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 p.	211 211 211 211 211 211 211 211 211 211	1 q12.1 q13 q14.2 q14.3 q21.2 q22.1 q - 4 790 bp	2223 q23.3 q24.2 q31.1 q31.2 q32.1	q32.3 q33.1 q33.3 q34 q35 q36.1 q37.1 q37.2 158 593 000 bp
HG000erage			How ma cover th	any reads his position?	
RefSeq Genes			LCT	LCT-AS1	

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
	INF0	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.



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?



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?

	- 41 bp			
136 545 840 bp			136 545 850 bp	
	G G G G G G G G G G G	Chr: chr2 Position: 136545844 ID: . Cenotype Information Sample: HC00097 Genotype: C/G Quality: 33 Type: HOM, VAR Is Filtered Out: No Cenotype Atributes 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 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.

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 -I` 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	
HGØ	0097	HGØØ	0100	HGØ	0101				

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

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

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

136 608 640 bp	136 608 650 bp	136 608 660 bp
	Chr: chr2 Position: 136608646 10: . Reference: G* Alternate: A Qual: 693,93 Type: SNP Is Filtered Out: No	
A A	Alleles: Alternate Alleles: A	
	Allele Count: 3 Total # Alleles: 6 Allele Frequency: 0.5	C
	Variant Attributes	
	Allele Count: 3 MQRankSum: 0.00 Mapping Quality: 60.00	
	MLEAC: 3 BaseQRankSum: -2.123e+00 ExcessHet: 1.5490 MLEAE: 0.500	
	Depth: 38 ReadPosRankSum: -3.300e-02 Total Alleles: 6	
	F5: 3.211 QD: 22.38 SOR: 1.037	

Reference allele = GAlternative allele = A

23. 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).

<u> </u>			
cohort.vcf			
HG00097 HG00100		Chr: chr2	
HG00101		Position: 136608646	
		Reference: G*	
		Qual: 693,93	
		Type: SNP	
		is Filtered Out. No	
	Â	Alleles: Alternate Alleles: A	
		Allele Count: 3	6
	Â	Allele Frequency: 0.5	Ŭ
		Variant Attributes	
HG00097.bam		Allele Frequency: 0.500	
	A	Allele Count: 3 MQRankSum: 0.00	
		Mapping Quality: 60.00	
		BaseQRankSum: -2.123e+00	
		ExcessHet: 1.5490 MLFAF: 0.500	
	0-12	Depth: 38	
HG001erage		ReadPosRankSum: -3.300e-02 Total Alleles: 6	
		FS: 3.211 OD: 22.38	
		SOR: 1.037	
HG00101.bam			
	A		
UC00100 hom			
HG00100.bam			-
	<u></u>		
	^		
	A		
Sequence		0 T A 0	
RefSeq Genes	MCM6		
HG00097 is			
HG00100 is	SA/G		

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.

24. 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.

25. 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 1/1:0,9:9:27:350,27,00/1:11,11:22:99:360,0,405 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.

136 608 640 bp	136 608 650 bp 	I	
	Chr: chr2 Position: 136608646 ID: . Genotype Information Sample: HG00097 Genotype: A/A Quality: 27 Type: HOM_VAR Is Filtered Out: No Genotype Attributes AD: 0,9 Genotype Quality: 27 Depth: 9 PL: 350,27,0		
	A		

26. 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.

136 608 640 bp	136 608 650 bp	136 608 660 bp
	136 668 660 hp Chr: chr2 Position: 136608646 ID: Reference: C* Alternate: A Qui: 693,93 Type: SNP Filtered Out: No. A Alternate: A Quie leo Court: 3 Total # Alleles: A Allele Frequency: 0.500 Variant Attributes Allele Frequency: 0.500 MRRankSum: -0.000 MRRankSum: -2.123e+00 Excessite: 1.5490 Excessite: 1.5490	136 666 660 hp
	QD: 22.38 SOR: 1.037	

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

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

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

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

716 variants

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

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

711 variants

30. 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

#CHRC	M	POS	ID	REF	ALT	QUAL	FILTEF	2	INF0	FORMAT	HG00097
HG001	L00	HG001	.01								
2	13616	4548		С	G	129.2	6	MQfil ¹	er		
	AC=1;	AF=0.1	167 ; AN=	=6;Base	eQRanks	Sum=0.	431;DP=	=32;Ex	cessHe	et=3.0103;FS	=4.771;MLEAC=1;MLEA
F=0.1	L67;MQ=	34.64	;MQRanl	<sum=0< td=""><td>.431;Q</td><td>D=21.5</td><td>4;Read</td><td>PosRan</td><td>kSum=-</td><td>-4.310e-01;S</td><td>SOR=3.258</td></sum=0<>	.431;Q	D=21.5	4;Read	PosRan	kSum=-	-4.310e-01;S	SOR=3.258
	GT:AD	:DP:GO):PL	0/0:1	5,0:15	:36:0,	36,540	0/0:11	l,0:11	:33:0,33,44	4
	0/1:2	,4:6:5	51:138,	0,51							
2	13617	4478		G	Α	120.2	6	MQfilt	er		
	AC=1;	AF=0.1	L67;AN=	=6;Base	eQRanks	Sum=-					
1.981	le+00;D	P=33;	Excessl	let=3.	0103;F	S=2.43	0;MLEA	C=1;ML	EAF=0	.167;MQ=38.8	80;MQRankSum=0.524;Q
D=12.	03;Rea	dPosRa	ankSum=	-8.16	0e-01;	SOR=1.	60GT:AI	D:DP:G	Q:PL	0/0:12,0:12	2:33:0,33,495
	0/0:1	1,0:11	L:33:0,	33,420	00/1:6	,4:10:	99:129	,0,205			
2	13617	4552	•	С	Т	42.41	MQfilt	er			
	AC=1;	AF=0.1	167;AN=	=6;Base	eQRanks	Sum=0.	792;DP=	=20;Ex	cessHe	et=3.0103;FS	=0.000;MLEAC=1;MLEA
F=0.1	L67;MQ=	36.93	;MQRanl	<sum=-< td=""><td>7.920e</td><td>-01;QD</td><td>=6.06;I</td><td>ReadPo</td><td>sRanks</td><td>Sum=-</td><td></td></sum=-<>	7 . 920e	-01;QD	=6.06;I	ReadPo	sRanks	Sum=-	
1.068	3e+00;S	0R=0.	446GT:/	AD:DP:	GQ:PL	0/0:5	,0:5:9	:0,9,1	35	0/0:8,0:8:2	24:0,24,316
	0/1:5	,2:7:5	51:51 , 0),148							
2	13626	9833	•	Α	Т	30.37	QDfilt	er	AC=1;	AF=0.167;AN=	=6;BaseQRankSum=-
7.530)e-										
01;DF	P=25;Ex	cessH	et=3.02	103;FS	=0.000	;MLEAC	=1;MLE/	4F=0.1	67;MQ=	=60 . 00;MQRan	1kSum=0.00;QD=1.90;R
eadPo	sRankS	um=1.	19;SOR=	=1.445	GT:AD	:DP:GC	PGT:P	ID:PL:	PS	0/0:5,0:5:1	15:.:.:0,15,187
	0 1:1	4,2:10	5:39:0	1:1362	269833 <u>-</u>	_A_T:3	9,0,617	7:1362	69833	0/0:4,0:4:1	12:.:.:0,12,172
2	136	269844	ł.			Α		Т		30.37	QDfilter
	AC=	1;AF=0).167;A	N=6;Ba	aseQRar	∩kSum=	-9 . 720e	<u>)</u> –			
01;DF	P=26;Ex	cessH	et=3.02	103;FS	=0.000	;MLEAC	=1;MLE	4F=0.1	67;MQ=	=60 . 00;MQRan	1kSum=0.00;QD=1.90;R
eadPo	sRankS	um=-5	.570e-0	01;SOR	=1.445	GT:AD	:DP:GQ:	PGT:P	ID:PL:	PS 0/0:5,0	:5:12:.:.0,12,180
	0 1	:14,2:	16:39:	0 1:13	3626983	33_A_T	:39,0,6	517:13	526983	3 0/0:5 , 0	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">