Variant Calling Workflow

Answers to questions

PART1

1A

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

SN:2 means that sequence name is 2 (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

1B

1. ***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
```

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

Genotype

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

3. ***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">
```

4. ***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.

5. ***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.

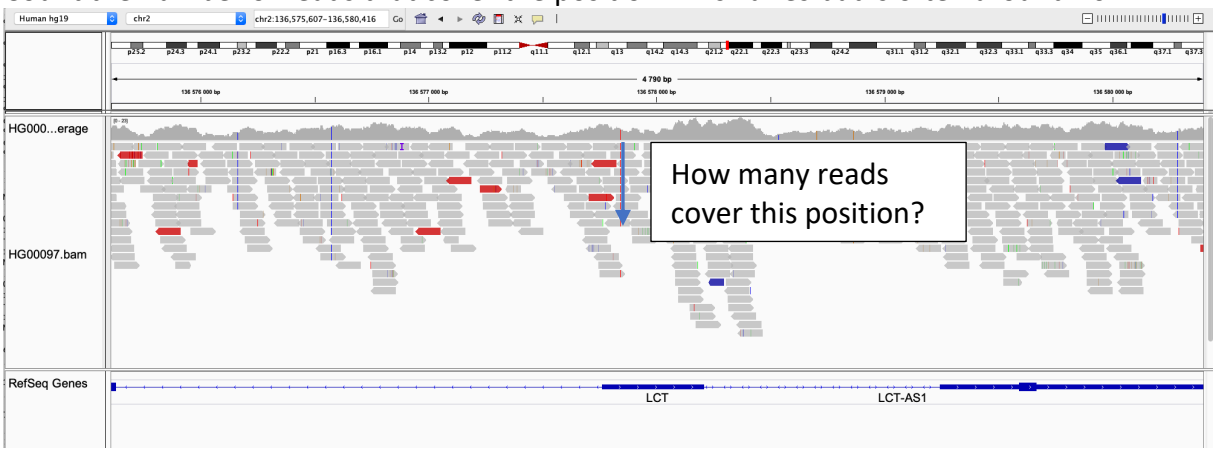
```
2       136545844       .       C       G       427.02
        .
        AC=2;AF=1.00;AN=2;DP=11;ExcessHet=3.0103;FS=0.000;M
LEAC=2;MLEAF=1.00;MQ=60.00;QD=34.86;SOR=1.270
        GT:AD:DP:GQ:PL  1/1:0,11:11:33:441,33,0
```

1C

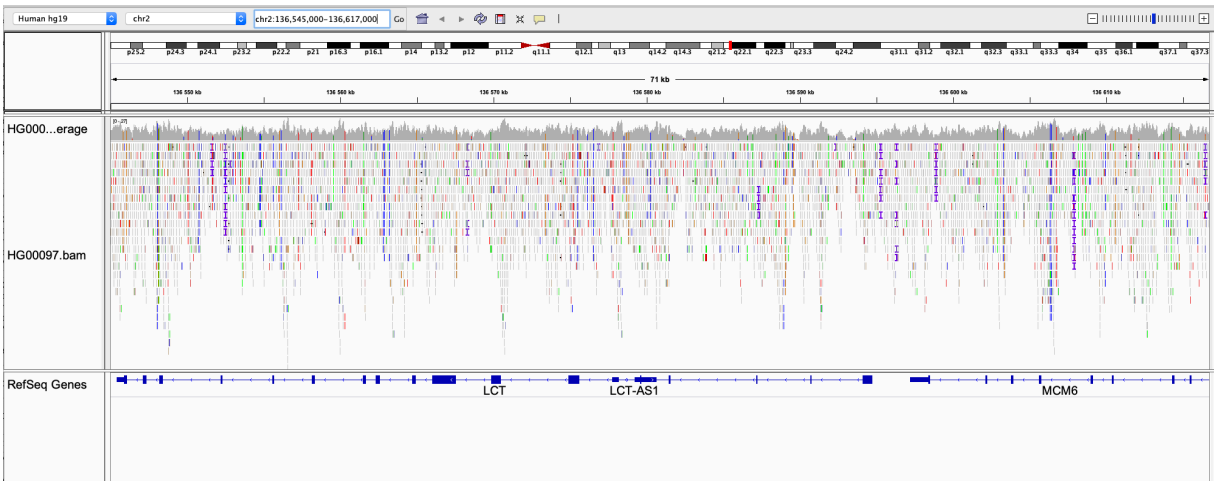
1. ***What is the read length?***
   101

2. ***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.

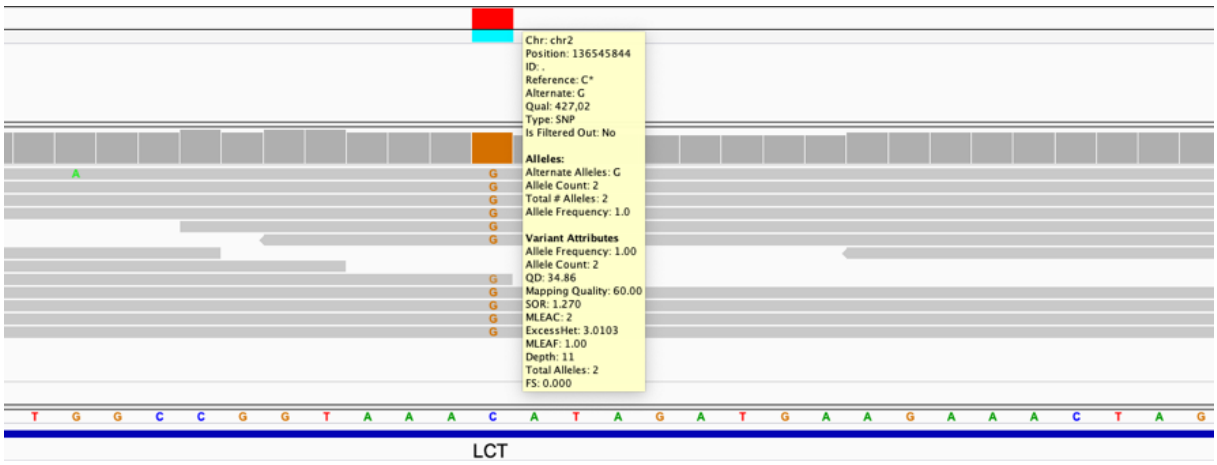


3. **Which genes are located within the region chr2:136545000-136617000?**



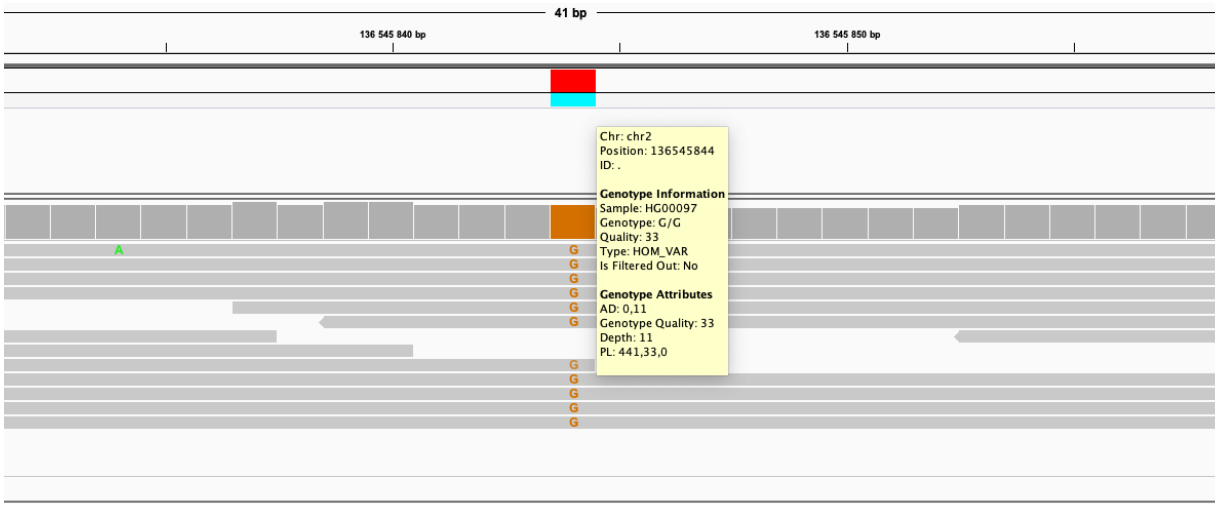
LCT, LCT-AS1 (also called NR_045486) and MCM6

4. ***Hover 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

5. ***Hover 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.

6. ***Look at 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"? (Hover 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.

PART2

2A

1. ***How many data lines does the cohort.g.vcf file have? You can use the 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 348269 lines (or similar)

2. ***How many data lines does the cohort.vcf file have? Explain the difference!***

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

This returns 706 lines (or similar)

Cohort.g.vcf contains information about every position in the analyzed region (although some positions are merged into blocks), cohort.vcf contains information about sites where genetic variants were detected.

3. ***What is encoded in the last three columns of the data lines?***
   Genotypes and genotype attributes of the samples HG00097, HG00100 and HG00101

2B

1. Check how many variants in total that are present in the cohort.filtered.vcf file and how many that have passed the filters. Is the difference big?

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

   706 variants (but this can differ a bit).

How many variants passed the filters?

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

   Returns 564 variants

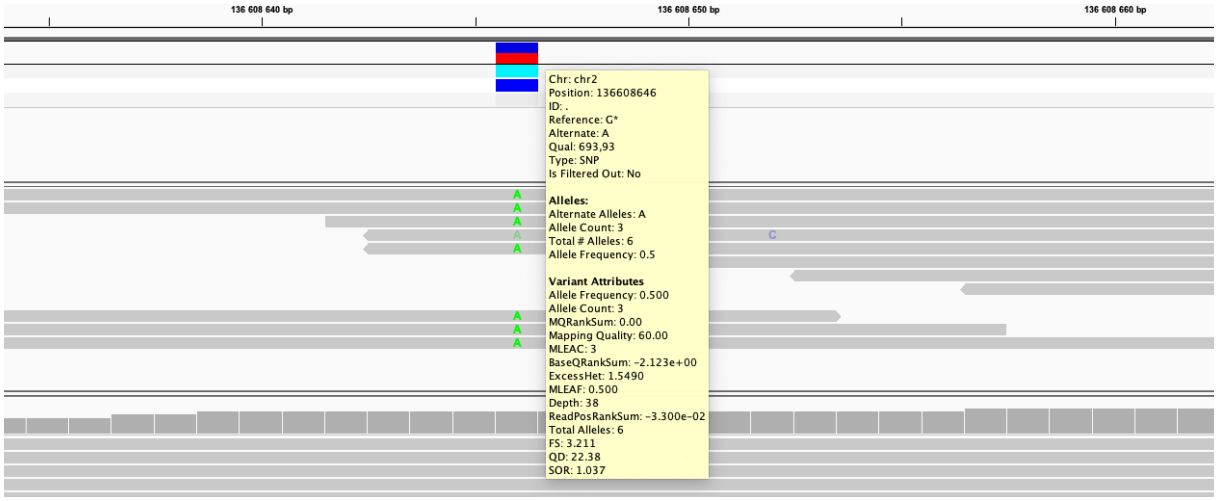
2. **Look at the variants that did not pass the filters using `grep -v 'PASS' cohort.filtered.vcf`. (Do you understand why these variants didn't pass the filter?)**
   They have been removed by the QUALfilt (due to low quality in the QUAL column)

3. **How many variants were removed by vcftools (in cohort.PASS.recode.vcf)?**
   539 of 706 variants remain, so 167 were removed. The numbers can be seen in the cohort.PASS.log. (We added a new depth filter, which explains the lower number than before.)

4. **Try to remove all indels using vcftools. How many variants remain?**
   473 variants remain (use vcftools with --remove-indels)

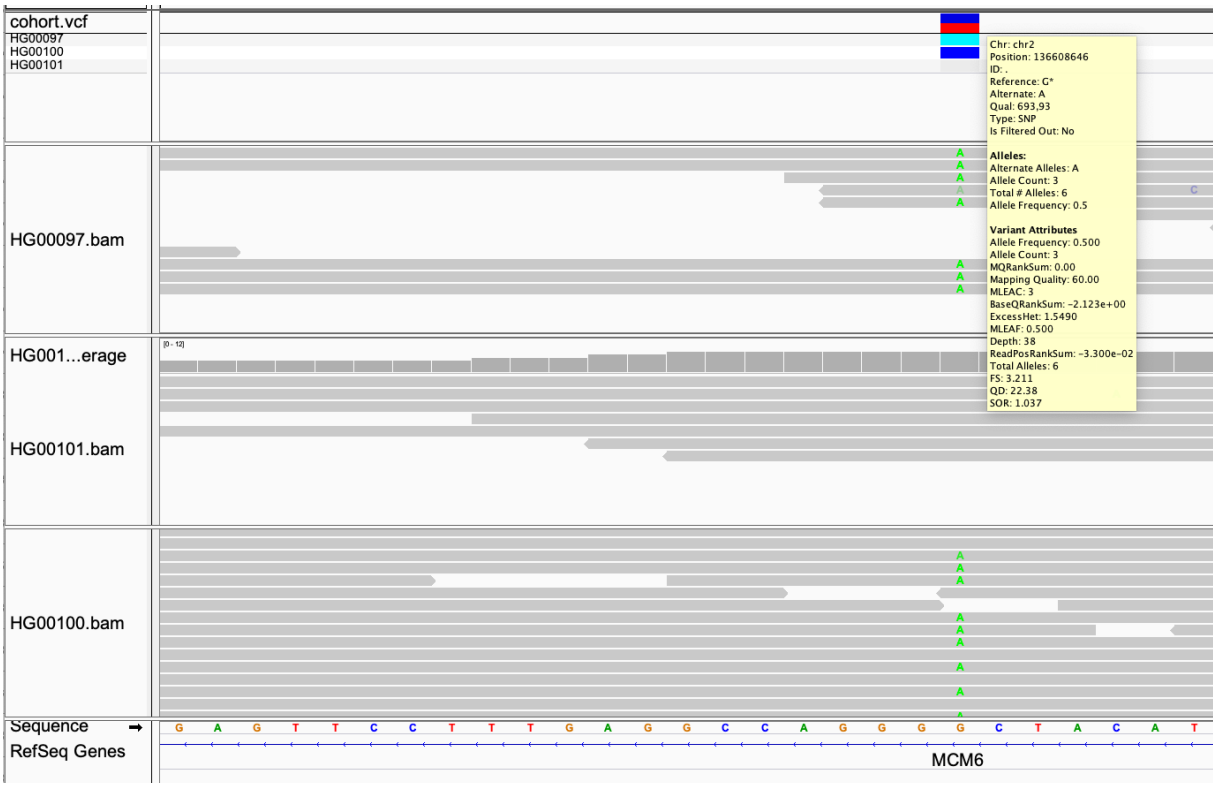
2C

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



Reference allele = G
Alternative allele = A

2. ***What genotype do the three samples have at chr2:136608646? Note how genotypes are color coded in IGV.***



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.

3. Should any of the individuals avoid drinking milk?

Yes, HG00101 is homozygote for the G/G allele, and therefore cannot upregulate *LCT* in adulthood.

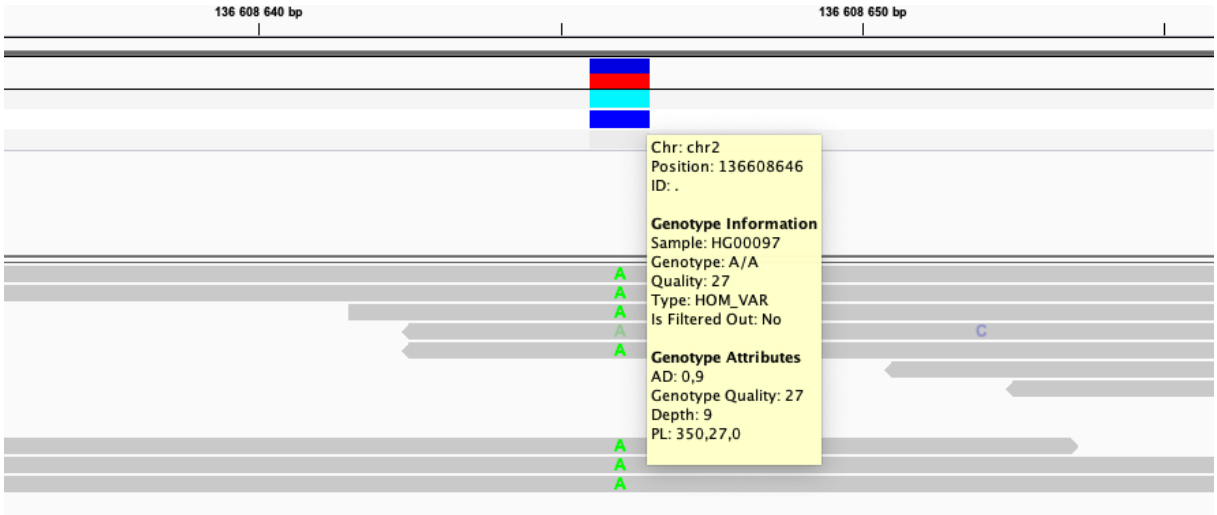
4. Now let's compare the data shown in IGV with the data in the VCF file. Extract the row for the chr2:136608646 variant in the cohort.vcf file, for example using `grep '136608646' cohort.vcf`. What columns of the vcf file contain the information shown in the upper part of the vcf track in IGV?

```
grep '136608646' cohort.vcf
```

results in

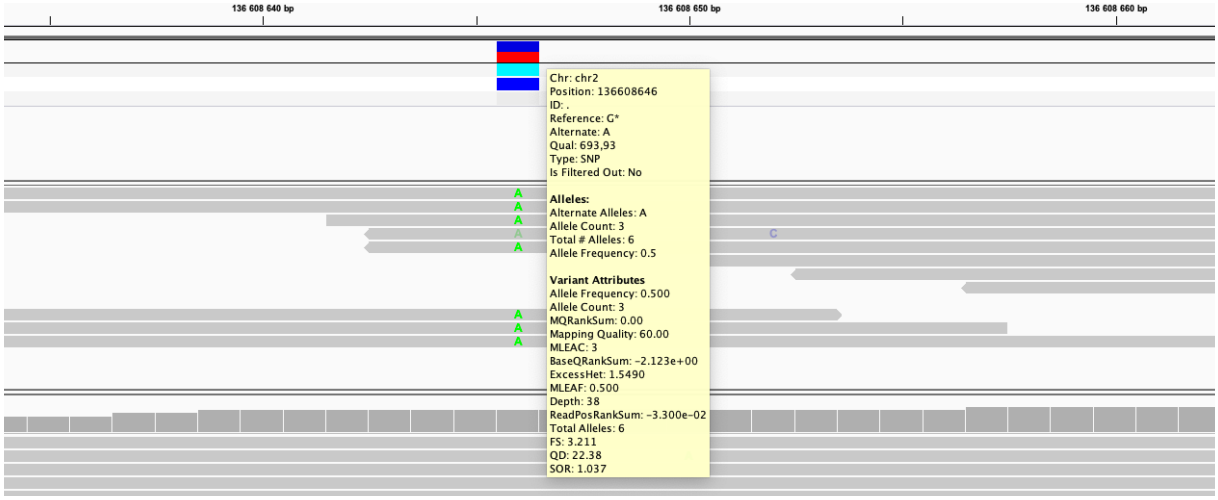
```
2136608646  .          G          A          693.93     .
 AC=3;AF=0.500;AN=6;BaseQRankSum=-
2.123e+00;DP=38;ExcessHet=1.5490;FS=3.211;MLEAC=3;MLEAF=0.500;MQ=60.00;MQ
RankSum=0.00;QD=22.38;ReadPosRankSum=-3.300e-02;SOR=1.037
 GT:AD:DP:GQ:PL        1/1:0,9:9:27:350,27,0 0/1:11,11:22:99:360,0,405
 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.



5. 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.



6. ***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 intron of MCM6 which acts as enhancer for LCT.