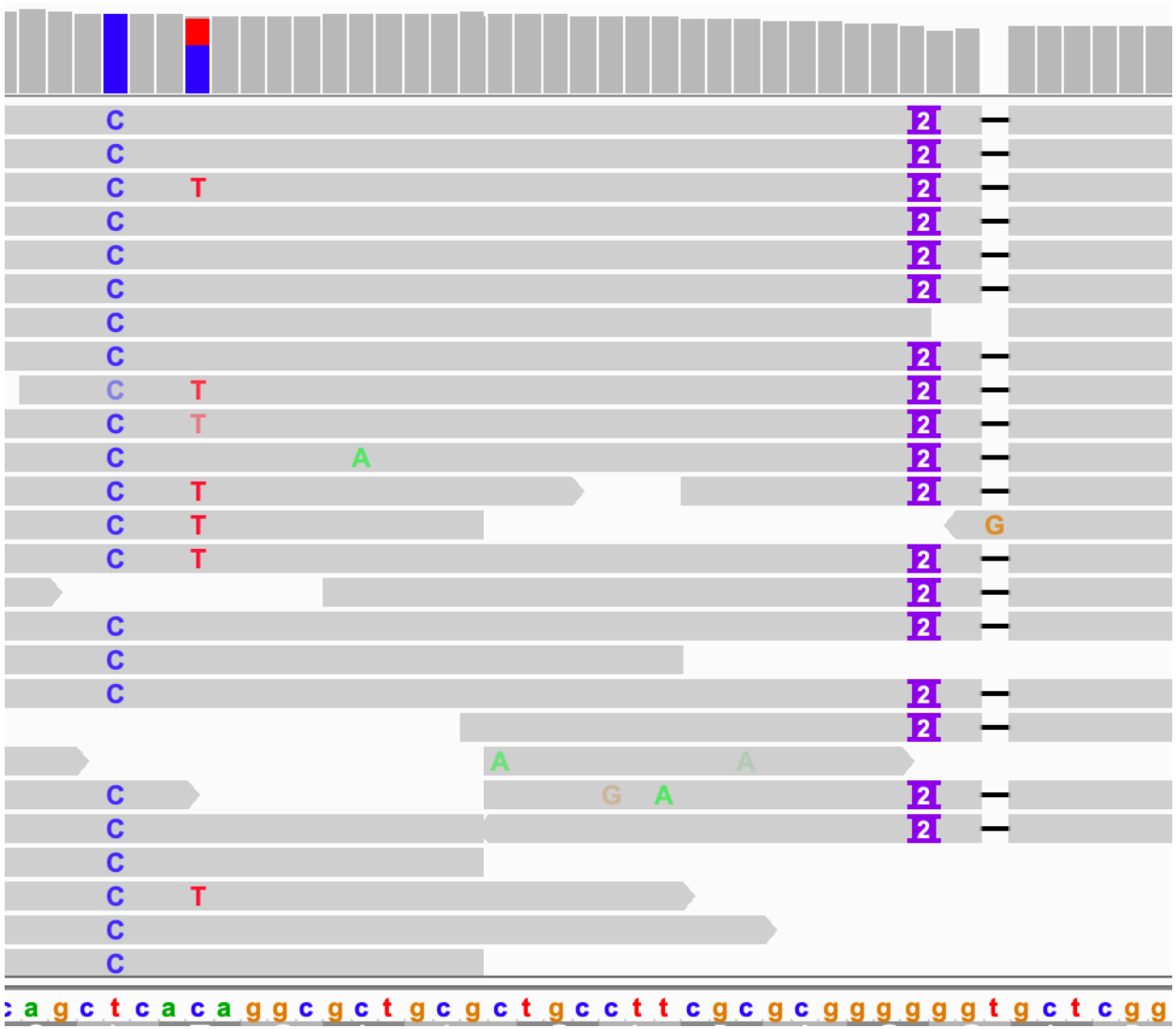
NGS – variant analysis

Variant calling



vcf

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

samples

Question

What would be a command to get the number of variants in a vcf file? (the -v option will invert the match; the -c option will give the count of matches)

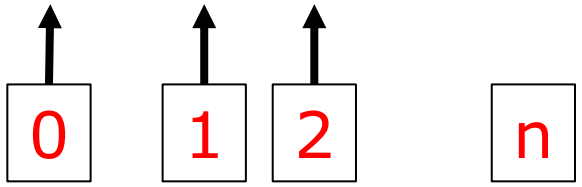
A. `grep -v ^# my.vcf | wc -l`

B. `wc -l my.vcf | grep -c`

C. `grep -c ^# my.vcf`

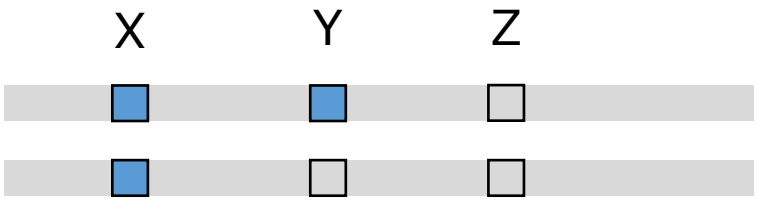
D. `grep -c CHROM my.vcf`

#CHROM	POS	ID	REF	ALT
20	14370	rs6054257	G	A
20	17330	.	T	A
20	1110696	rs6040355	A	G,T
20	1230237	.	T	.
20	1234567	microsat1	GTC	G,GTCT

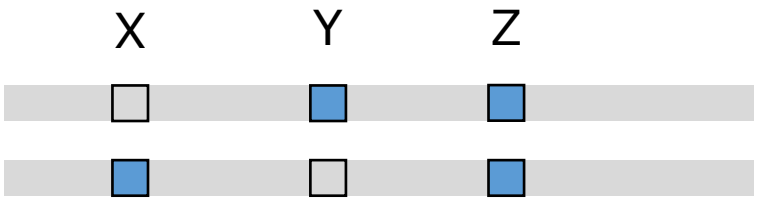


FORMAT	NA00001	NA00002
GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
GT:GQ:DP	0/1:35:4	0/2:17:2

sample 1



sample 2



sample1.vcf

CHROM	POS	ID	SAMP1
20	1101	SNPX	1|1
20	1203	SNPY	0|1

sample2.vcf

CHROM	POS	ID	SAMP2
20	1101	SNPX	1|0
20	1203	SNPY	0|1
20	1253	SNPZ	1|1

combined.vcf

CHROM	POS	ID	SAMP1	SAMP2
20	1101	SNPX	1|1	1|0
20	1203	SNPY	0|1	0|1
20	1253	SNPZ	?	1|1

Question

What would be solution for this ‘missing genotype’ problem?

A. Do a variant call on all samples in one go
B. Store information on non-variant regions in the vcf
C. Fill in missing values at missing genotypes

Missing genotype problem

- Most variant callers genotype all samples in one go. But:
  - variant calling process can become very computational intensive
  - new sample? Redo entire variant call
- GATK uses GVCF:
  - Store information on non-variant regions

Other software

- **freebayes**: haplotype-aware variant calling -> good alternative to gatk
- **bcftools**: working with vcfs (part of samtools)
- **vcftools**: working with vcfs
- **whatshap**: haplotyping
- **medaka**: SNP calling in Oxford nanopore data

GATK workflow

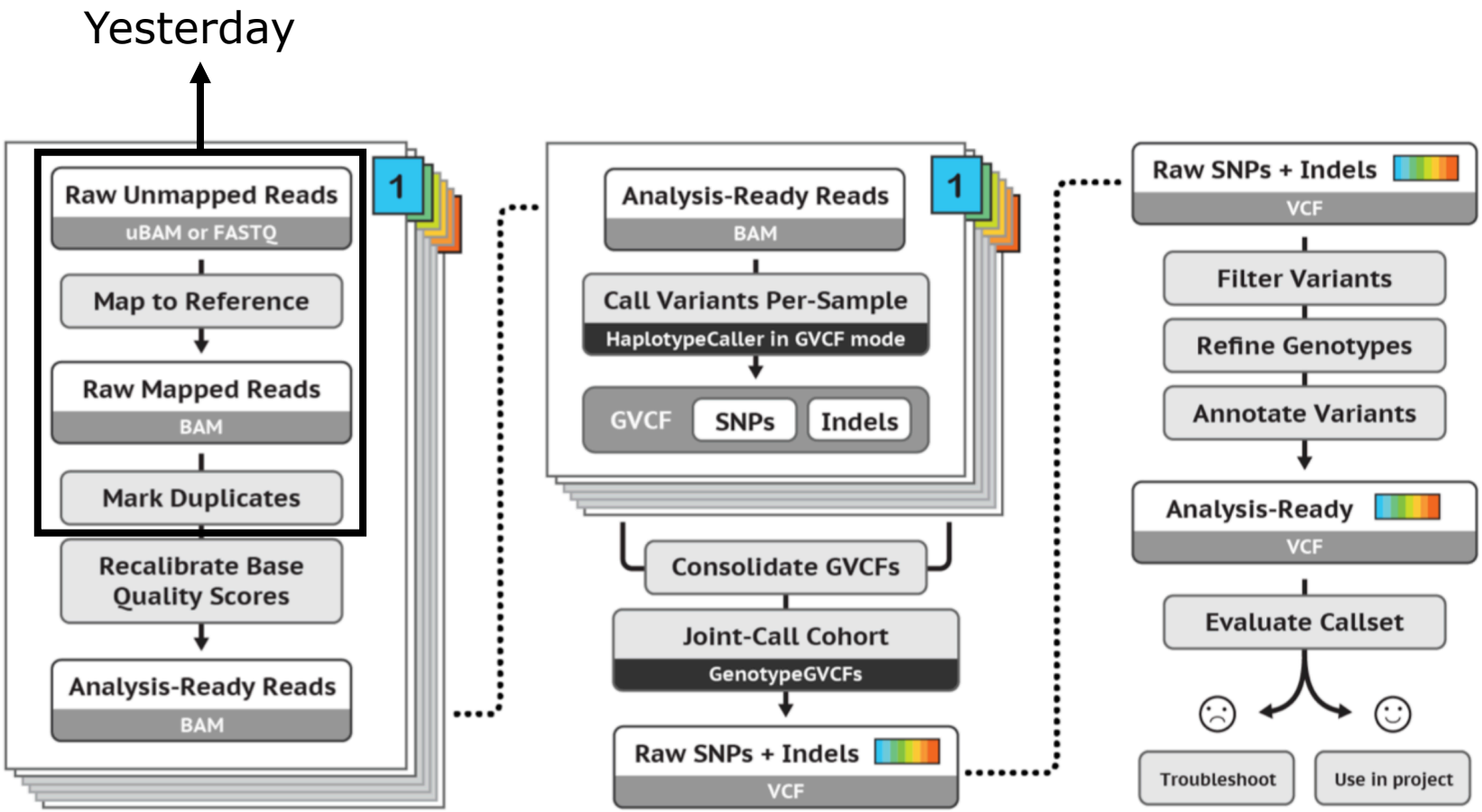


image source: https://gatk.broadinstitute.org