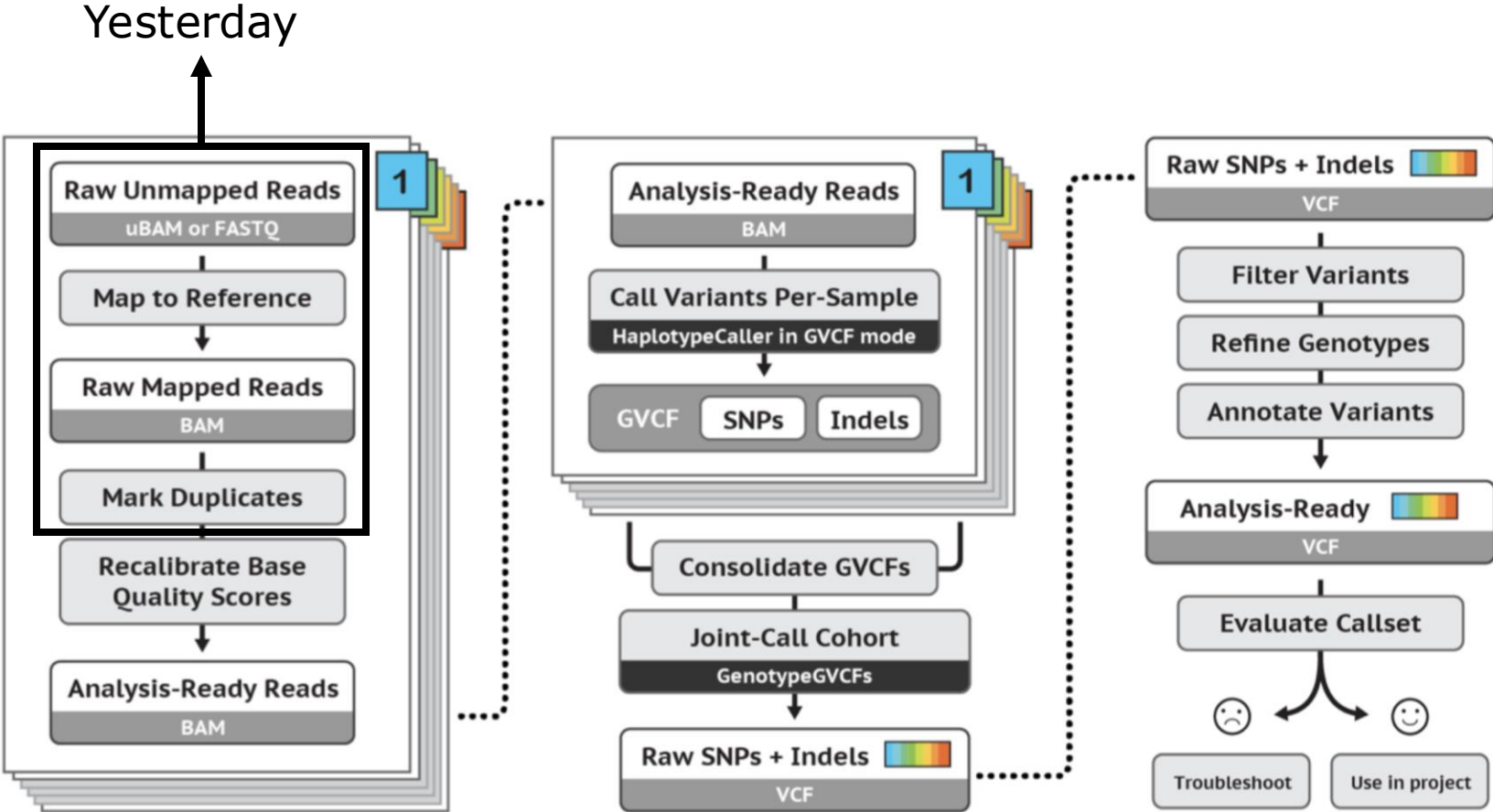
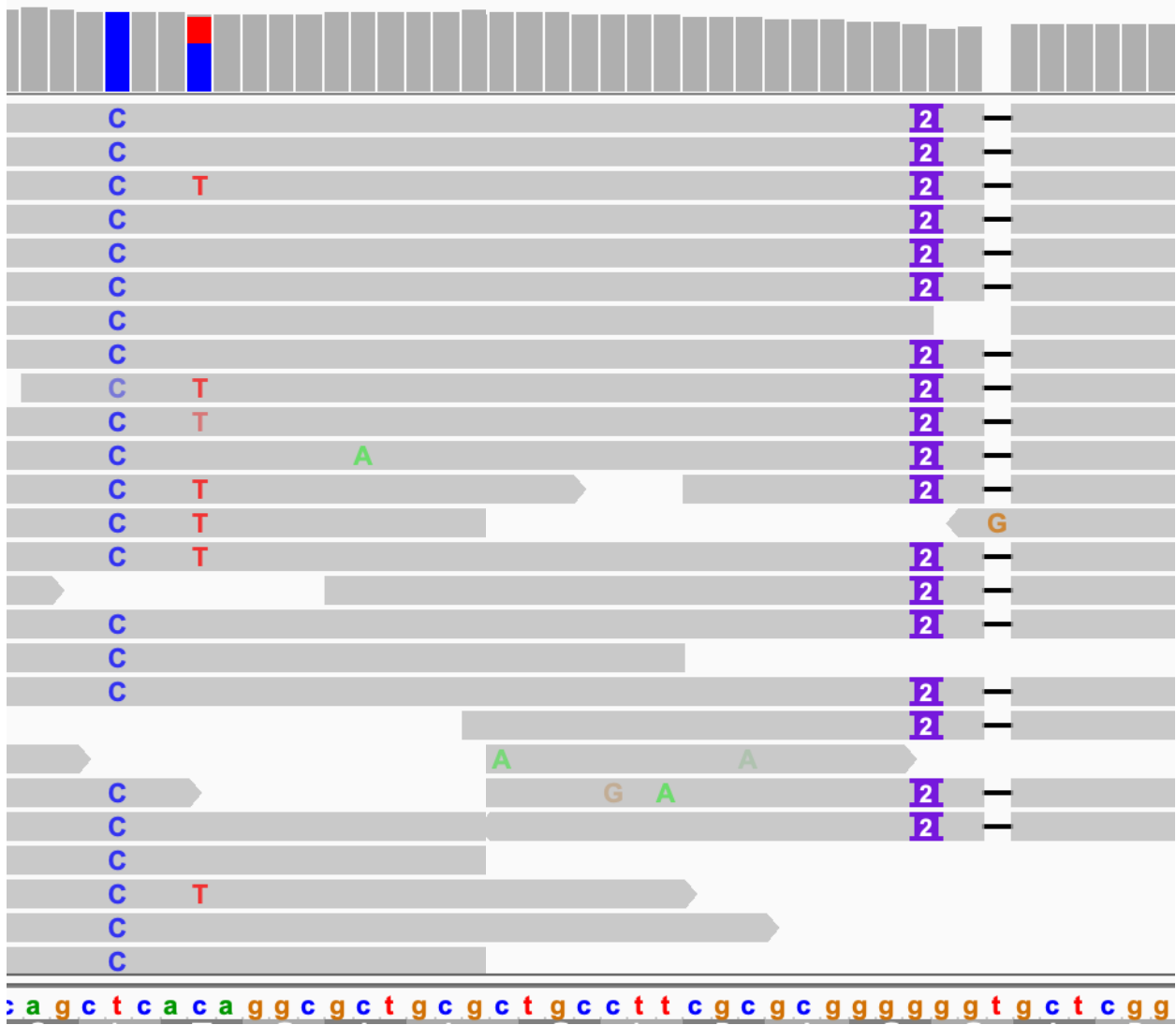


NGS - variant analysis

Variant calling

GATK workflow

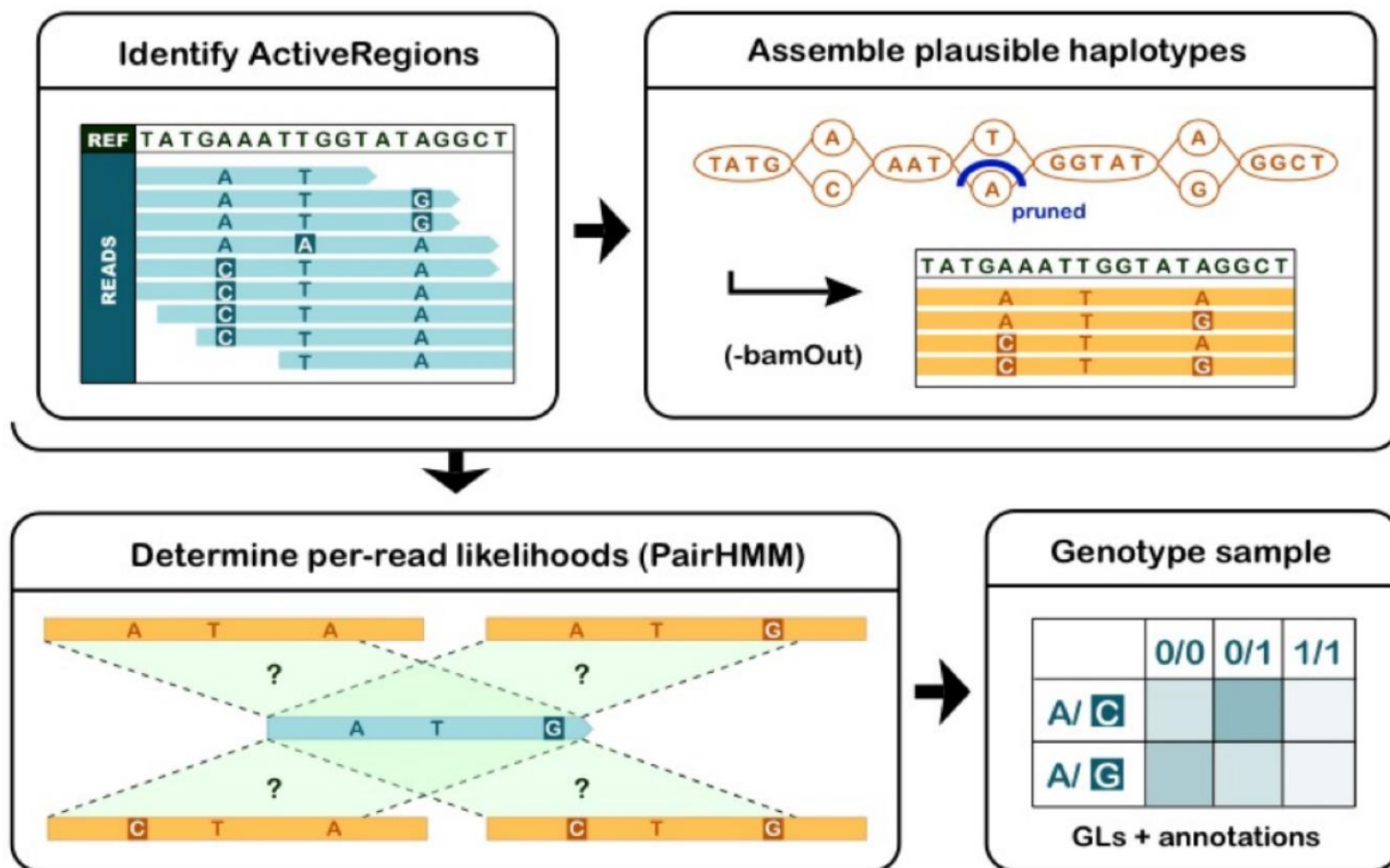




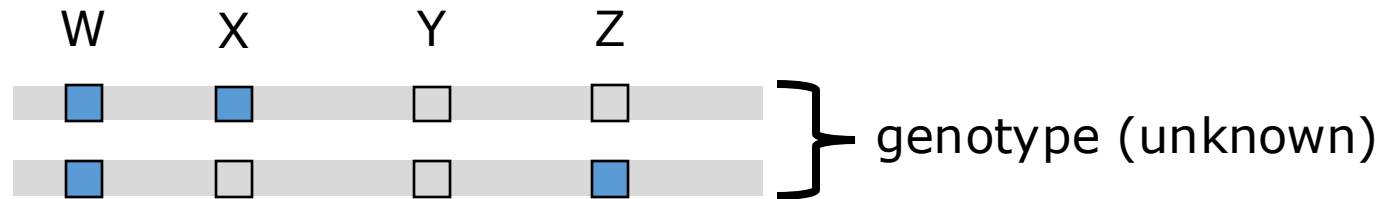
Three important questions

- Is there a variant at location X?
 - Deviation from REF in the alignments
- What are the alleles?
 - The variation in sequence in these deviations
- What is the genotype (HomRef, Heterozygote or HomAlt)?
 - Estimating the allele counts in the sample

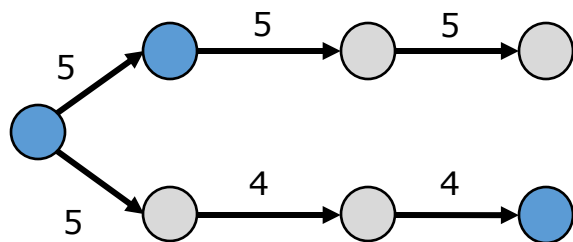
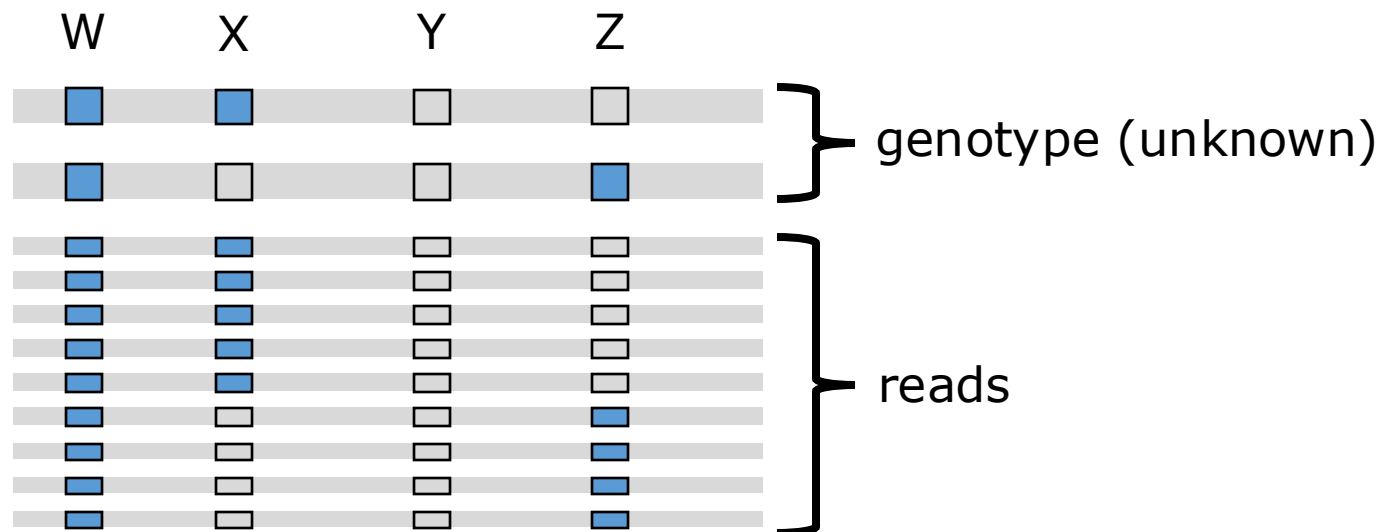
HaplotypeCaller



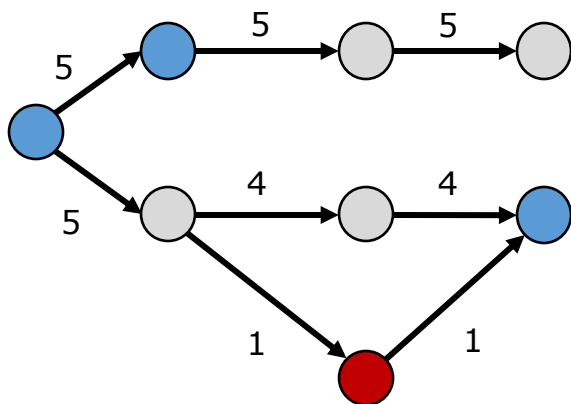
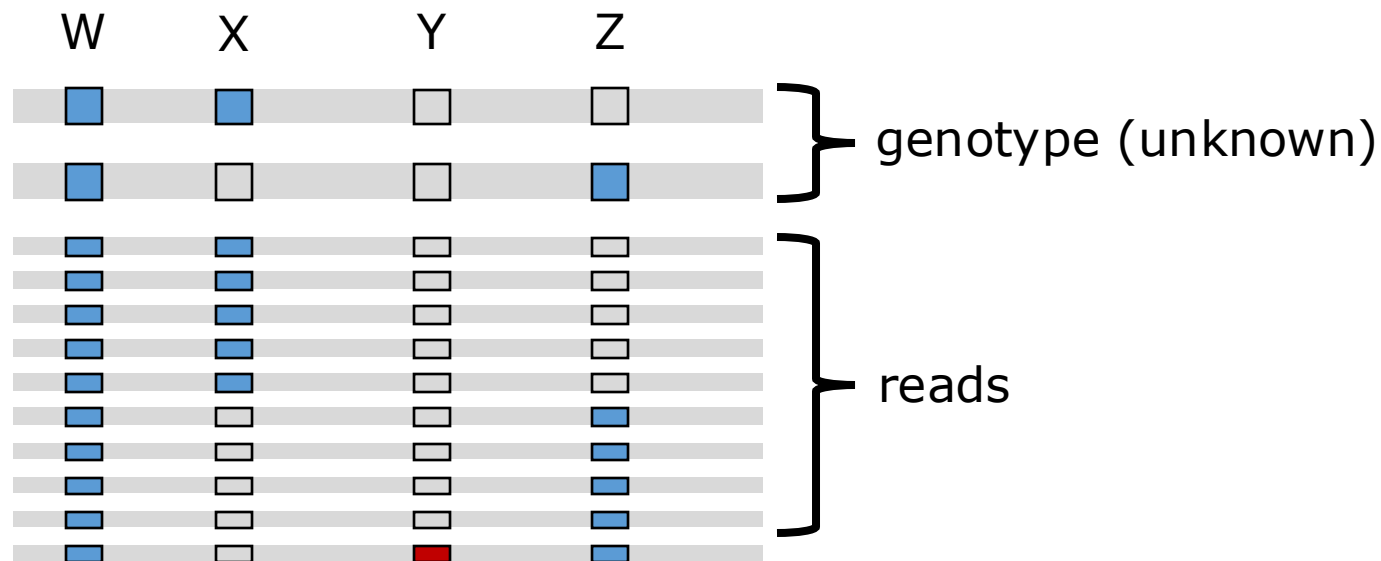
What are the alleles?



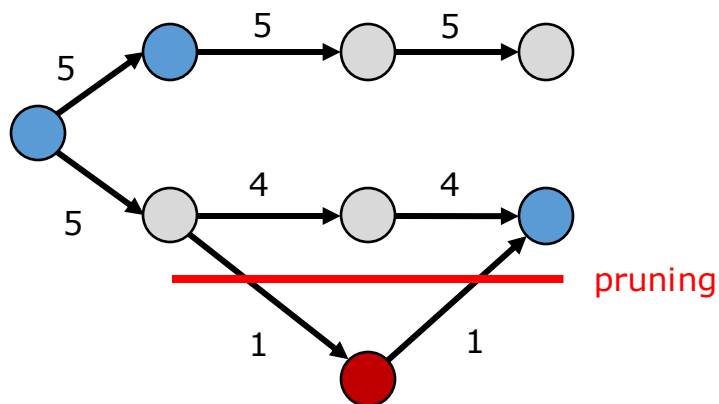
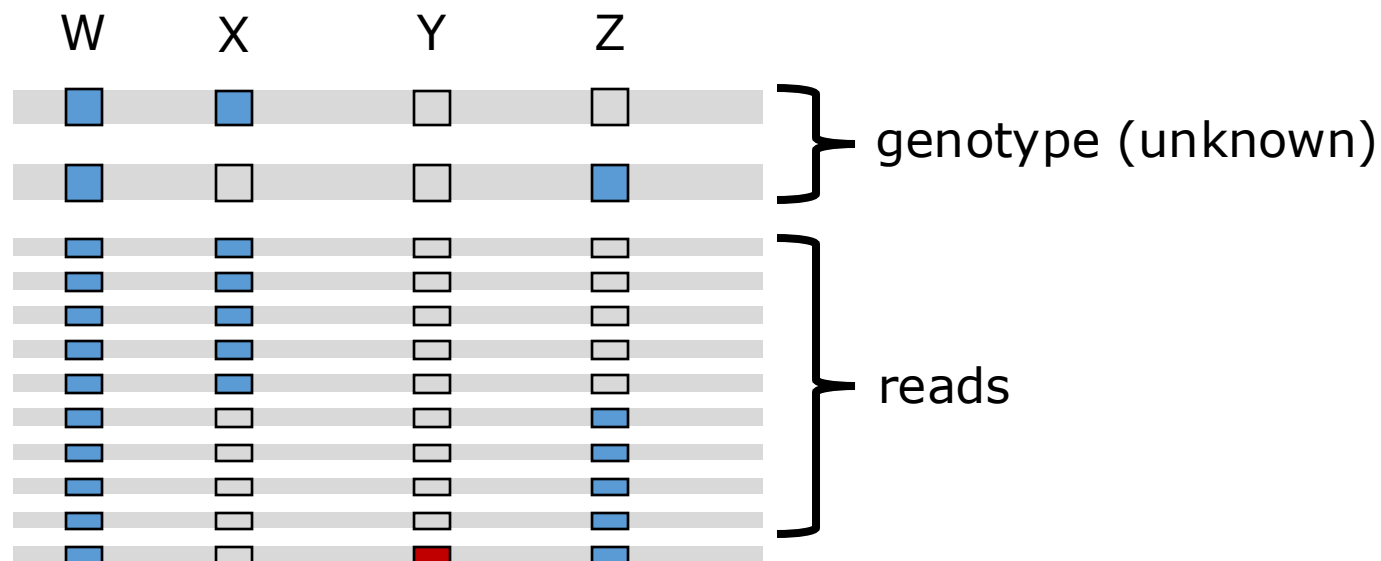
What are the alleles?



What are the alleles?



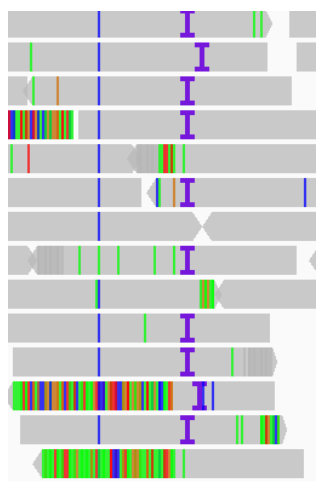
What are the alleles?



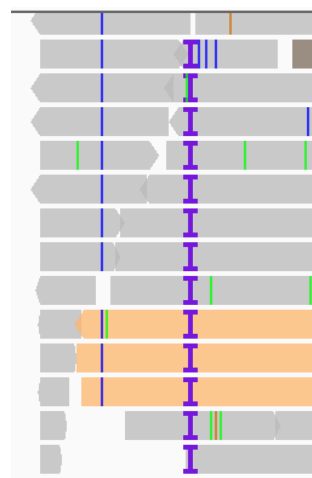
What are the alleles?

- Indel realignment
- Expensive process, but only on 'active' regions

bwa alignment



re-aligned



What is the genotype?

At a site we count 9 bases

5 REF and 4 ALT

$\Pr(X=4) = 0.25$ if heterozygous

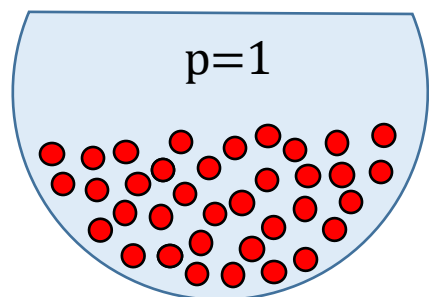
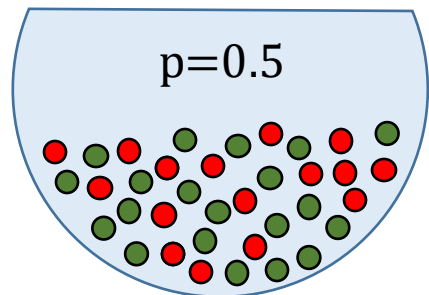
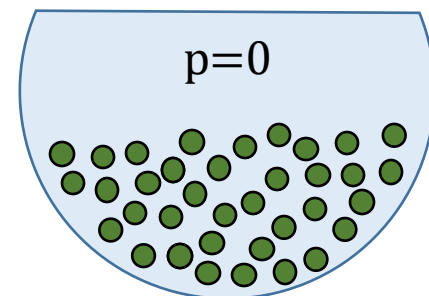
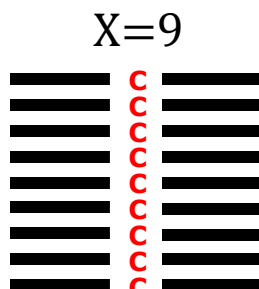
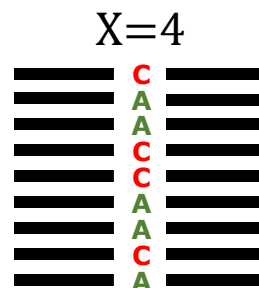
so: $\mathcal{L}(p=0.5 \mid X=4) = 0.25$

0 REF and 9 ALT

$\mathcal{L}(p=0.5 \mid X=9) = 0.002$

$\mathcal{L}(p=1 \mid X=9) = 1$

$\mathcal{L}(p=0 \mid X=9) = 0$



Question 10

Estimating genotype

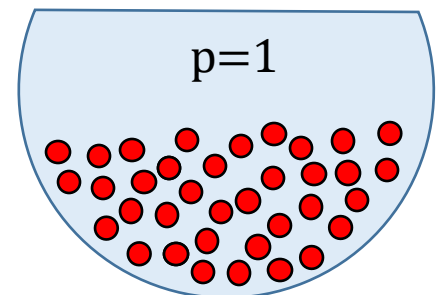
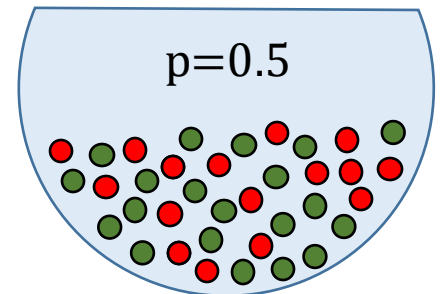
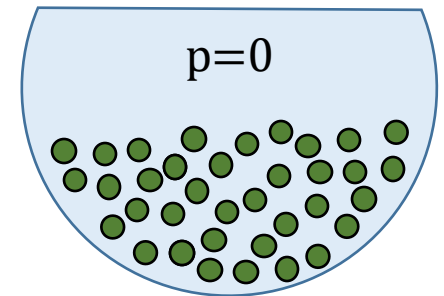
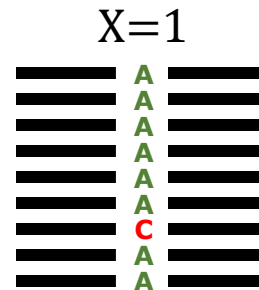
What are the likely genotypes?

At a site we count 9 bases

8 REF and 1 ALT

$$\mathcal{L}(p=0.5 \mid X=1) = 0.017$$

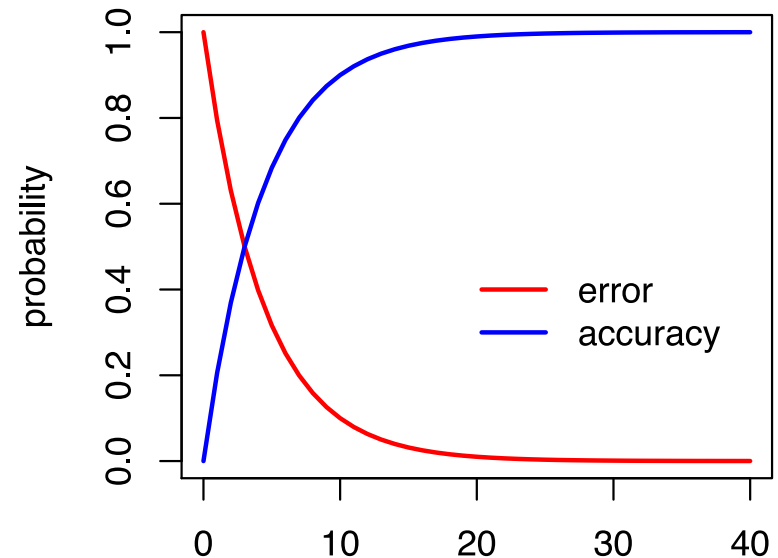
$$\mathcal{L}(p=0 \mid X=1) = 0$$



Strict binomial distribution
would only work with error-free
data

Base quality and error

- Base quality: 20 = error probability 0.01
- 100 samples with 40x coverage
- In total 40 errors expected



Estimating the genotype

Genotype likelihood (simplified):

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^l \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^k \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$

g: genotype (i.e. 0, 1 or 2)

m: ploidy (2 for human)

ϵ : base error

k: number of bases at the site

l: number of bases that equal reference

In GATK:

$$PL = -10 \cdot \log_{10}(\mathcal{L}(g))$$

PL and GQ

Our example: 8 REF and 1 ALT

Assuming base error probability $\epsilon = 0.01$

$$PL = -10 \cdot \log_{10}(\mathcal{L}(g))$$

Genotype	HomRef	Heterozygous	HomAlt
$\mathcal{L}(g)$	0.0092	0.0020	9.9E-17
PL	20	27	160

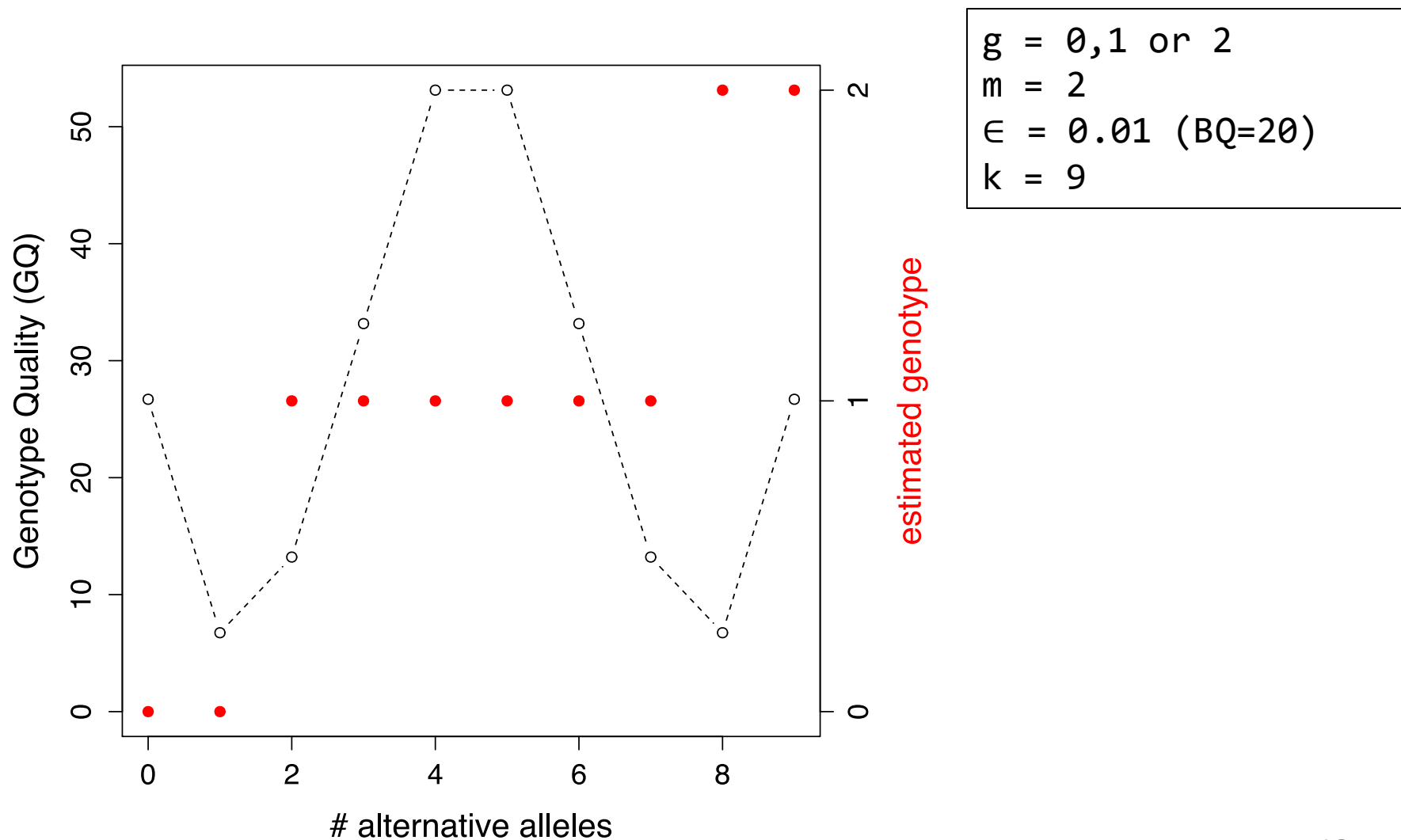
Lowest PL = most likely genotype

$$GQ = \text{Second lowest PL} - \text{Lowest PL} = 27 - 20 = 7$$

$$p(\text{genotype error}) = 10^{\frac{-7}{10}} = 0.2$$

Question 11

Estimating the genotype

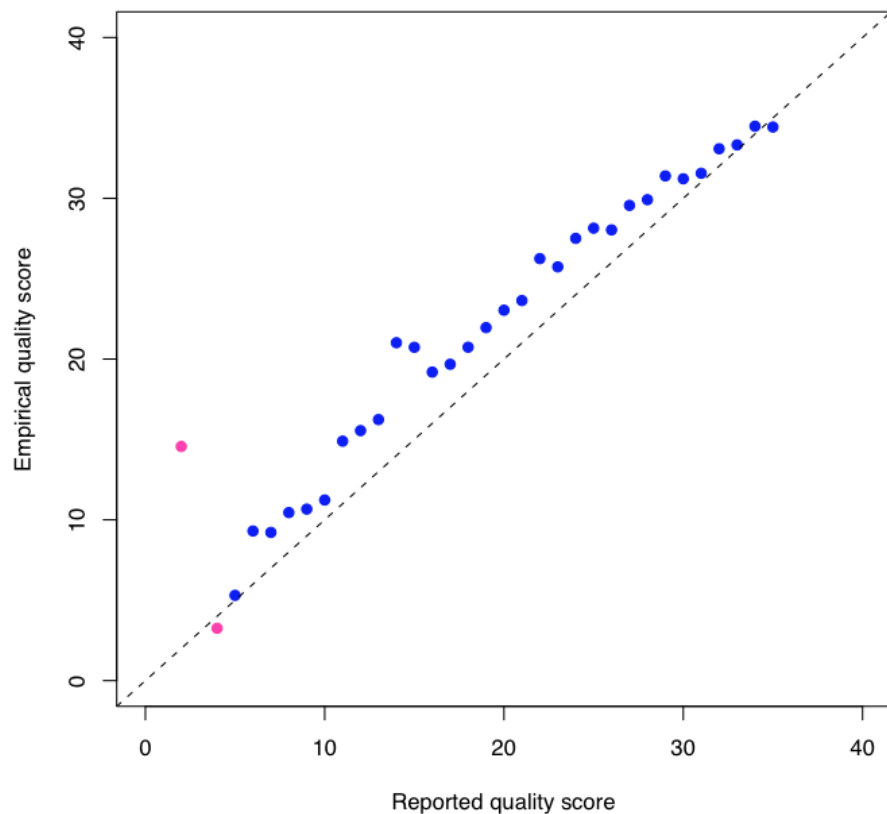


Base quality correction

- Essential for estimating genotype likelihood
- Context can affect base quality, e.g.:
 - homopolymers
 - cycle
- estimated error rate \neq 'real' error rate
- Base quality score recalibration (BQSR) takes this context into account

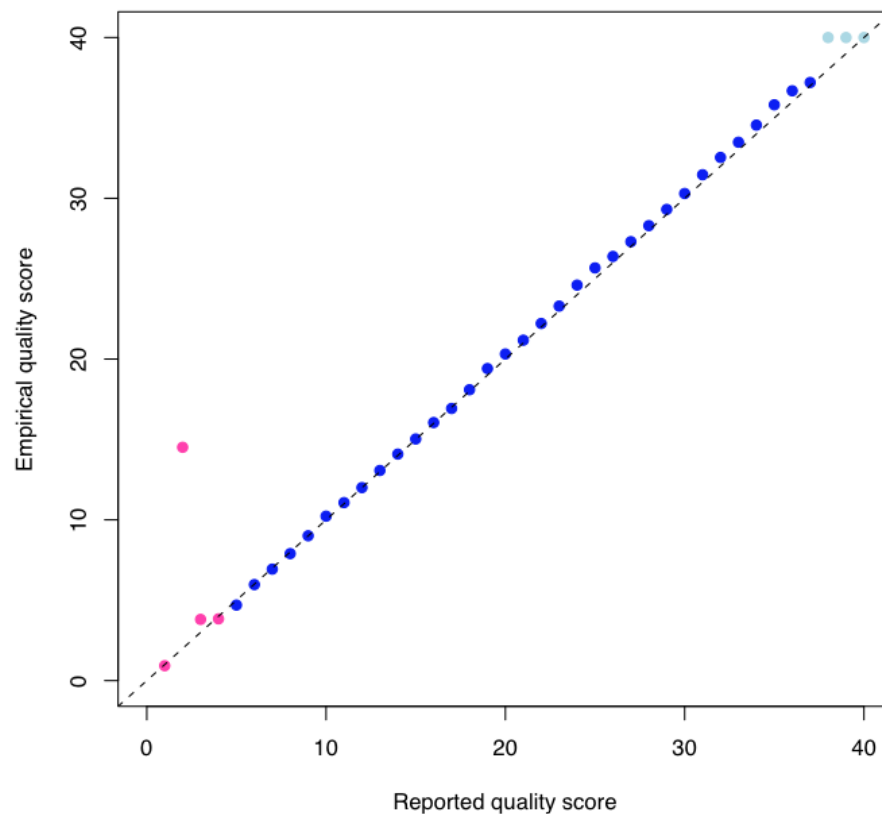
BQSR

RMSE = 1.221



Before BQSR

RMSE_good = 0.599 , RMSE_all = 0.599



After BQSR

vcf

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=PASS,Description="All filters passed">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specs">
##GATKCommandLine=<ID=GenotypeGVCFs,CommandLine="GenotypeGVCFs --output
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller --bam-output
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##contig=<ID=chr20,length=64444167>
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT father mother
chr20 10019252 . G C 134.68 . AC=1;AF=0.167;AN=6;DP=15 GT:AD:DP:GQ:PL 0/1:3,5:8:58:143,0,58 0/0:2,0:2:6:0,6,48
chr20 10019348 . A ACT 1587.89 . AC=5;AF=0.833;AN=6;DP=45 GT:AD:DP:GQ:PL 0/1:7,6:13:99:231,0,256 1/1:0,13:13:39:573,39,0
chr20 10019469 . C T 1792.98 . AC=4;AF=0.667;AN=6;DP=89 GT:AD:DP:GQ:PL 0/1:17,15:32:99:465,0,503 0/1:11,12:23:99:289,0,289
```

vcf

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=PASS,Description="All filters passed">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specs">
##GATKCommandLine=<ID=GenotypeGVCFs,CommandLine="GenotypeGVCFs --output
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller --bam-output
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##contig=<ID=chr20,length=64444167>
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT father mother
chr20 10019252 . G C 134.68 . AC=1;AF=0.167;AN=6;DP=15 GT:AD:DP:GQ:PL 0/1:3,5:8:58:143,0,58 0/0:2,0:2:6:0,6,48
chr20 10019348 . A ACT 1587.89 . AC=5;AF=0.833;AN=6;DP=45 GT:AD:DP:GQ:PL 0/1:7,6:13:99:231,0,256 1/1:0,13:13:39:573,39,0
chr20 10019469 . C T 1792.98 . AC=4;AF=0.667;AN=6;DP=89 GT:AD:DP:GQ:PL 0/1:17,15:32:99:465,0,503 0/1:11,12:23:99:289,0,289
```



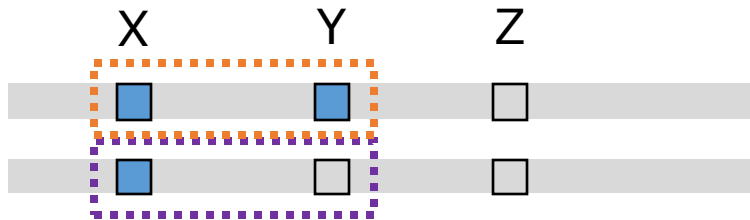
vcf

#CHROM	POS	ID	REF	ALT	FORMAT	NA00001	NA00002
20	14370	.	G	A	GT:GQ	0 0:48	1 0:48
20	17330	.	T	A	GT:GQ	0 0:49	0 1:99
20	1110696	.	A	G,T	GT:GQ	1 2:21	2 1:27
20	1230237	.	T	.	GT:GQ	0 0:54	0 0:48
20	1234567	.	GTC	G,GTCT	GT:GQ	0/1:35	0/2:17

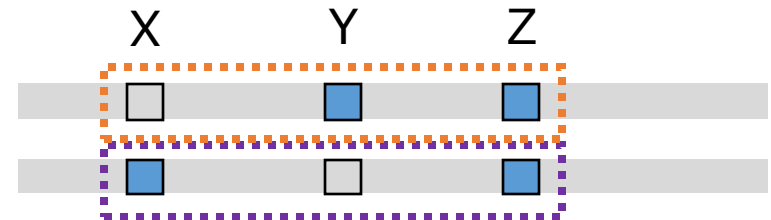
↑ ↑ ↑
0 1 2

Question 12

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 13

Missing genotype problem

- Most variant callers genotype all samples in one go. But:
 - variant calling process can become very computationally 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
- **DeepVariant**: variant calling in short and long reads

GATK workflow

