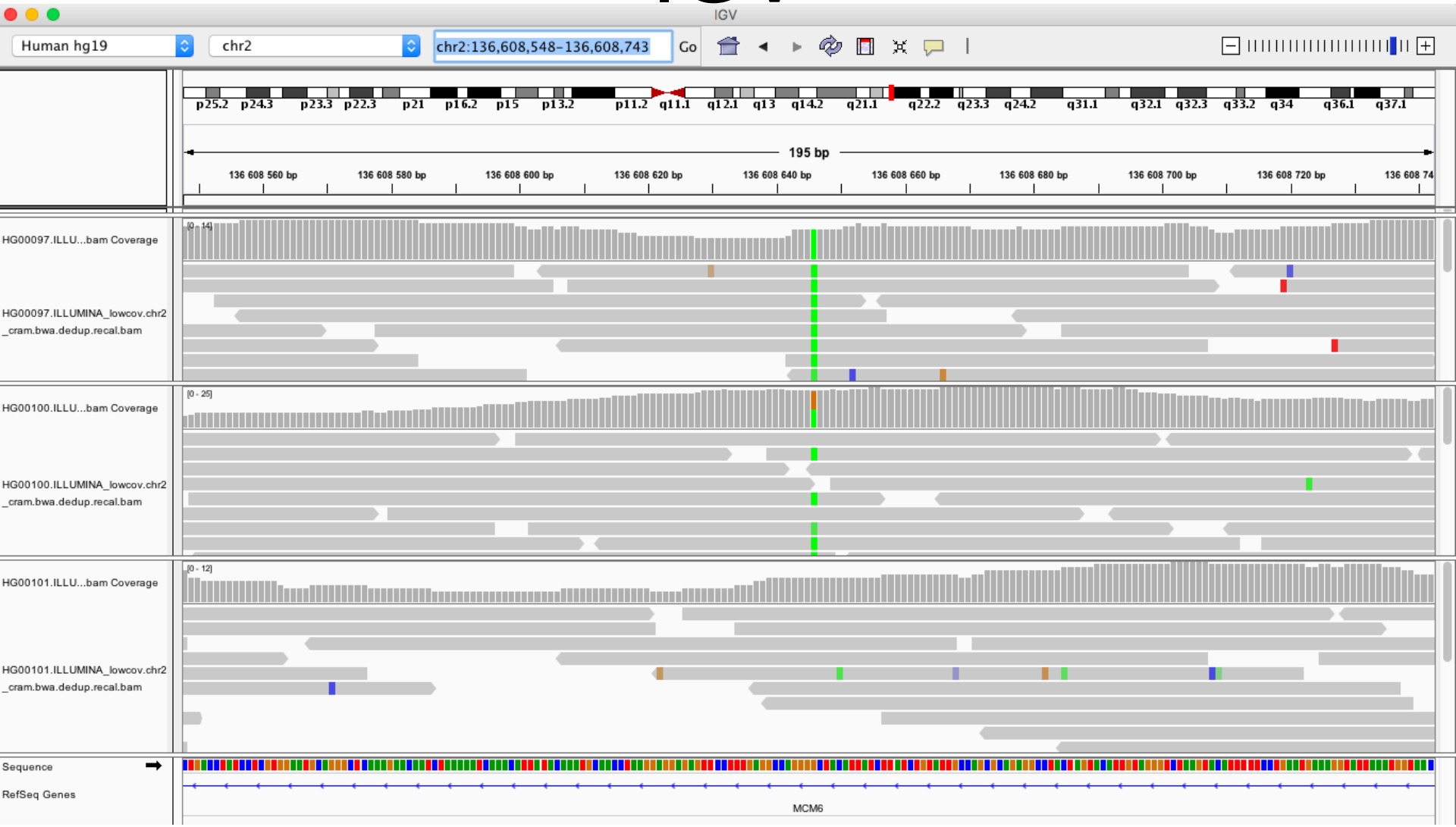
NGS workflow

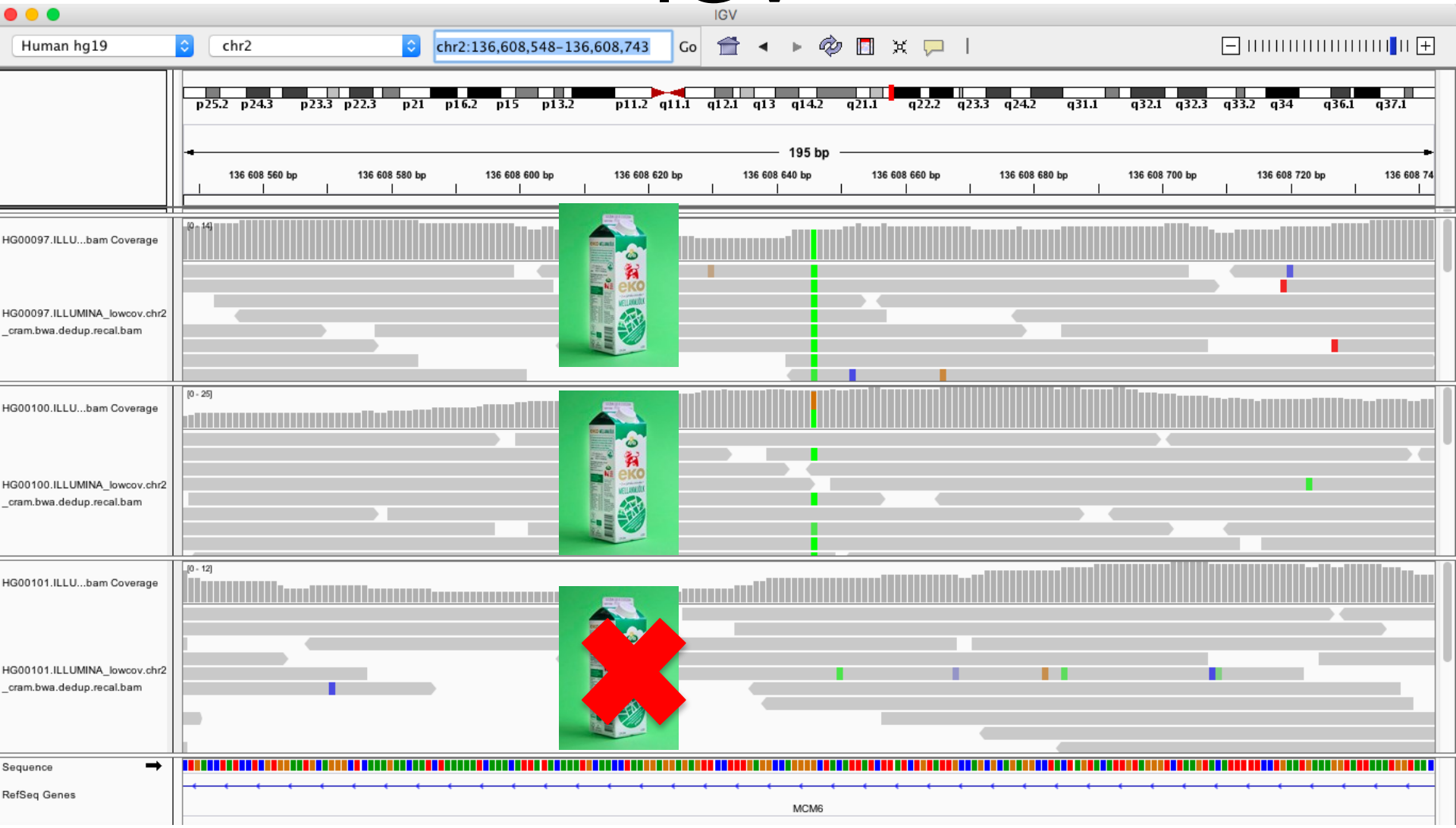
Lab discussion

IGV



Read length?
Do we have decent coverage across the variant position?
Homozygotes/heterozygotes?
Who can drink milk?

IGV


IGV
Human hg19
chr2
chr2:136,608,548-136,608,743
Go
p25.2 p24.3 p23.3 p22.3 p21 p16.2 p15 p13.2 p11.2 q11.1 q12.1 q13 q14.2 q21.1 q22.2 q23.3 q24.2 q31.1 q32.1 q32.3 q33.2 q34 q36.1 q37.1
195 bp
136 608 560 bp
136 608 580 bp
136 608 600 bp
136 608 620 bp
136 608 640 bp
136 608 660 bp
136 608 680 bp
136 608 700 bp
136 608 720 bp
HG00097.ILLU...bam Coverage
HG00097.ILLUMINA_lowcov.chr2_cram.bwa.dedup.recal.bam
HG00100.ILLU...bam Coverage
HG00100.ILLUMINA_lowcov.chr2_cram.bwa.dedup.recal.bam
HG00101.ILLU...bam Coverage
HG00101.ILLUMINA_lowcov.chr2_cram.bwa.dedup.recal.bam
Sequence
RefSeq Genes
MCM6

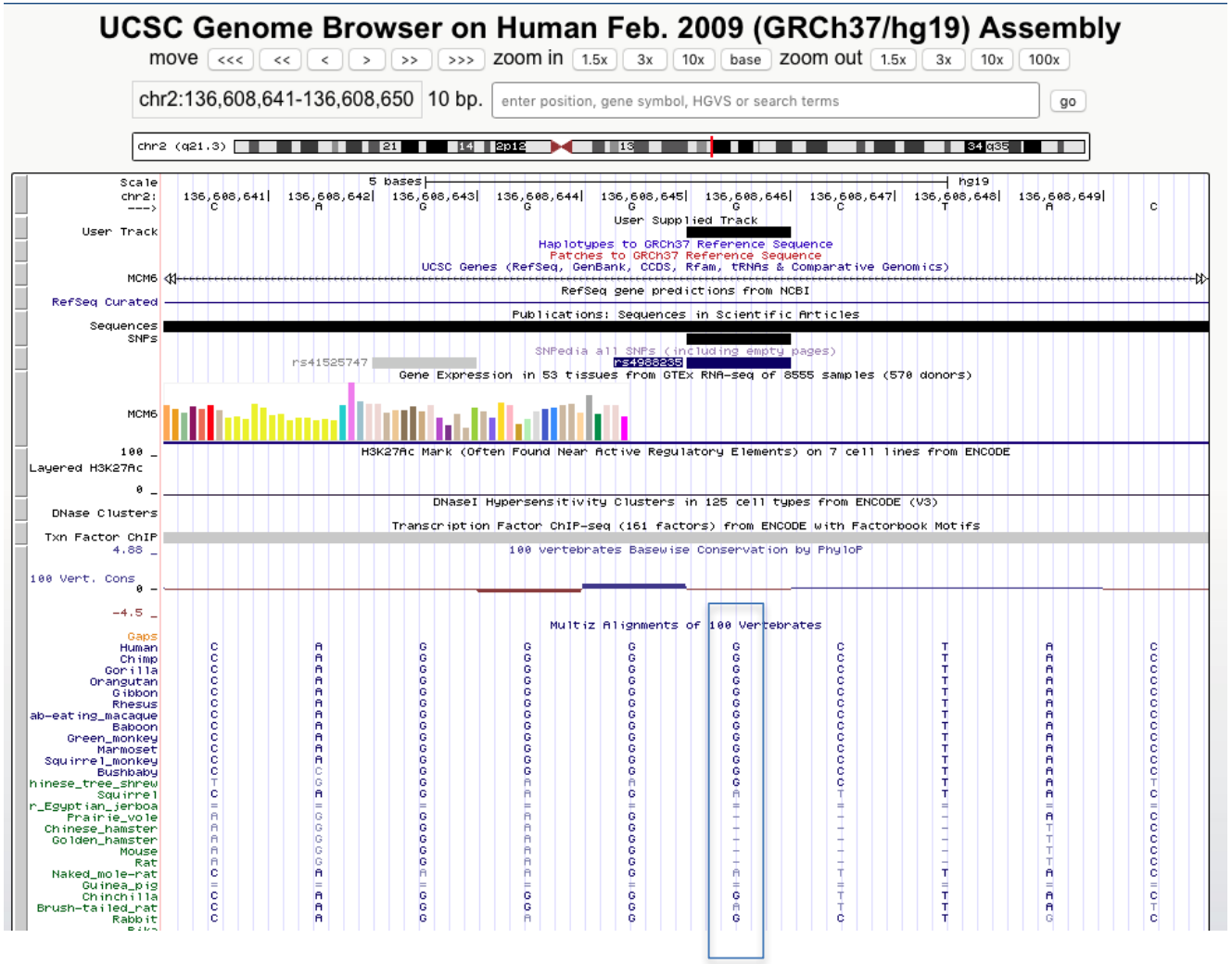
Vcf files

Descriptions of each field can be found in the header section (lines starting with '#')

```
CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00097 HG00100 HG00101
2       136608646   .       G       A       628.02  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-3.410e-01;ClippingRankSum=0.00;DP=38;ExcessHet=1.5490;FS=3.211;MLEAC=3;MLEAF=0.500;MQ=60.00;MQRankSum=0.00;QD=20.26;ReadPosRankSum=-3.970e-01;SOR=1.037;set=snp
GT:AD:DP:GQ:PL  1/1:0,9:9:27:316,27,0       0/1:11,11:22:99:348,0,382       0/0:7,0:7:21:0,21,223
```

UCSC

- Most species have the G allele, however squirrels and naked mole rat apparently have the A allele



Questions

- MarkDuplicates: 588 duplicate reads
- 697 of the 711 variants passed the filters
- 108 variants are indels (use: --keep-only-indels)

Warnings

- Warnings are common, google them if you are not sure what they mean
  - JointGenotyping: Some filters require heterozygote variants
  - You probably got a warning that optical duplicates were not discovered. This is because the read names are not properly formatted to be able to do this
  - Annovar: “1 invalid alternative alleles found in input file” (it is a '*', which is a symbol used by JointGenotyping when some samples have a SNP and others have an indel covering that site.