilter
    AC=1;AF=0.167;AN=6;BaseQRankSum=0.792;DP=20;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEA
F=0.167;MQ=36.93;MQRankSum=-7.920e-01;QD=6.06;ReadPosRankSum=-
1.068e+00;SOR=0.446GT:AD:DP:GQ:PL  0/0:5,0:5:9:0,9,135    0/0:8,0:8:24:0,24,316
    0/1:5,2:7:51:51,0,148
2    136269833   .    A    T    30.37 QDfilter    AC=1;AF=0.167;AN=6;BaseQRankSum=-
7.530e-
01;DP=25;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.167;MQ=60.00;MQRankSum=0.00;QD=1.90;R
eadPosRankSum=1.19;SOR=1.445 GT:AD:DP:GQ:PGT:PID:PL:PS    0/0:5,0:5:15:.:.:0,15,187
    0|1:14,2:16:39:0|1:136269833_A_T:39,0,617:136269833 0/0:4,0:4:12:.:.:0,12,172
2    136269844   .    A    T    30.37    QDfilter
    AC=1;AF=0.167;AN=6;BaseQRankSum=-9.720e-
01;DP=26;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.167;MQ=60.00;MQRankSum=0.00;QD=1.90;R
eadPosRankSum=-5.570e-01;SOR=1.445 GT:AD:DP:GQ:PGT:PID:PL:PS  0/0:5,0:5:12:.:.:0,12,180
    0|1:14,2:16:39:0|1:136269833_A_T:39,0,617:136269833   0/0:5,0:5:15:.:.:0,15,214
```

Reasons for not passing filter:

##FILTER=<ID=MQfilter,Description="MQ < 40.0">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">

##FILTER=<ID=QDfilter,Description="QD < 2.0">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">