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?***

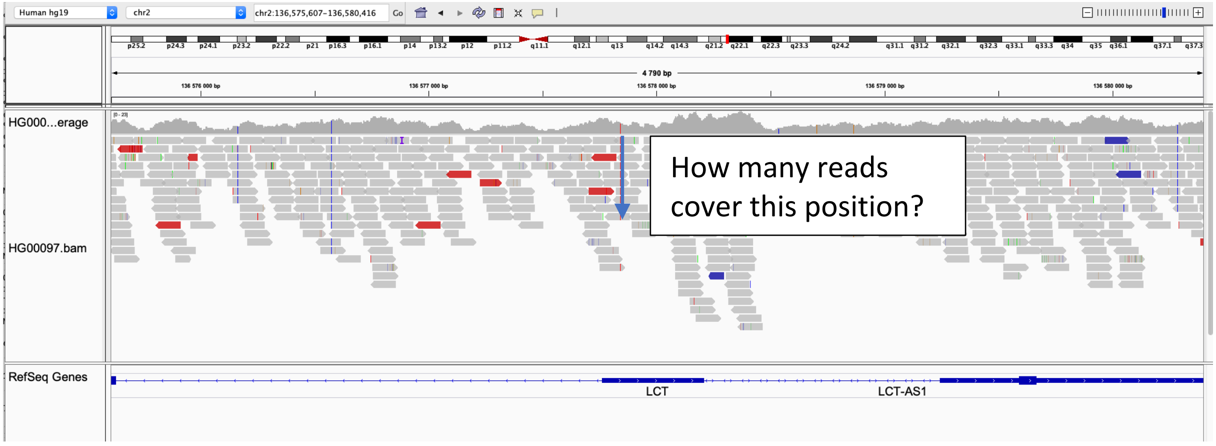
Chromosome 2, position 3843448

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

101 bp

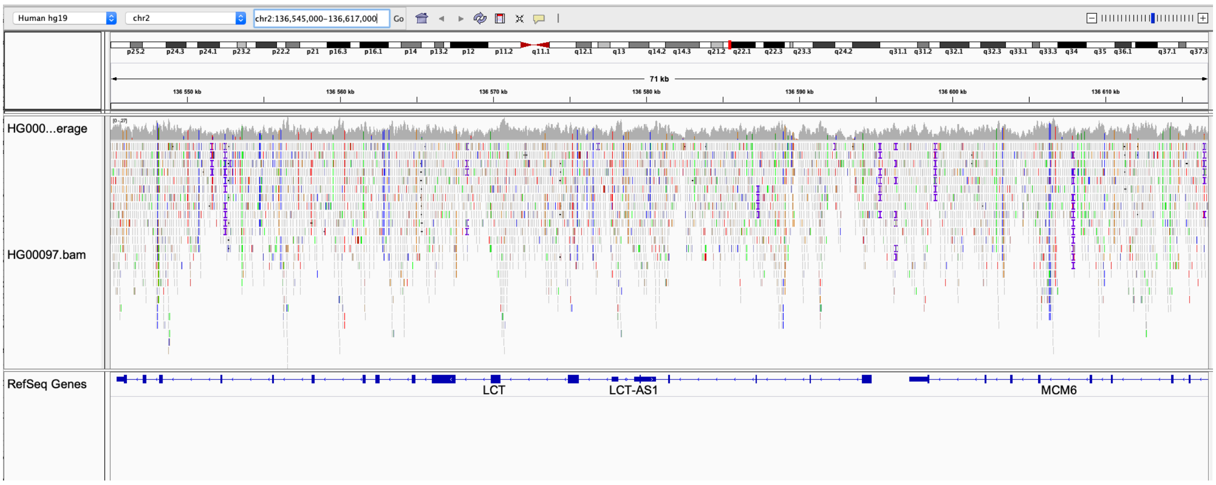
6. ***How can you estimate the coverage in IGV?***

Count the number of reads that cover the position.


Human hg19
chr2
chr2:136,575,607-136,580,416
Go
p25.2 p24.3 p24.1 p23.2 p22.2 p21 p16.3 p16.1 p14 p13.2 p12 p11.2 q11.1 q12.1 q13 q14.2 q14.3 q21.2 q22.1 q22.3 q23.3 q24.2 q31.1 q31.2 q32.1 q32.3 q33.1 q33.3 q34 q35 q36.1 q37.1 q37.3
4 790 bp
136 576 000 bp
136 577 000 bp
136 578 000 bp
136 579 000 bp
136 580 000 bp
HG000...erage
[0 - 23]
HG00097.bam
RefSeq Genes
LCT
LCT-AS1

How many reads cover this position?

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



LCT, LCT-AS1 and MCM6

8. ***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
```

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

Genotype

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

10. ***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.

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.

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

12. ***What is 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
```

13. How many genetic variants was detected in the sample?

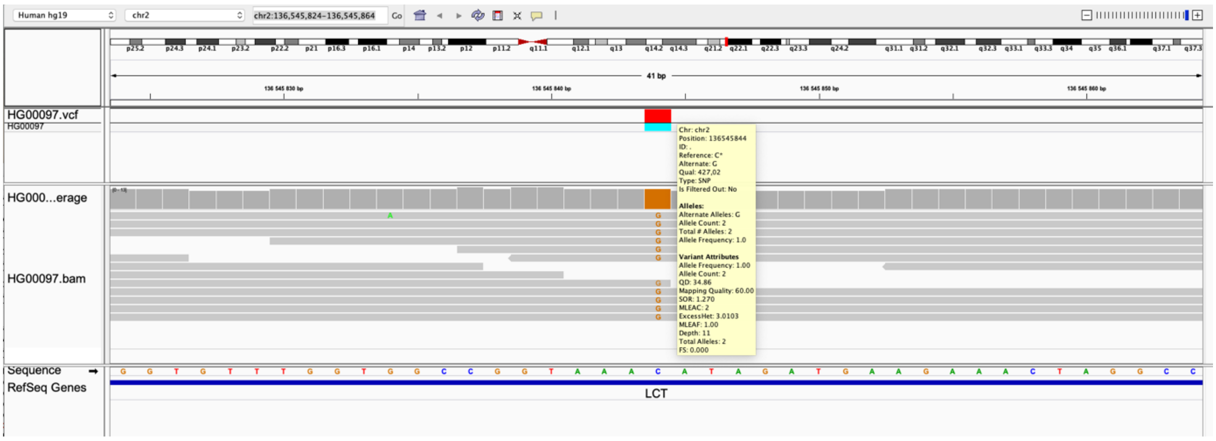
The linux command

grep -v "#" HG00097.vcf | wc -l

extracts all lines in that don't start with "#", and then counts these lines.

206 variants

14. 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?



Chr: chr2
Position: 136545844
ID: .
Reference: C*
Alternate: G
Qual: 427,02
Type: SNP
Is Filtered Out: No

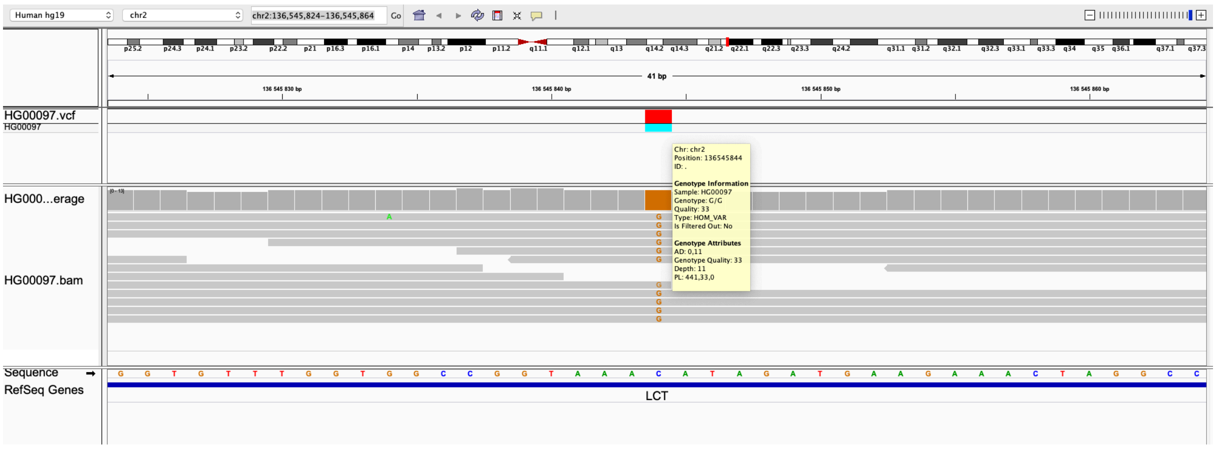
Alleles:
Alternate Alleles: G
Allele Count: 2
Total # Alleles: 2
Allele Frequency: 1.0

Variant Attributes
Allele Frequency: 1.00
Allele Count: 2
QD: 34.86
Mapping Quality: 60.00
SOR: 1.270
MLEAC: 2
ExcessHet: 3.0103
MLEAF: 1.00
Depth: 11
Total Alleles: 2
FS: 0.000

Referenec allele = C

Alternative allele = G

15. 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?



Chr: chr2
Position: 136545844
ID: .

Genotype Information
Sample: HG00097
Genotype: G/G
Quality: 33
Type: HOM_VAR
Is Filtered Out: No

Genotype Attributes
AD: 0,11
Genotype Quality: 33
Depth: 11
PL: 441,33,0

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.

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

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

17. 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

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

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

This returns 718 lines

19. 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.

20. 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
```