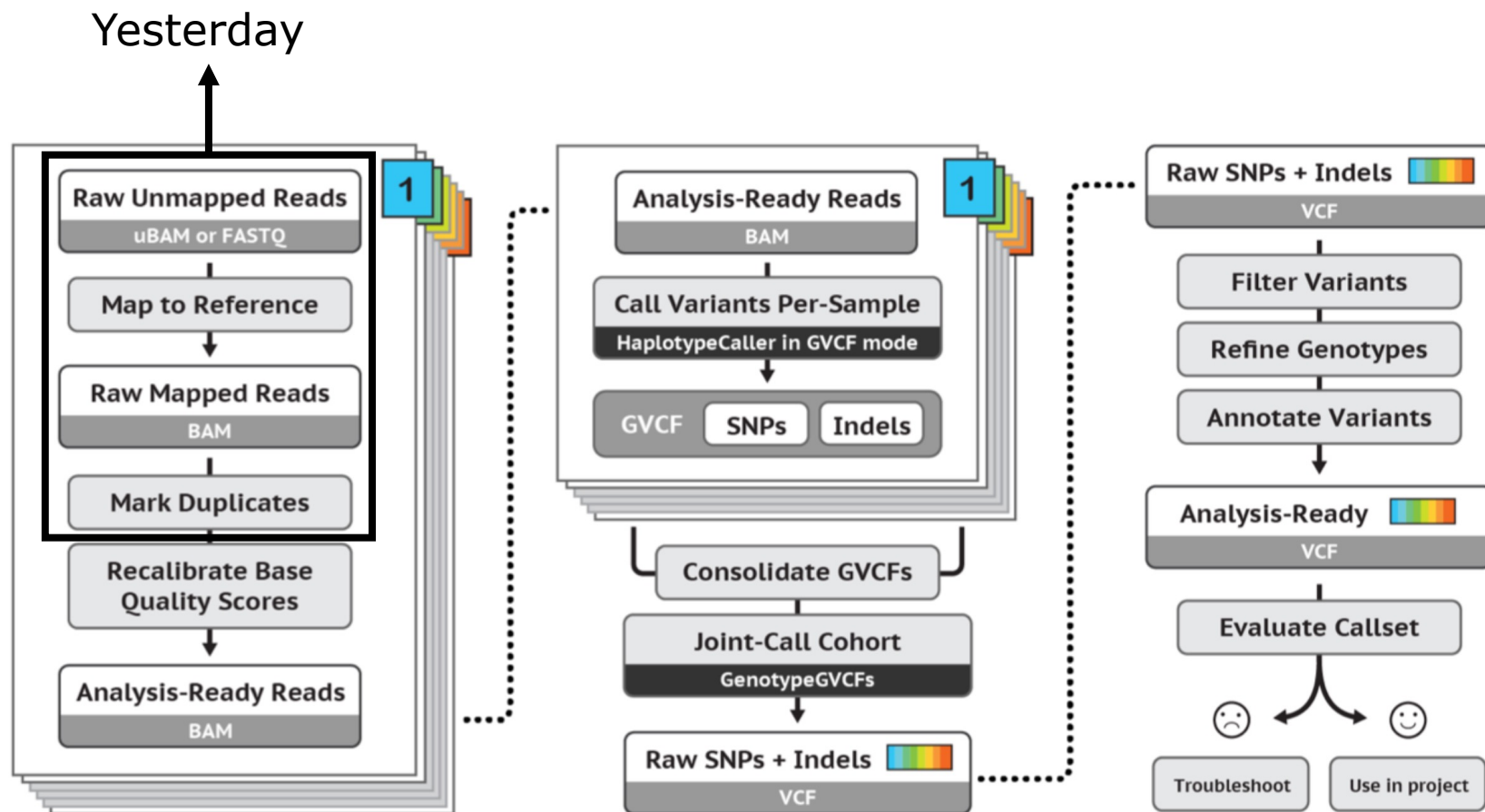
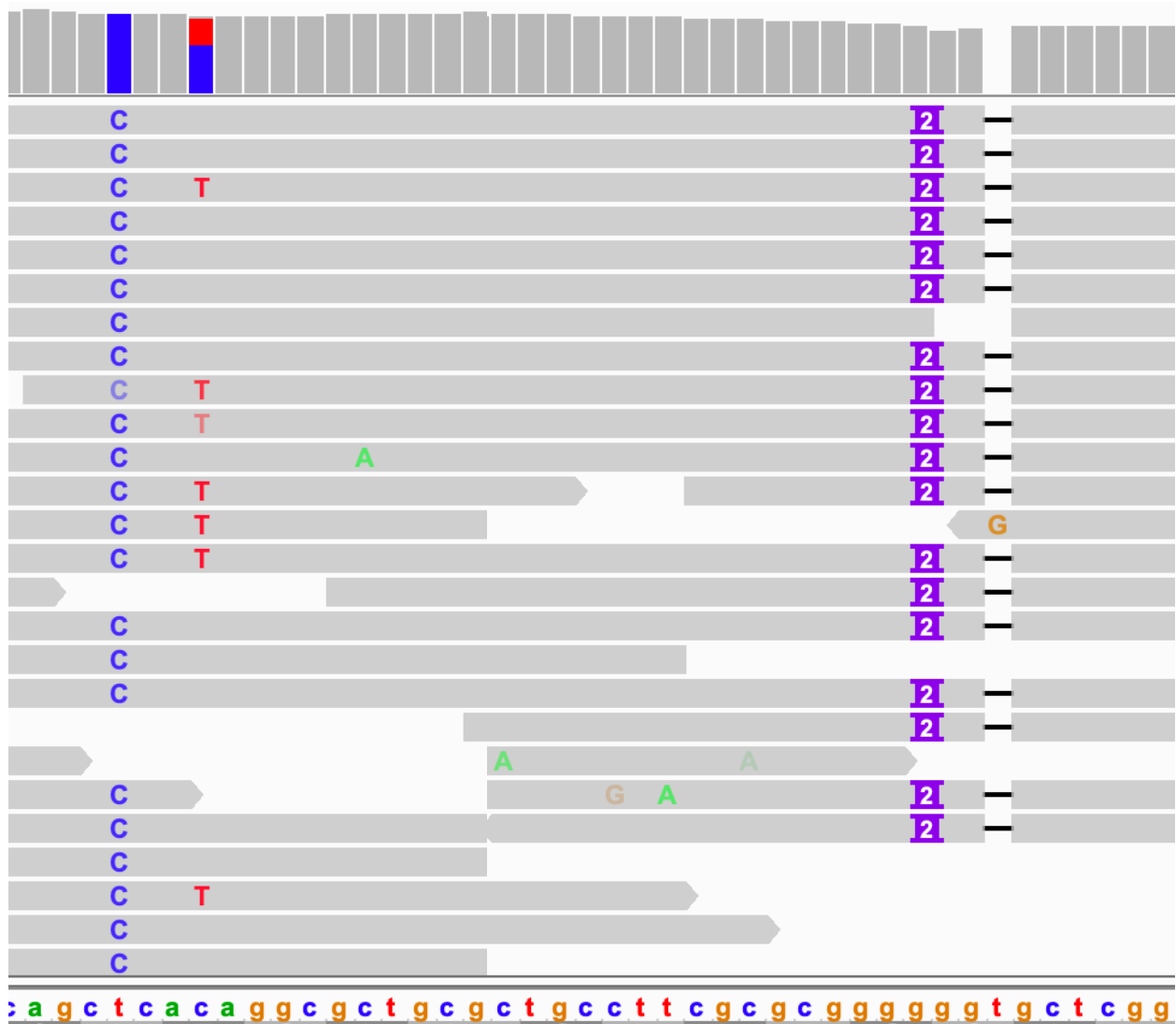


NGS – variant analysis

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

Estimating genotype

What are the likely genotypes?

At site X in sample Y we count 9 bases:

5 REF and 4 ALT

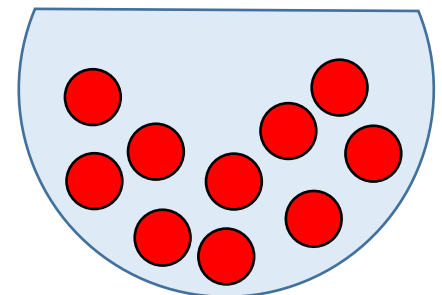
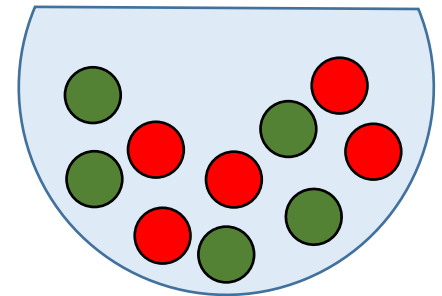
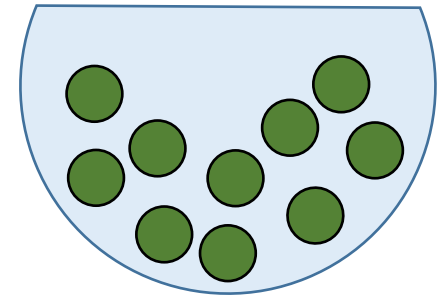
$\Pr(X=4) = 0.25$ if heterozygous

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

0 REF and 9 ALT

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

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



Quiz Question 6

Estimating genotype

What are the likely genotypes?:

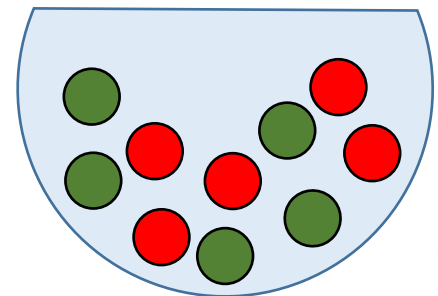
At site X in sample Y we count 9 bases:

8 REF and 1 ALT

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

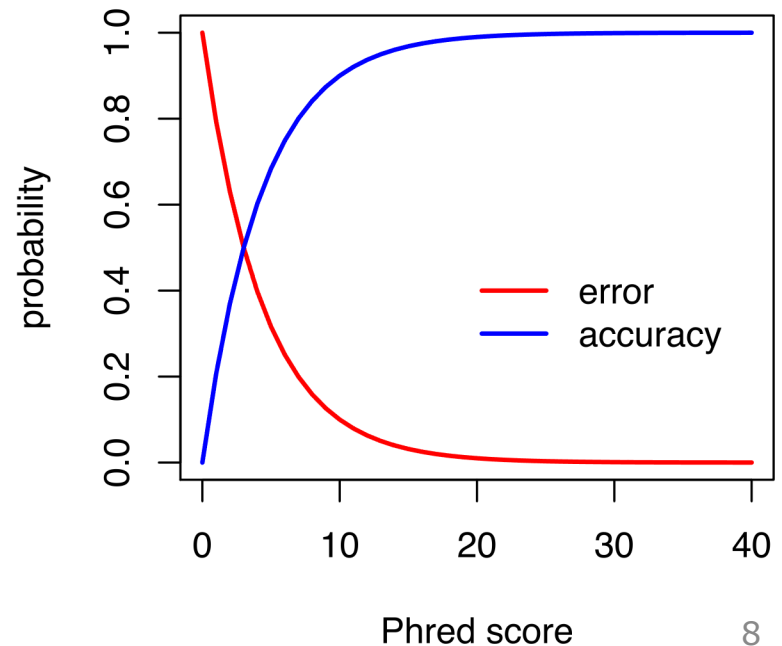
$$\mathcal{L}(p=0 \mid X=1) = 0$$

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 $\epsilon = 0.01$

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

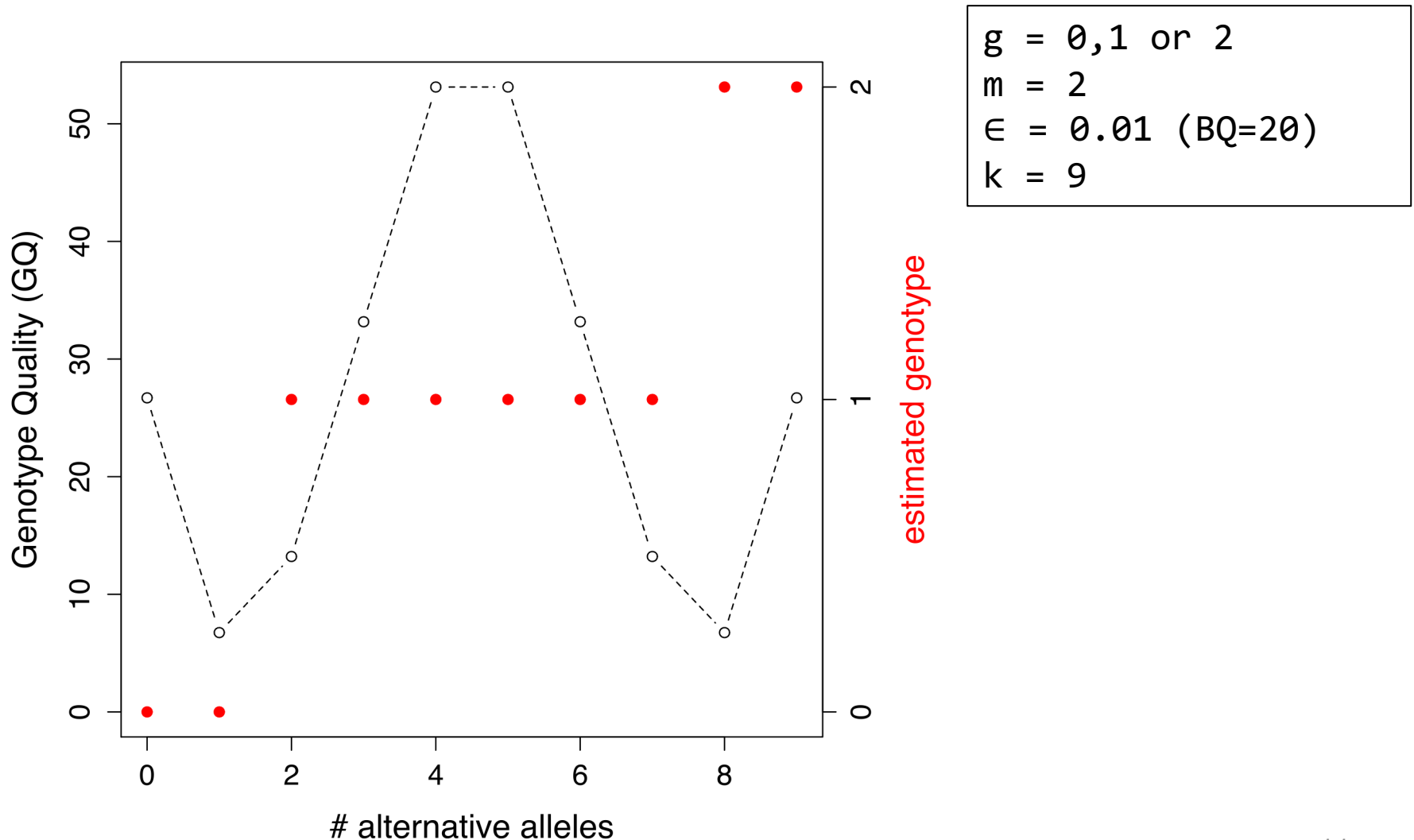
Genotype	HomRef	Heterozygous	HomAlt
$\mathcal{L}(g)$	0.0092	0.0020	9.9E-17
PL	20	27	160

Lowest PL = most likely genotype

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

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

Estimating the genotype

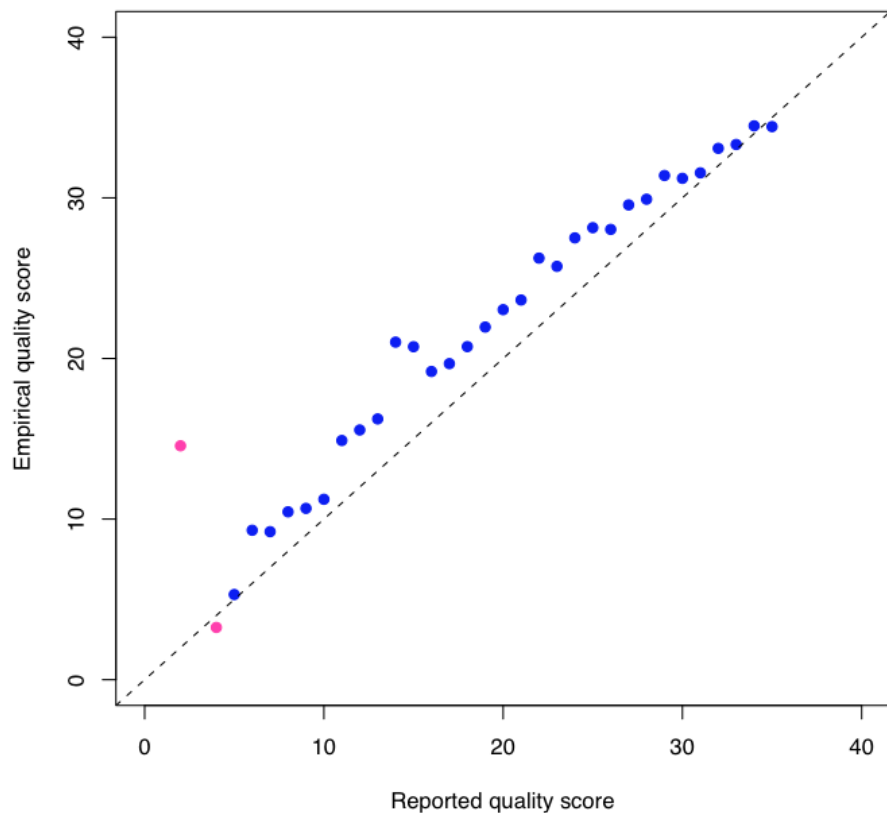


Base quality correction

- 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

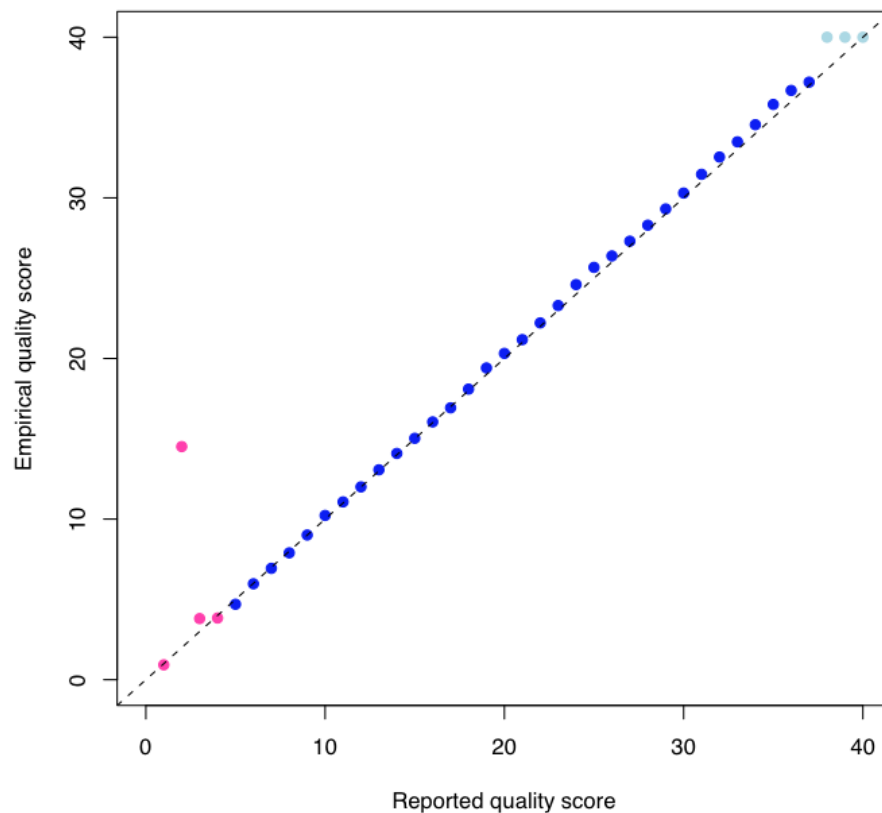
BQSR

RMSE = 1.221



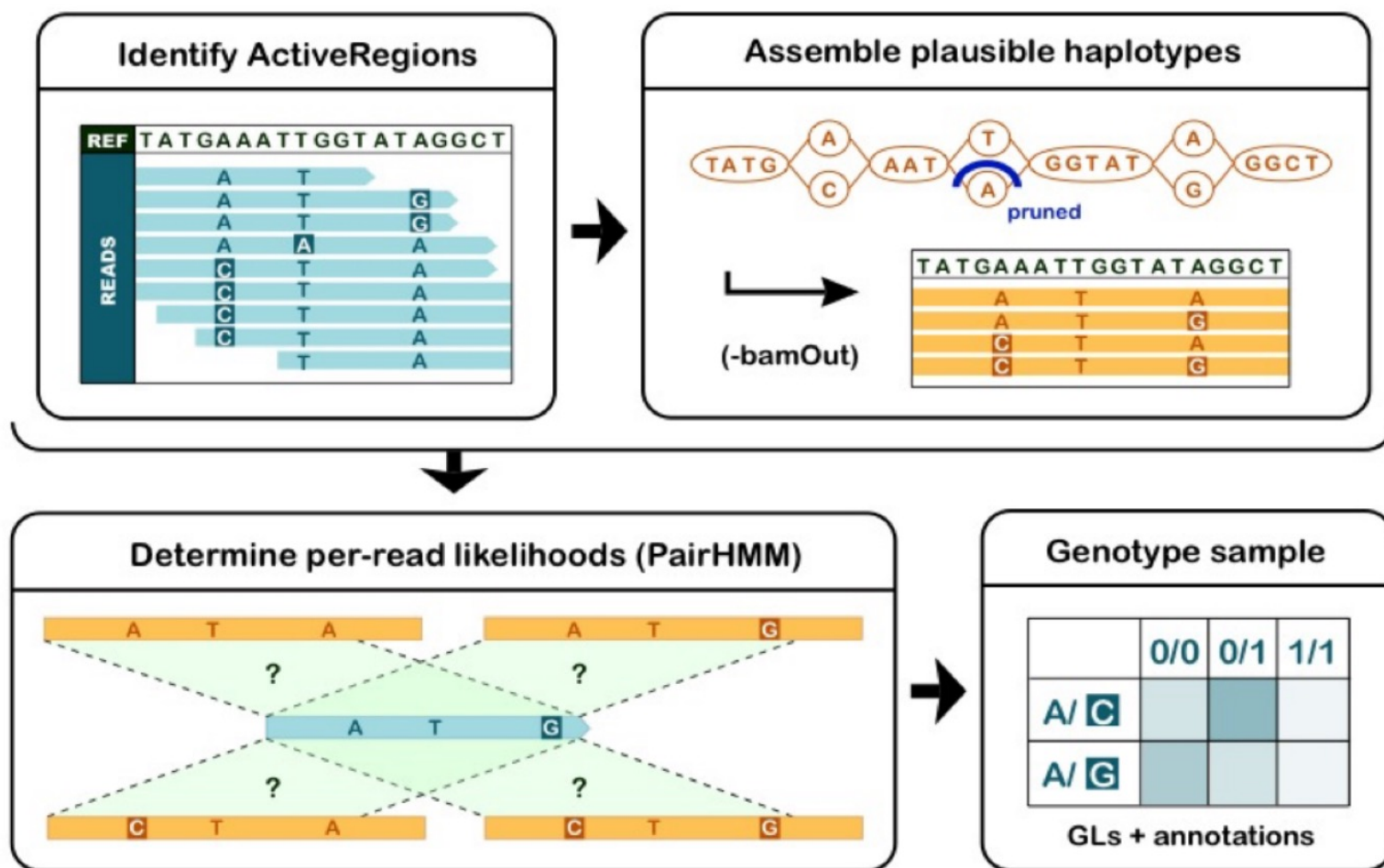
Before BQSR

RMSE_good = 0.599 , RMSE_all = 0.599

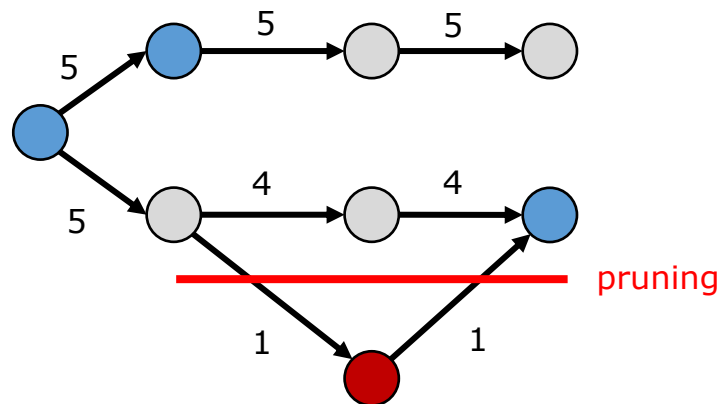
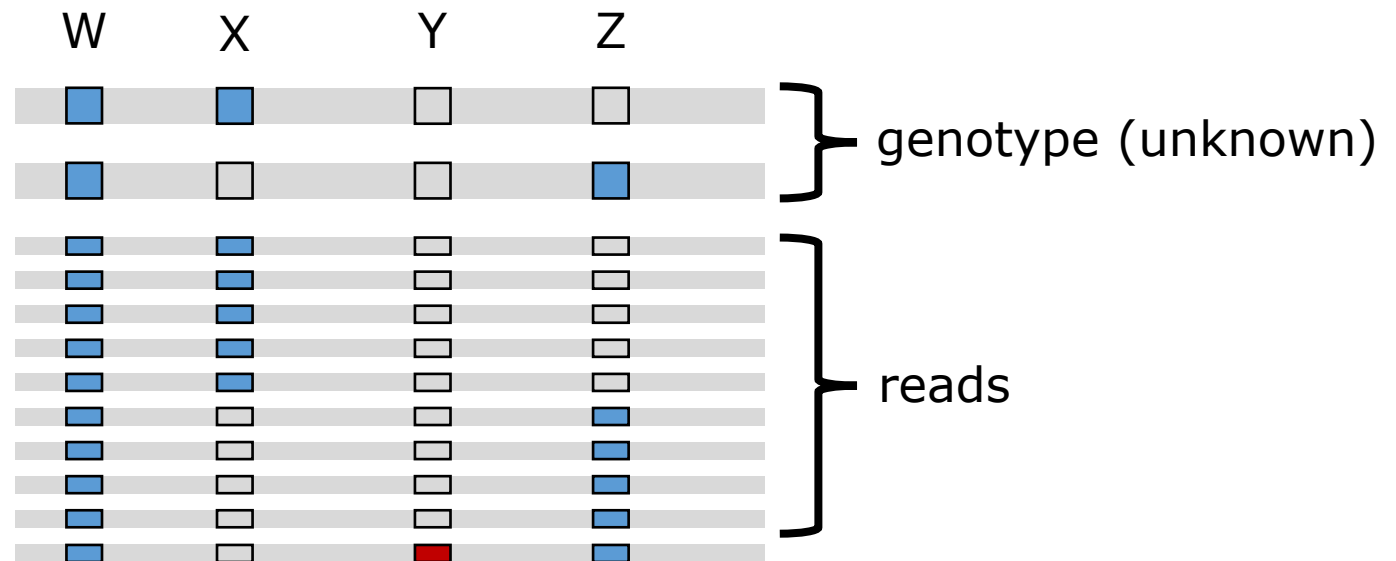


After BQSR

HaplotypeCaller



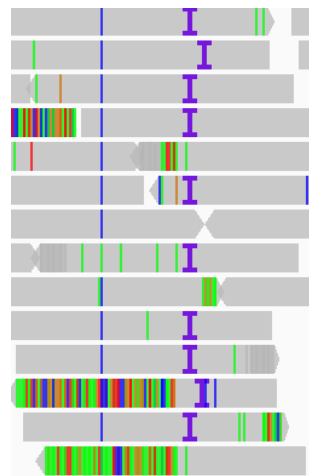
HaplotypeCaller



HaplotypeCaller

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

bwa alignment



re-aligned



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">
```

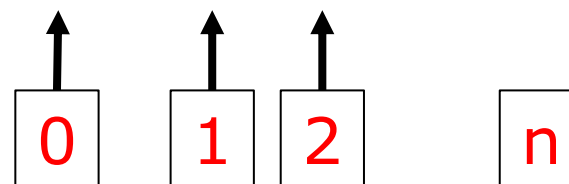
```
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA000001	NA000002
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

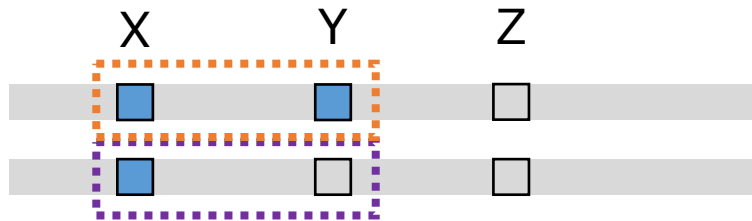
Quiz Question 7

#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

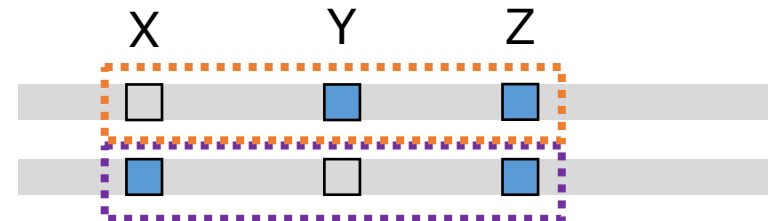
sample 1



sample1.vcf

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

sample 2



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

Quiz Question 8

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
- **DeepVariant**: variant calling in long reads

GATK workflow

