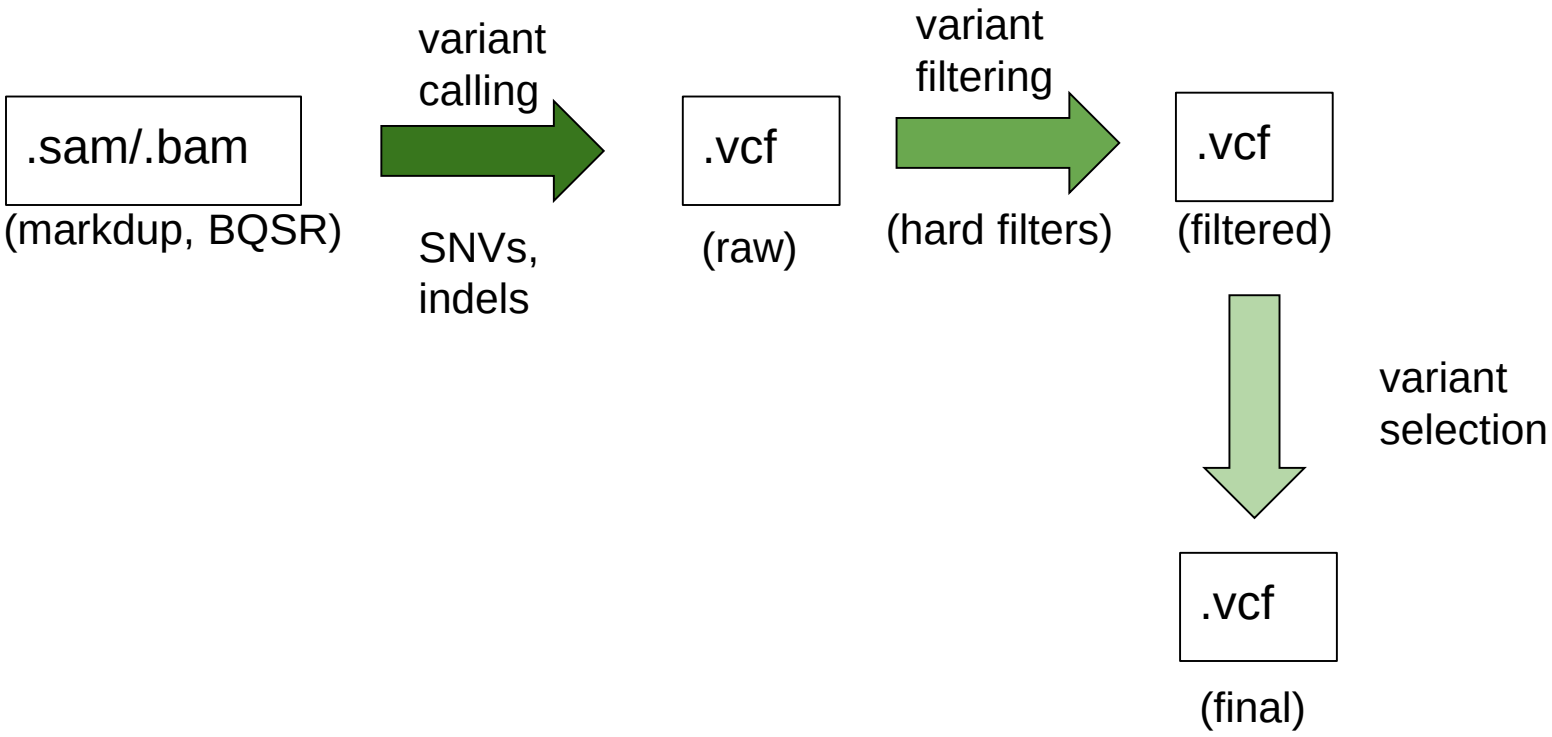
Chamada, filtragem e avaliação de qualidade de variantes

José Ferrão e Hugo Martiniano

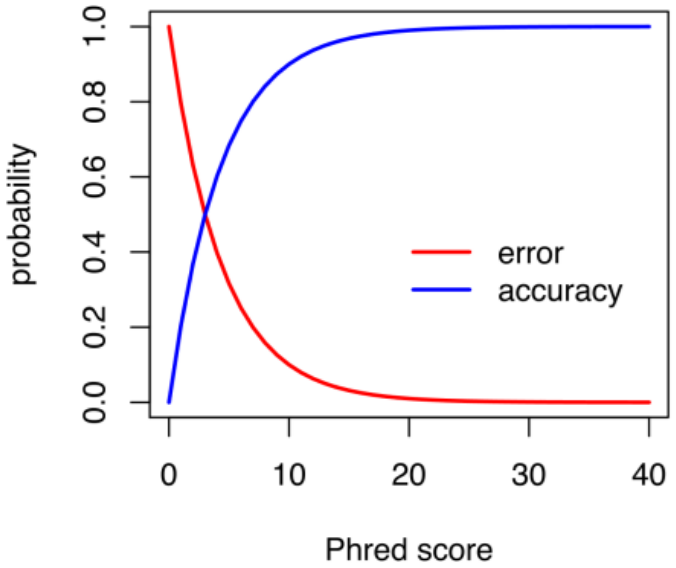
Variant calling

- Variant calling is the process through which variants are identified from aligned reads.



Base quality and error

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



Calculation of PL and GQ by HaplotypeCaller

P(Genotype | Data) is the conditional probability of the Genotype given the sequence Data that we have observed.

Alleles
Reference: A
Read: T

Conditional probabilities calculated by HC

P(AA | Data) = 0.000001
P(AT | Data) = 0.000100
P(TT | Data) = 0.010000

Genotype	A/A	A/T	T/T
Raw PL	-10 * log(0.000001) = 60	-10 * log(0.000100) = 40	-10 * log(0.010000) = 20

PL is the **P**hred-scaled **L**ikelihood of the genotype

PL: the probability that the genotype is not correct. In other words, low PL values mean a genotype is more likely, and high PL values means it's less likely.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890451-Calculation-of-PL-and-GQ-by-HaplotypeCaller-and-GenotypeGVCFs

Calculation of PL and GQ by HaplotypeCaller

PL is the **P**hred-scaled **L**ikelihood of the genotype

PL: the probability that the genotype is not correct. In other words, low PL values mean a genotype is more likely, and high PL values means it's less likely.

Genotype	A/A	A/T	T/T
Normalized PL	60 - 20 = 40	40 - 20 = 20	20 - 20 = 0

GQ: **G**enotype **Q**uality

The value of GQ is simply the difference between the second lowest PL and the lowest PL (which is always 0)

https://gatk.broadinstitute.org/hc/en-us/articles/360035890451-Calculation-of-PL-and-GQ-by-HaplotypeCaller-and-GenotypeGVCFs

Estimating the genotype

Genotype likelihood (simplified):

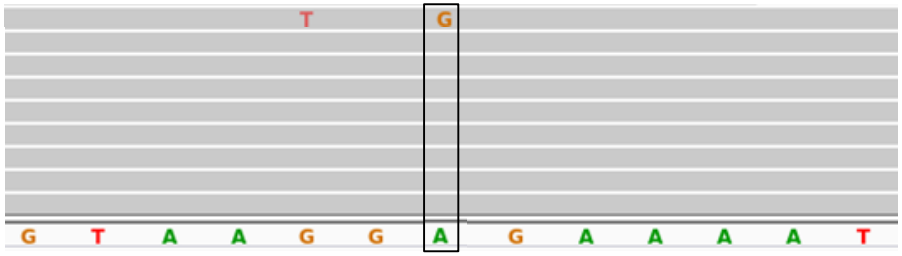
P(Genotype | Data) ⇨

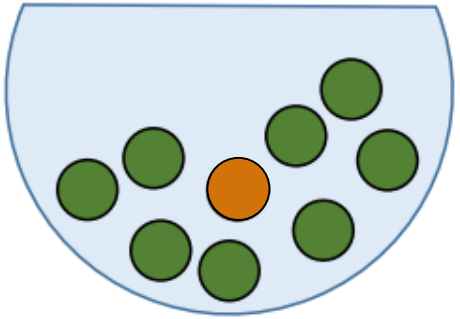
$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^{k} \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$

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:
$PL = -10*\log10(\mathcal{L}(g))$

Li H. Bioinformatics. 2011;27:2987–93.





PL and GQ

Our example: 8 REF and 1 ALT

Assuming base error probability $\epsilon = 0.01$

$PL = -10*\log10(\mathcal{L}(g))$

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

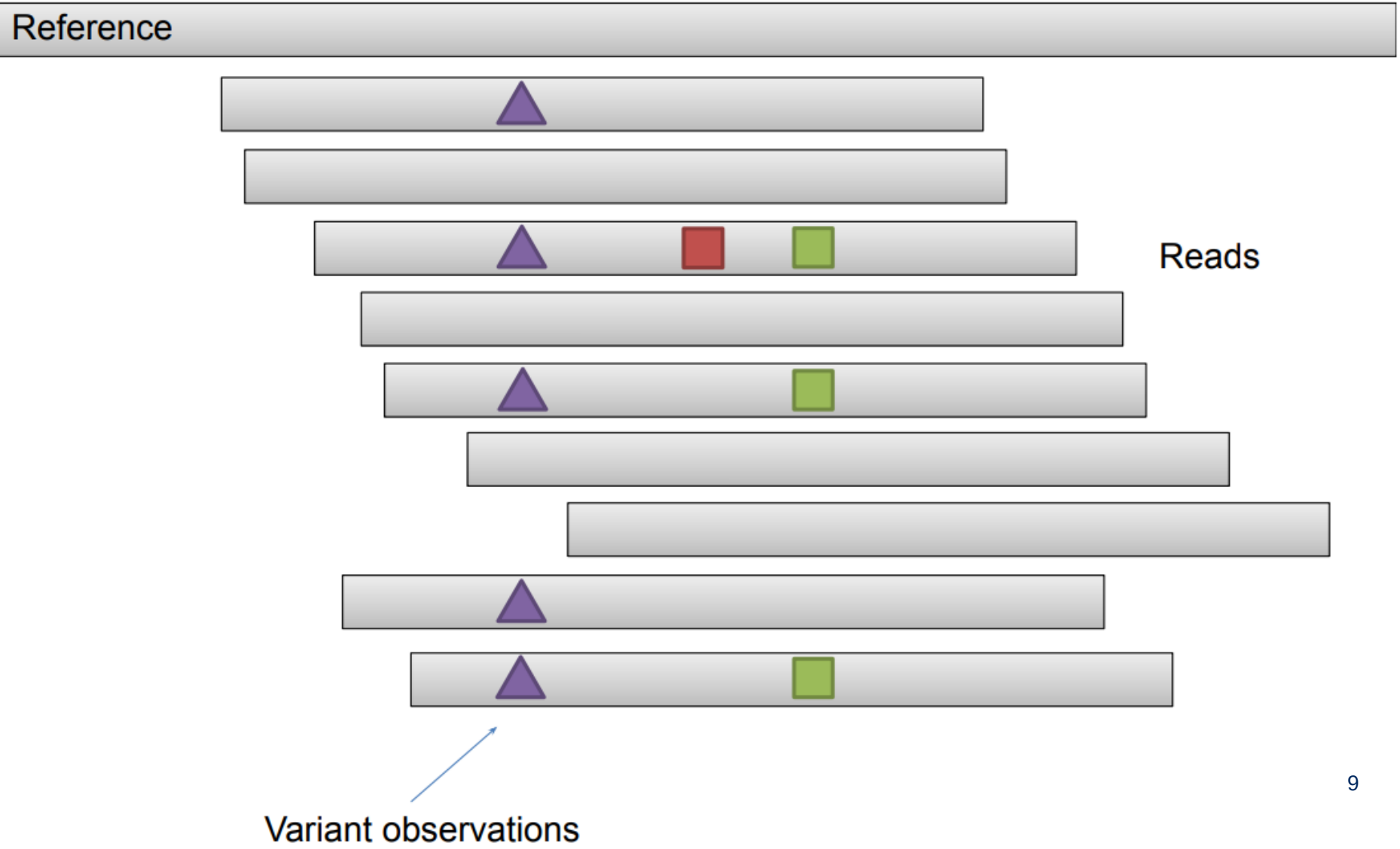
Lowest PL = most likely genotype

GQ = Second lowest PL – Lowest PL = 27 -20 = 7

Base quality correction

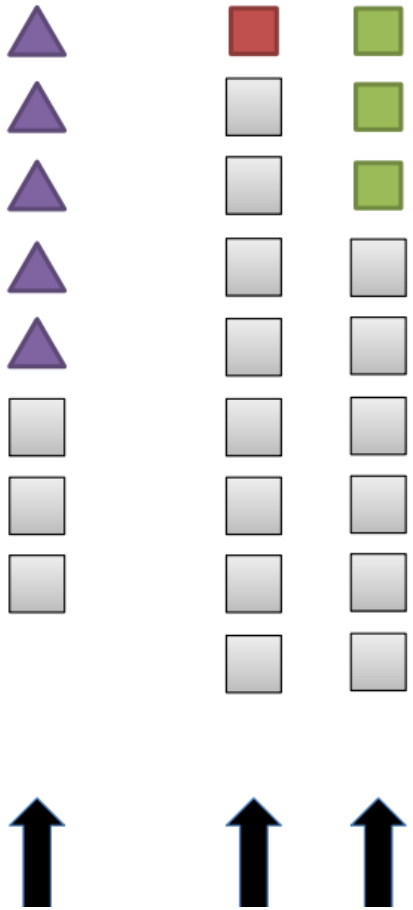
- Essential for estimating genotype likelihood
- Context can affect base quality, e.g.:
  - homopolymers
  - cycle
- estimated error rate ≠ 'real' error rate
- Base quality score recalibration (BQSR) takes this context into account

Alignments to candidates

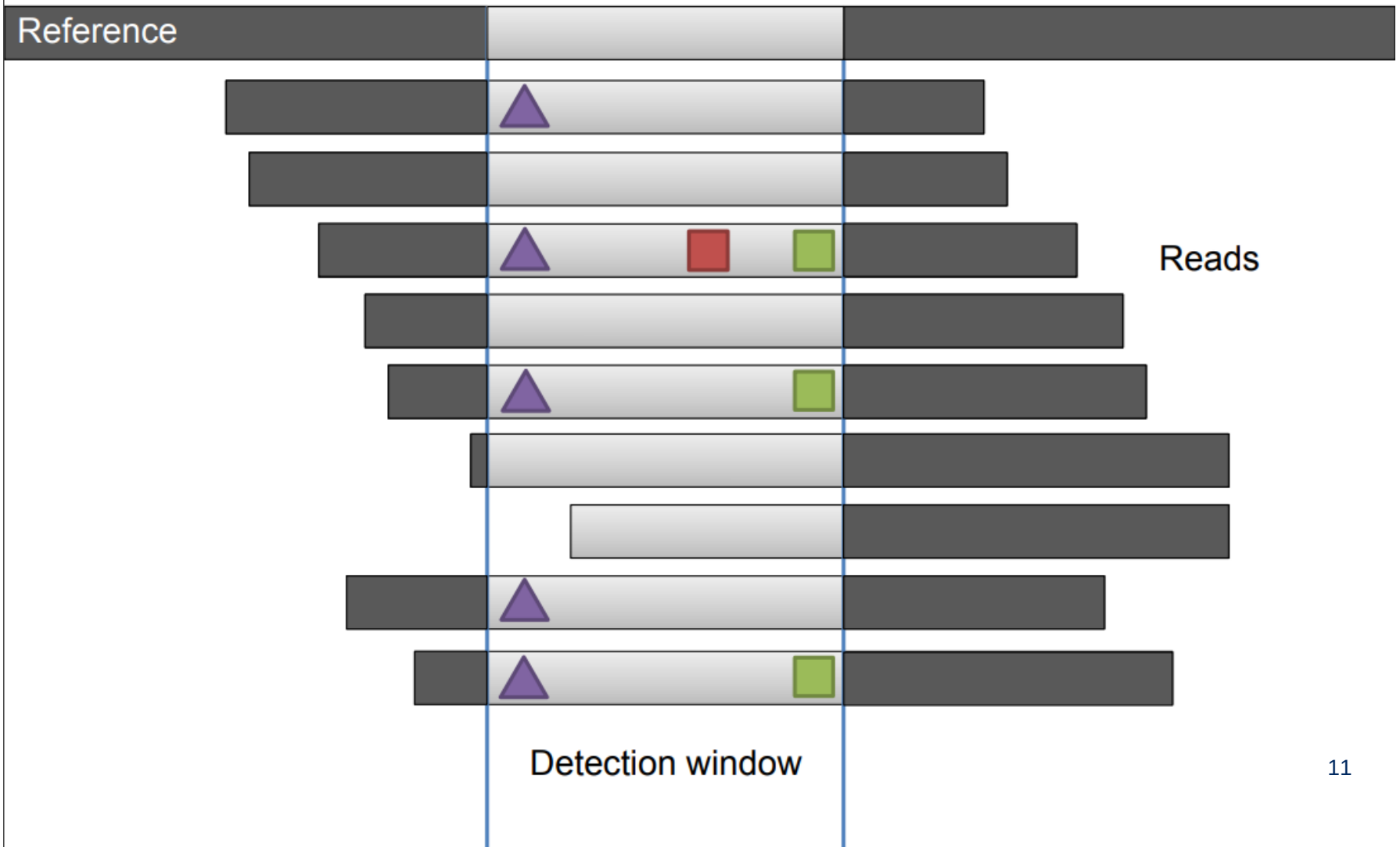


The data exposed to the caller

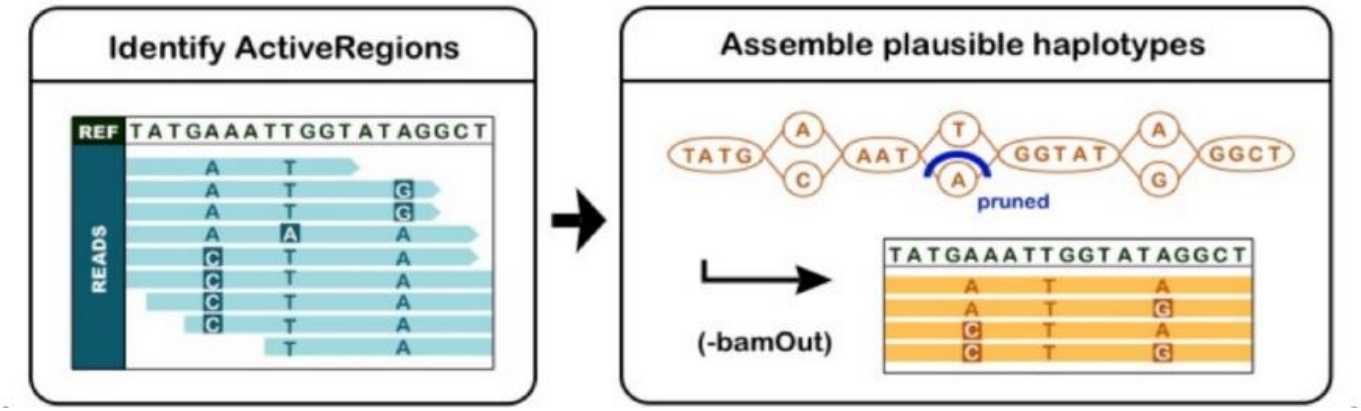
Reference



Direct detection of haplotypes



HaplotypeCaller

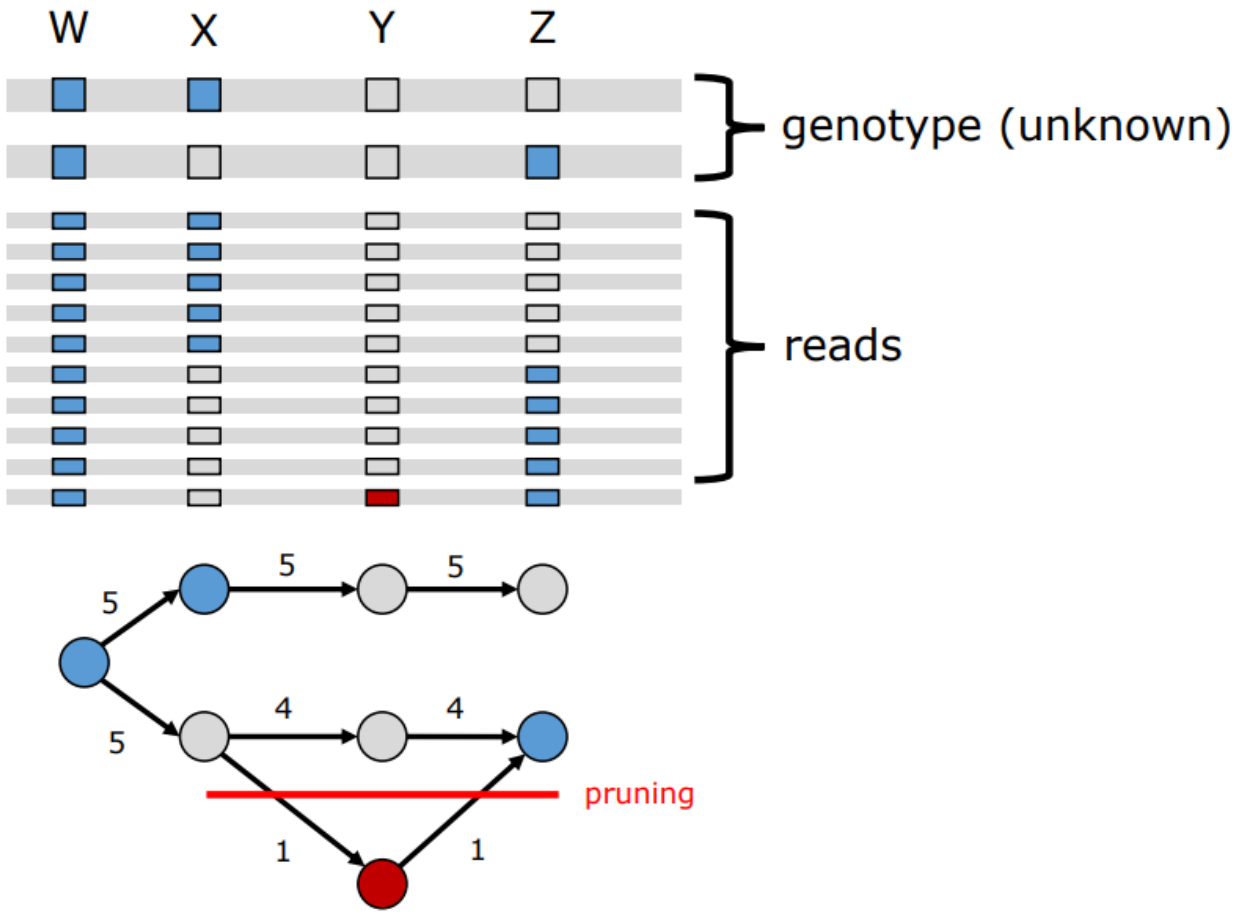


identifies what are the possible haplotypes present in the data

identifies potentially variant sites

Poplin R et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017;1–22.

HaplotypeCaller



vcf

Header
lines starting with ##: arbitrary number of meta-information lines
line starting with #: column definition – mandatory columns include:

CHROM	chromosome
POS	position of the start of the variant
ID	unique identifier of the variant (e.g. rs number for SNPs)
REF	reference allele
ALT	comma separated list of alternate non-reference alleles
QUAL	phred-scaled quality score
FILTER	site filtering information
INFO	user extensible annotation (e.g. samtools and GATK may differ in this)

samples follow

Data
one line per site (all columns described above per line); useful information per site and per sample

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

reference allele (GT: 0)

alternative allele (GT: 1)

alternative allele (GT: 2)

vcf

#CHROM POS ID REF ALT QUAL FILTER INFO **FORMAT** sample1

4 1031852 . A T 9473.06 PASS AC=2;AF=1.00;AN=2;BaseQRankSum=2.038;DP=298;ExcessHet=3.0103;MQRankSum=0.000;RAW_MQandDP=1087200,302;ReadPosRankSum=-0.842 **GT:AD:DP:GQ:PL:SB** 1/1:4,293:298:99:9887,790,0:2,2,220,74

GT

0/0 - homozygous reference

0/1 - heterozygous

1/1 - homozygous alternative

variant calling – filtering variants

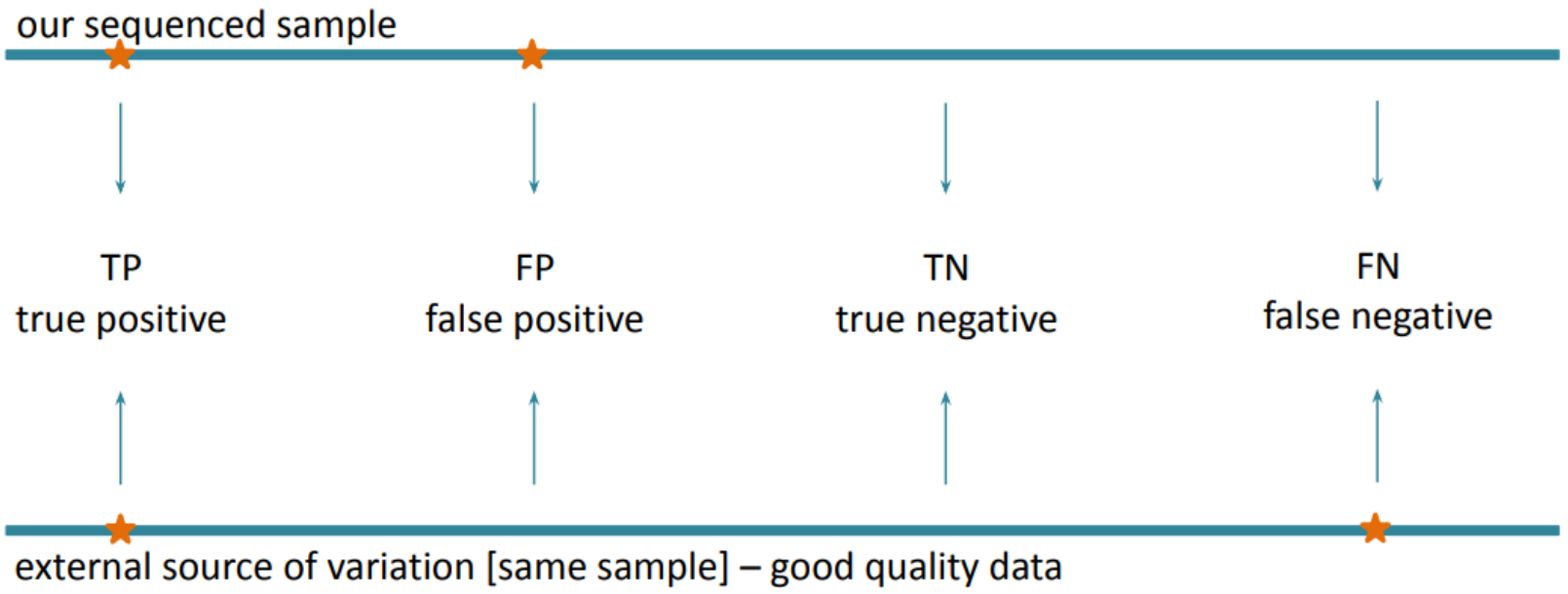
Common cautions:

- Base quality	BQ20
- Depth (min and max)	very dependent on your average
- Mapping quality	MQ50/60
- Strand-bias	p-value>0.05
- SNP density	dependent on the genome [e.g. no more than 1 SNP/4bp]
- QUAL	>20/25
- Genotype Quality (GQ)	>20/30

Keep in mind your project may have some specific requirements

Further reading: "Consensus Rules in Variant Detection from Next-Generation Sequencing Data" Jia et al 2012 PLoS One

variant calling – evaluating

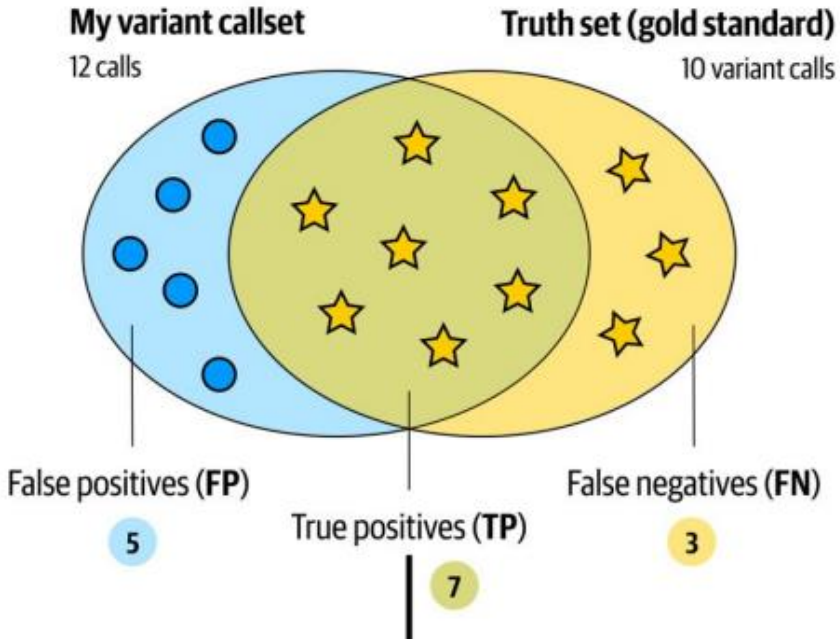


high specificity → low FP

high sensitivity → low FN

variant calling – evaluating

Site-level concordance with a truth set



Sensitivity (Recall)

$$\frac{TP}{TP + FN} = \frac{7}{7 + 3} = 70\%$$

Specificity (Precision)

$$\frac{TP}{TP + FP} = \frac{7}{7 + 5} = 58\%$$

DNA Sequencing – NGS (Germline)

PERFORMANCE CRITERIA

For the 2022, Performance Criteria have been applied to the results of this EQA[1].

The marking schema includes:

- NGS variant concordance using the F-score for single nucleotide polymorphisms (SNPs) only, located within the high-confidence (HC) regions of the genome.
- The performance outcome for this EQA is **Satisfactory** OR **Poor**. EMQN and GenQA staff will ensure consistency of scoring between and within the EQA rounds.

Poor performance is defined as follows:

- *Those participants having a submission with an F-score below 90% for SNPs within the high-confidence regions of the genome.*

Please note:

- 'High confidence' is defined as genomic regions exclusive of union of all tandem repeats, all homopolymers >6bp, all imperfect homopolymers >10bp, all difficult to map regions, all segmental duplications, GC <25% or >65%, "Bad Promoters", and "Other Difficult Regions" as published by NIST in Genome In A Bottle - Genome Stratifications (https://doi.org/10.18434/M32190).
- The F-score of indels (<50bp) is excluded from the current Performance Criteria

EMQN

SCHEME: DNA SEQUENCING - NGS (v Germline) **SEASON:** 2021

Variant call assessment

Region	all		high confidence	
Variant type	snp	indel	snp	indel
True positives	7423	216	5341	101
False positives	54	2	2	0
False negatives	503	171	223	10
Sensitivity	93.65%	55.81%	95.99%	90.99%
Precision	99.28%	99.08%	99.96%	100.00%
F-Score	96.38%	71.40%	**97.94%**	95.28%

SUMMARY OF YOUR PERFORMANCE IN THIS SCHEME

Assessment Category	Performance[3] (mean score)
Performance	2.00
Scheme result (SATISFACTORY or POOR)	SATISFACTORY

EMQN

SCHEME: DNA SEQUENCING - NGS (Germline) **SEASON:** 2022

Variant call assessment

Region	all		high confidence	
Variant type	snp	indel	snp	indel
True positives	7419	220	5343	88
False positives	147	48	40	4
False negatives	556	145	232	8
Sensitivity	93.03%	60.27%	95.84%	91.67%
Precision	98.06%	82.09%	99.26%	95.65%
F-Score	95.48%	69.51%	**97.52%**	93.62%

SUMMARY OF YOUR PERFORMANCE IN THIS SCHEME

Assessment Category	Performance[3] (mean score)
Genotyping	2.00
Scheme result (SATISFACTORY or POOR)	SATISFACTORY

Variant consensus analysis report (NGS EQA 2022) Germline.xlsx

Variant position	Type	Gene	Submitted genotype	EQA genotype	EQA consensus ratio	Classification	Notes	Region
1:1041950	snp	AGRN	C/C	C/C	50/50	Agree		low confidence
1:1046551	snp	AGRN	G/G	G/G	77/78	Agree		low confidence
1:1047614	snp	AGRN	C/C	C/C	78/78	Agree		high confidence
1:1048922	snp	AGRN	C/C	C/C	78/78	Agree		low confidence
1:1051820	snp	AGRN	T/T	T/T	54/54	Agree		low confidence
1:1054900	snp	-	T/T	T/T	78/78	Agree		low confidence
1:1212042	snp	TNFRSF4		T/T	72/73	Missing		low confidence
1:1334174	snp	DVL1	T/C	T/C	76/76	Agree		low confidence
1:1721589	snp	-	T/T	T/T	40/41	Agree		low confidence
1:2025598	snp	GABRD	T/C	T/C	77/77	Agree		high confidence