GATK Best Practices for Variant Discovery

Variant calling with HaplotypeCaller

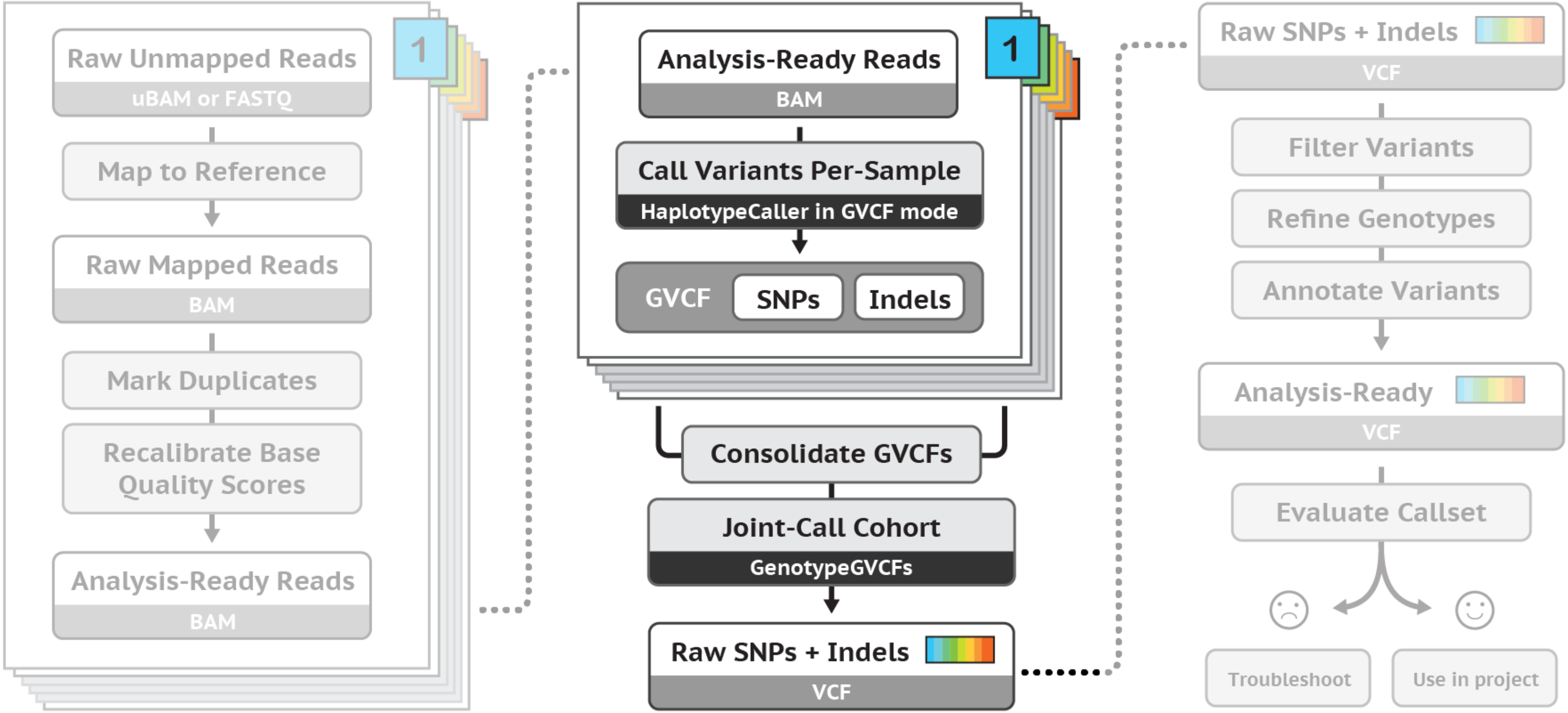
Basic operation and algorithm

BROAD INSTITUTE

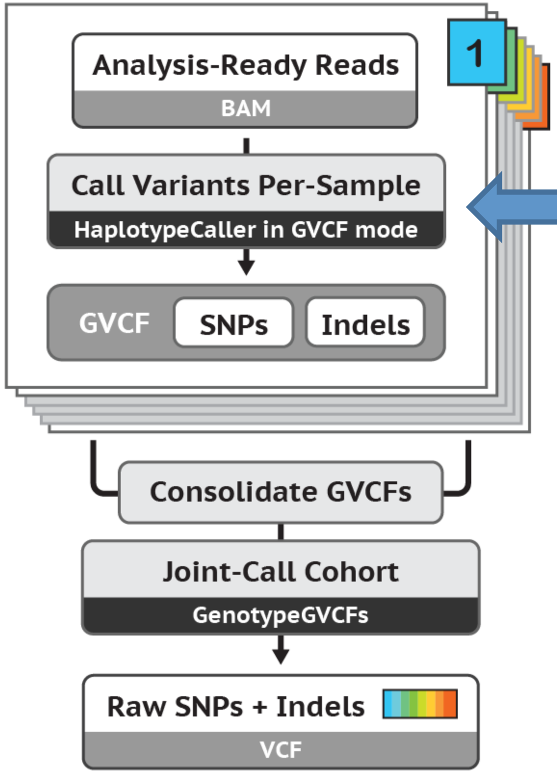
https://software.broadinstitute.org/gatk/

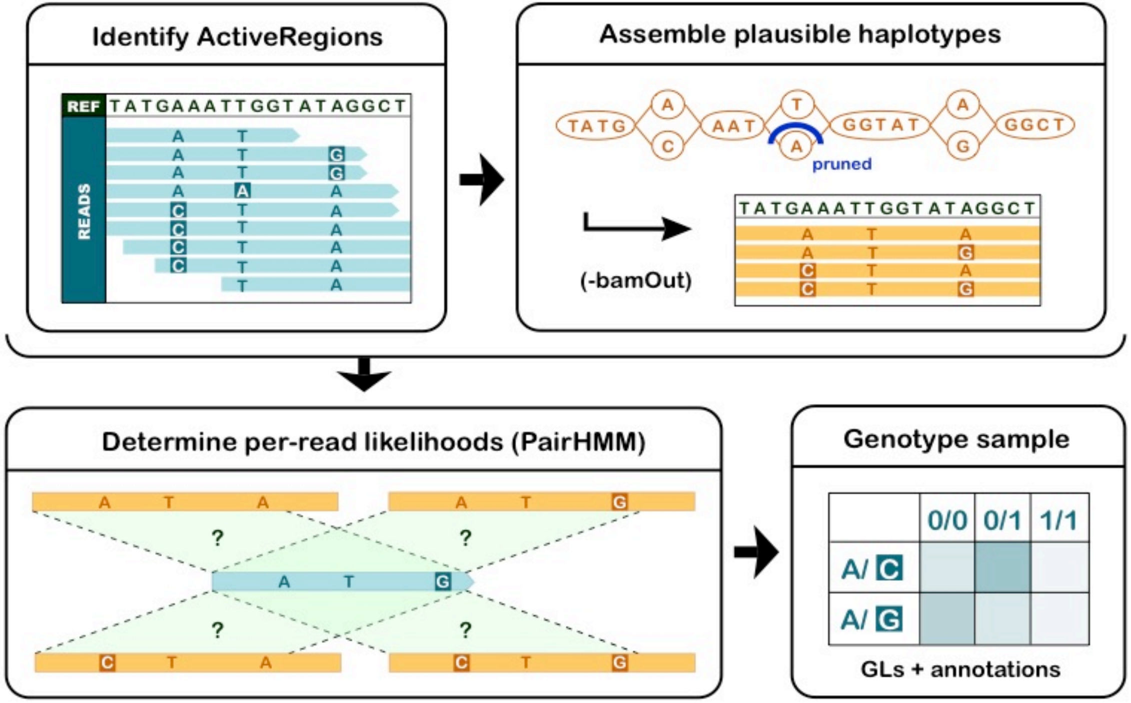
gatk

Best Practices for Germline SNP & INDEL Discovery



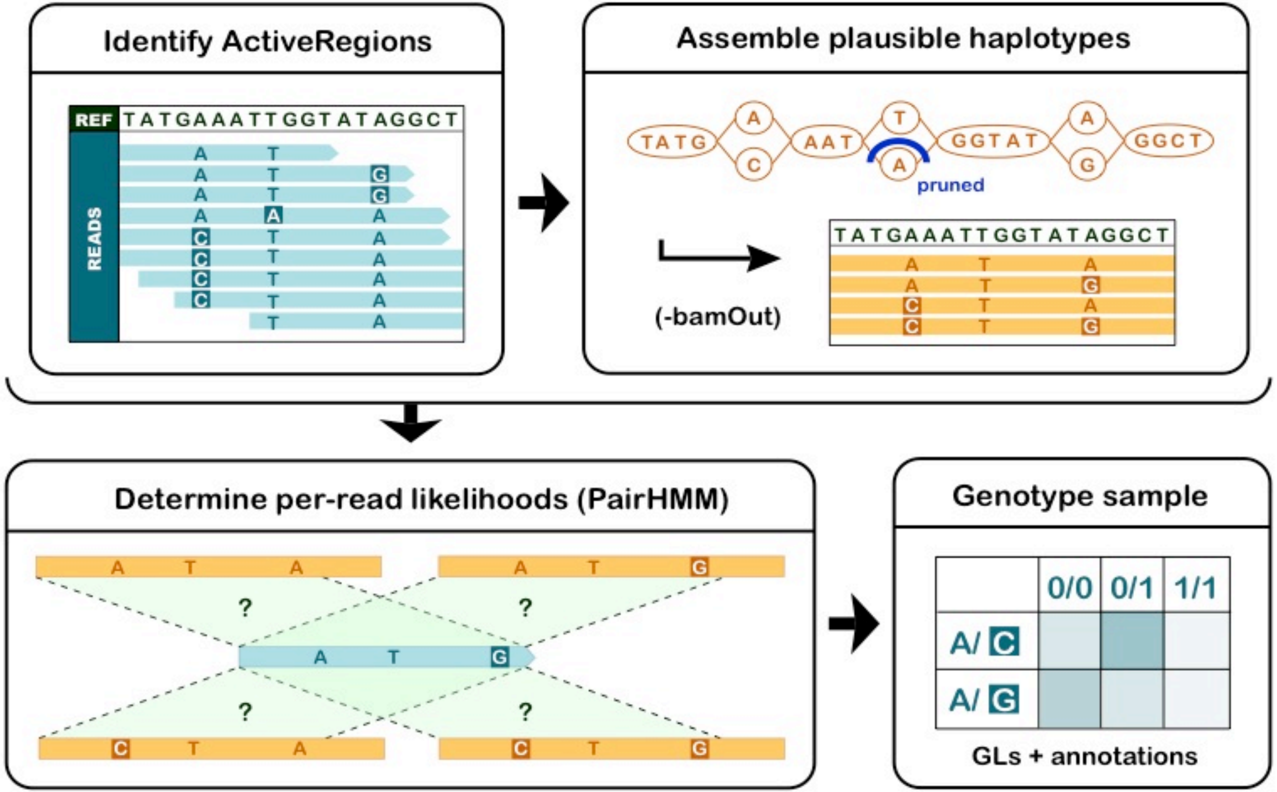
Call variants per-sample with HaplotypeCaller -> GVCF



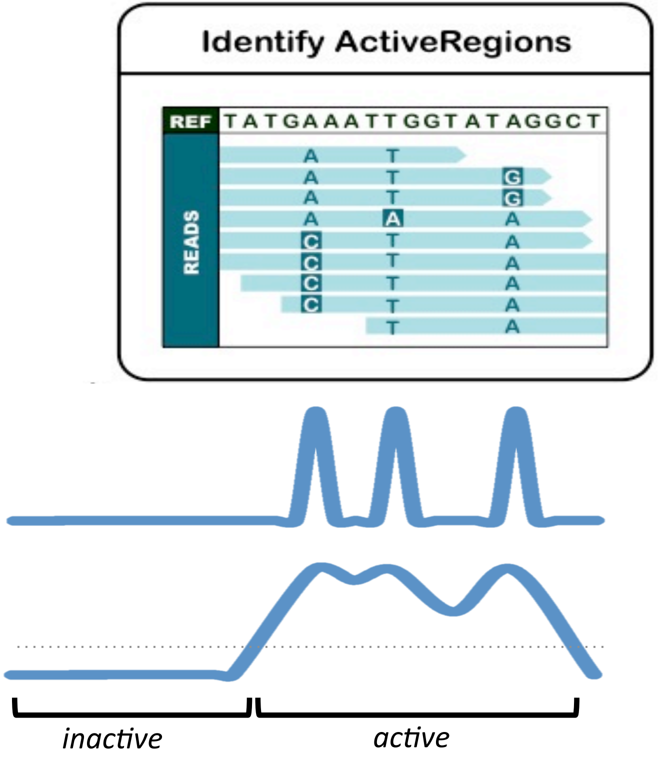


	0/0	0/1	1/1
A/ C			
A/ G			

HaplotypeCaller consists of 4 distinct operations



Step 1: Identify ActiveRegions



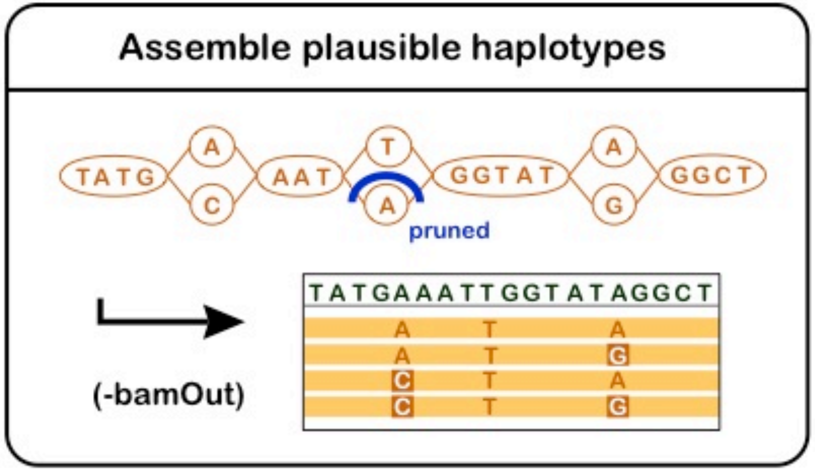
- Sliding window along the reference
- Count mismatches, indels and soft-clips
- Measure of entropy

Trim and continue with ActiveRegions over threshold

** Toy example*

Step 2: Assemble plausible haplotypes

- Local realignment via graph assembly
- Traverse graph to collect most likely haplotypes
- Align haplotypes to reference using Smith-Waterman

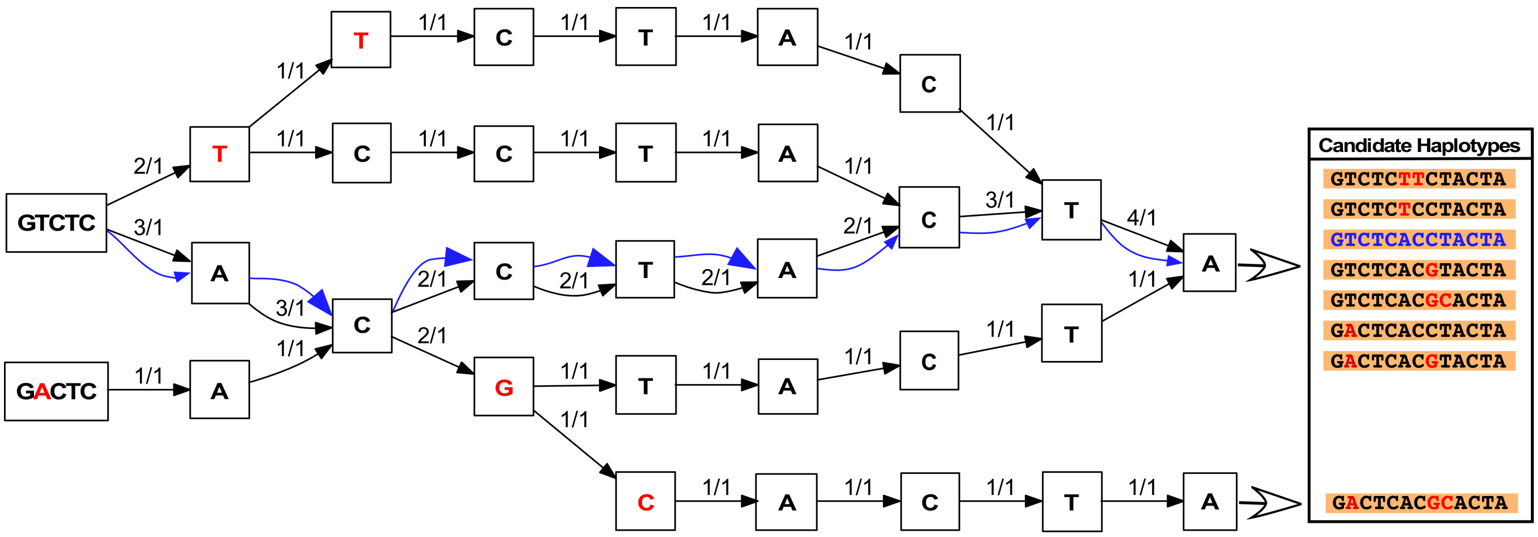


Assemble plausible haplotypes

Likely haplotypes + candidate variant sites

Can make HC output the reassembled reads and selected haplotypes using the –bamOut parameter

Example HaplotypeCaller assembly graph



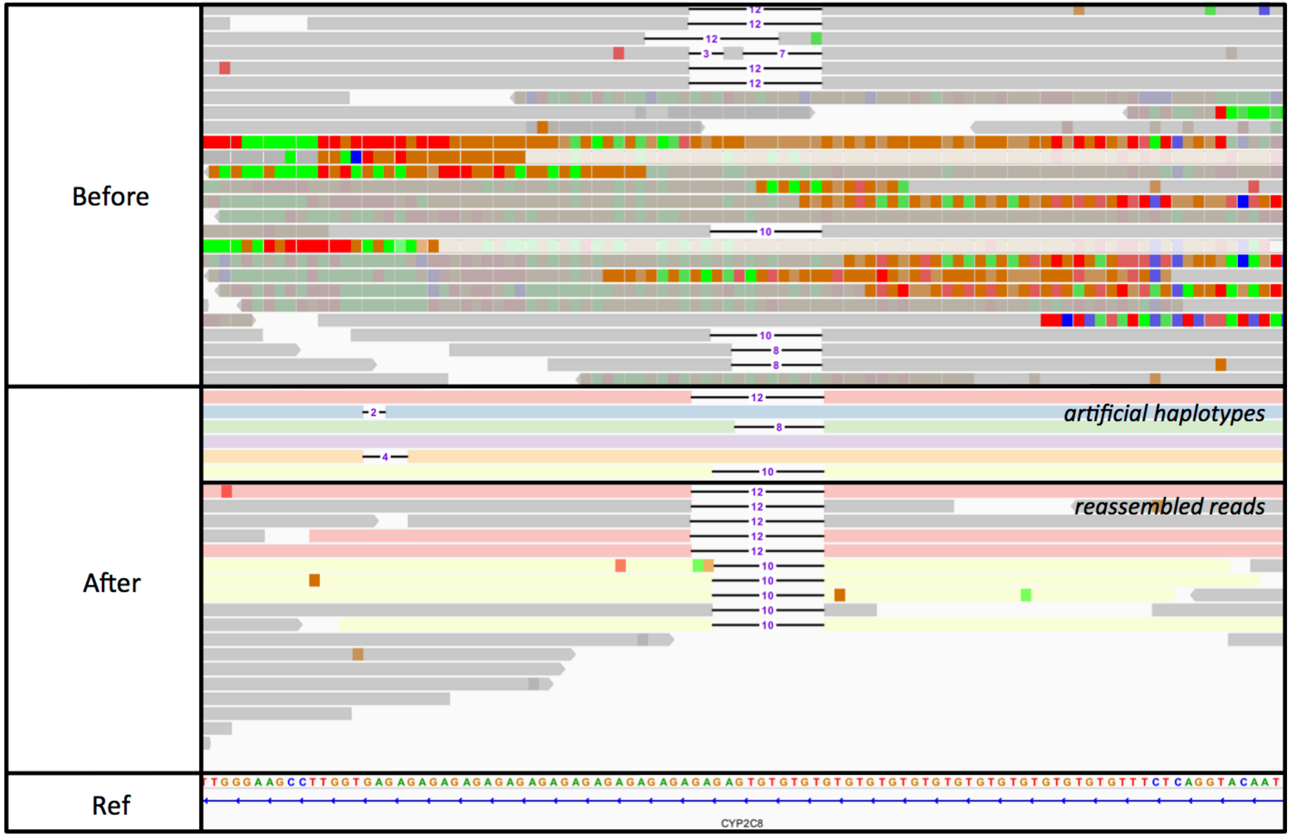
Candidate Haplotypes

GTCTCTTCTACTA
GTCTCTCCTACTA
GTCTCACCTACTA
GTCTCACGTACTA
GTCTCACGCACTA
GACTCACCTACTA
GACTCACGTACTA

GACTCACGCACTA

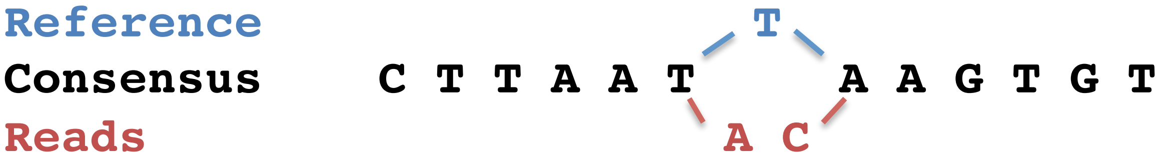
- Ignore previous alignments
- Graph consists of every possible sequence combination based on reads
- Count reads that support paths

Graph assembly recovers indels and removes artifacts


Before
After
Ref
12
12
12
3
7
12
12
10
10
8
8
12
2
8
4
10
artificial haplotypes
12
12
12
12
12
10
10
10
10
10
reassembled reads
TTGGGAAGCCTTGGTGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTCTCAGGTACAAT
CYP2C8

Showing 100bp region starting at 10:96,825,862 for NA12878

Resolves complexity caused by mapper limitations

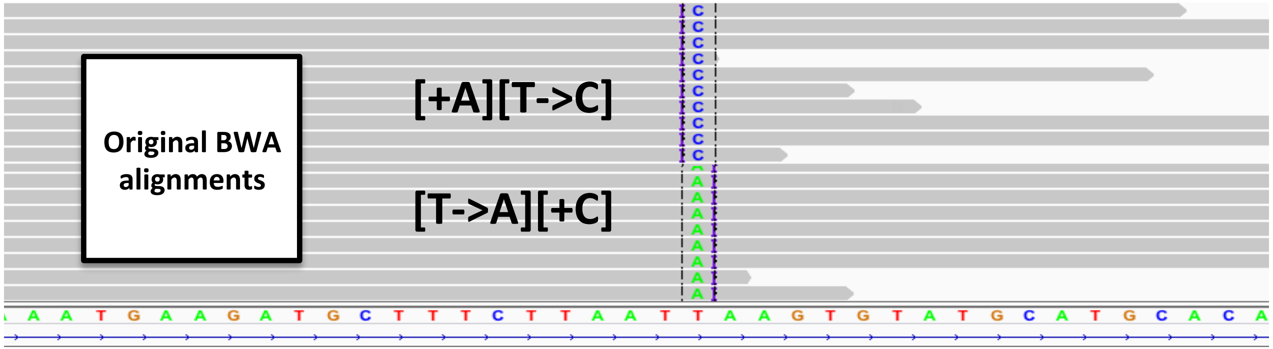


Reference T

Consensus C T T A A T A A G T G T

Reads A C

- Mapper can represent two different ways, at random:



Original BWA alignments

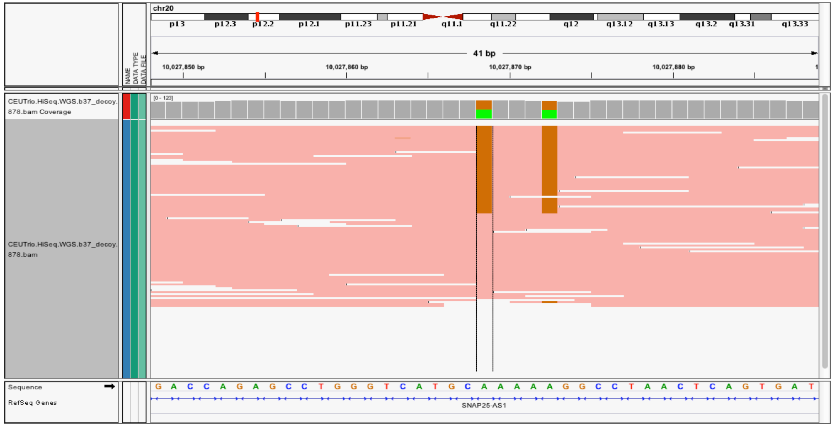
[+A][T->C]

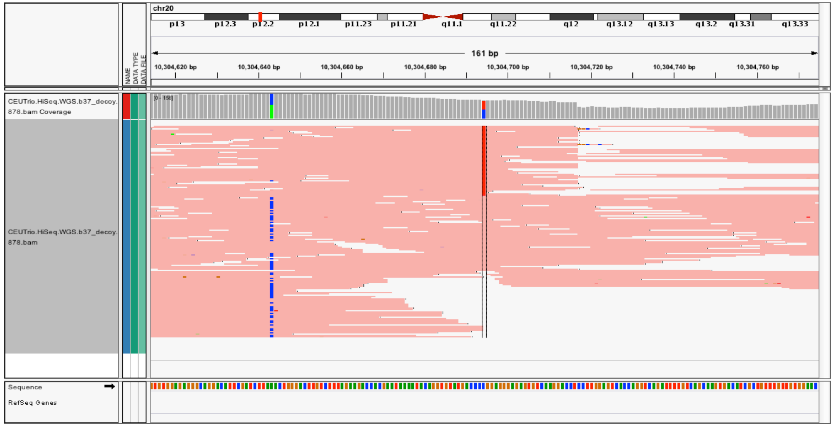
[T->A][+C]

C
A
A A T G A A G A T G C T T T C T T A A T T A A G T G T A T G C A T G C A C A

- HaplotypeCaller will settle on one representation -> cleaner output call

Bonus perk of haplotype calling: physical phasing

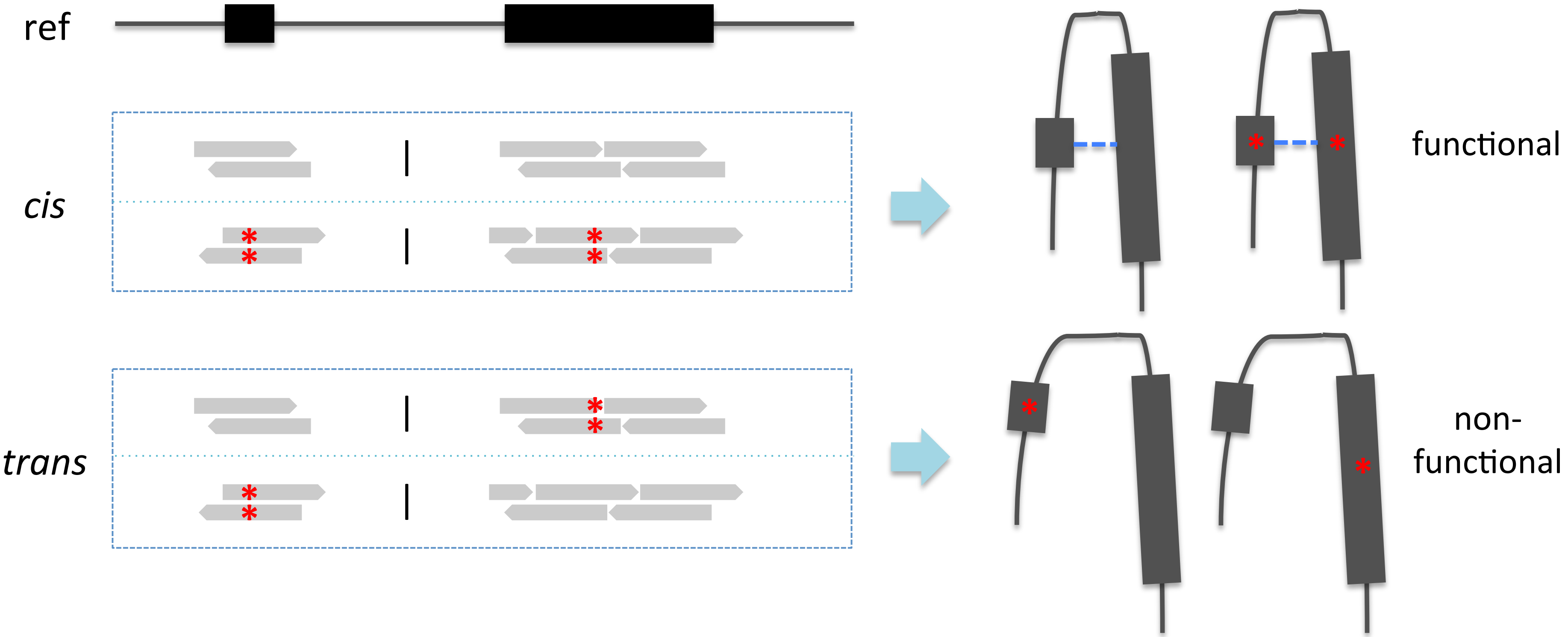




Two new sample-level annotations are PGT (phased genotype) and PID (phase identifier):

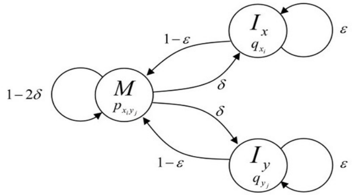
```
#CHROM POS … REF ALT            …  FORMAT        SAMPLE
1   1372268 . G   A,<NON_REF>  …  GT…:PGT:PID:…  0/1…:0|1:1372268_G_A:…
1   1372269 . G   T,<NON_REF>  …  GT…:PGT:PID:…  0/1…:0|1:1372268_G_A:…
```

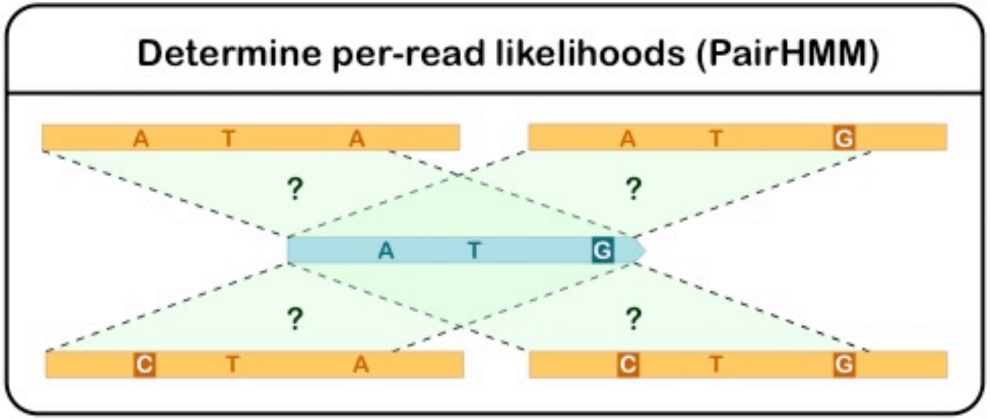
Functional implications of variant phasing



Step 3: Score haplotypes using PairHMM

- PairHMM* aligns each read to each haplotype
- Uses base qualities as the estimate of error

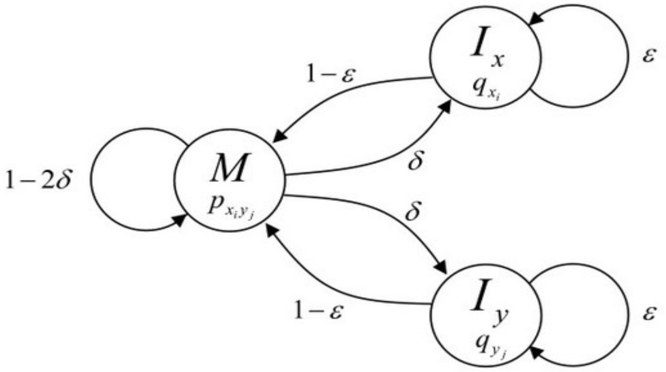



Determine per-read likelihoods (PairHMM)

Likelihoods of the haplotypes given reads

** Hardware-optimized versions of PairHMM are included and are activated automatically at runtime*

PairHMM uses base qualities to score alignments



	State
(*M*)	Match
(I_x)	Insertion
(I_y)	Deletion

Transition probabilities (derived from BQSR)

(ε) = Gap continuation

(δ) = Gap open penalty

(1 - ε) = Base precedes an insertion or a deletion

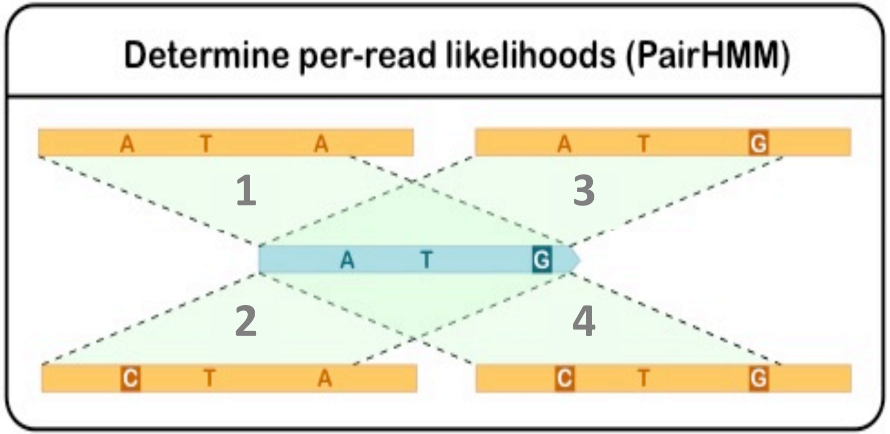
(1 - 2δ) = Base matches and continues

$$\begin{array}{cc} & \text{Haplotypes} \\ \text{Reads} & \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & A_{2n} \\ \vdots & & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \end{array}$$

A_{ij} = probability of haplotype-read pair

Matrix contains likelihoods of the haplotypes given the reads

Transforming **support for haplotypes** into **support for alleles**



Haplotypes

Reads	1	2	3	4
1	0.05	0.10	0.06	0.01
2	0.05	0.02	0.09	0.10
3	0.05	0.12	0.03	0.02

Alleles

	C	A
1	0.10	0.06
2	0.10	0.09
3	0.12	0.05

Take the highest per-read likelihood of haplotypes with allele

Numbers are artificially large for illustration.

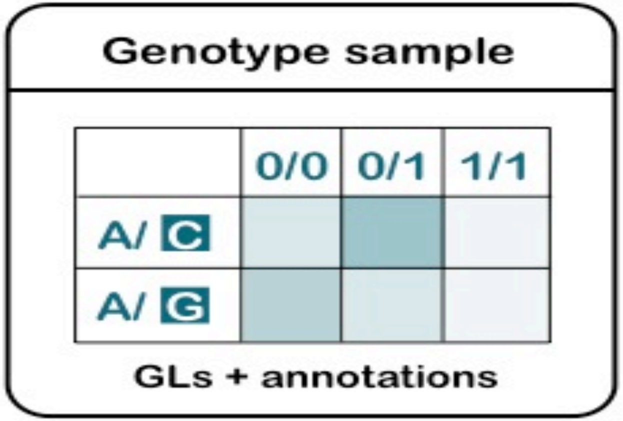
Step 4 : Genotype each sample at each potential variant site

- Determine most likely combination of allele(s) for each site
- Based on allele likelihoods (from PairHMM)
- Apply Bayes’ theorem with ploidy assumption*

$$P(G|R) = \frac{P(R|G)P(G)}{\sum_i P(R|G_i)P(G_i)}, \text{ where } P(R|G) = \mathcal{L}(G|R)$$

$$\mathcal{L}(G|R) = \prod_j \left(\frac{\mathcal{L}(H_1|R_j)}{2} + \frac{\mathcal{L}(H_2|R_j)}{2} \right), \; G = H_1 H_2 \text{ for diploids}$$

$\mathcal{L}(H_i|R_j)$ is the per read haploid likelihood



Genotype sample

	0/0	0/1	1/1
A/ C			
A/ G			

GLs + annotations

→ **Genotype calls**

** Default is diploid; can set desired ploidy in command line*

And finally, a bit of Bayesian math

Likelihood of the genotype

Genotype prior

Posterior probability of the genotype given the reads

$$\boxed{P(G|R)} = \frac{\boxed{P(R|G)}\boxed{P(G)}}{\sum_i P(R|G_i)P(G_i)}, \quad \text{where } P(R|G) = \mathcal{L}(G|R)$$

$$\mathcal{L}(G|R) = \prod_j \left(\frac{\mathcal{L}(H_1|R_j)}{2} + \frac{\mathcal{L}(H_2|R_j)}{2} \right), \quad G = H_1 H_2 \text{ for diploids}$$

$$\mathcal{L}(H_i|R_j) \text{ is the per read haploid likelihood}$$

Plug in the numbers!

		Alleles	
		C	**A**
Reads	**1**	**0.10**	**0.06**
	2	**0.10**	**0.09**
	3	**0.12**	**0.05**

Determines the most likely genotype of the sample at each event in the haplotypes

Follow through for genotype probability

$$\mathcal{L}(G|R) = \prod_j \left(\frac{\mathcal{L}(H_1|R_j)}{2} + \frac{\mathcal{L}(H_2|R_j)}{2} \right), \; G = H_1 H_2 \text{ for diploids}$$

Reads	Alleles C	Alleles A
1	0.10	0.06
2	0.10	0.09
3	0.12	0.05

Genotype likelihoods for $G_{C/C}$, $G_{C/A}$ and $G_{A/A}$ given reads $R_{1\text{-}3}$:

$L(G_{C/C}|R_{1\text{-}3}) = [(0.10+0.10)/2]*[(0.10+0.10)/2]*[(0.12+0.12)/2] = \mathbf{0.00120}$

$L(G_{C/A}|R_{1\text{-}3}) = [(0.10+0.06)/2]*[(0.10+0.09)/2]*[(0.12+0.05)/2] = 0.00065$

$L(G_{A/A}|R_{1\text{-}3}) = [(0.06+0.06)/2]*[(0.09+0.09)/2]*[(0.05+0.05)/2] = 0.00027$

Multiply by prior and divide by sum (0.002116)

Genotype probability :

$P(G_{C/C}|R_{1\text{-}3}) = \mathbf{0.567}$

$P(G_{C/A}|R_{1\text{-}3}) = 0.305$

$P(G_{A/A}|R_{1\text{-}3}) = 0.128$

- Assigns highest probability genotype C/C
- For variant genotypes, emit variant record

Example **PL** and **GQ** calculations

- PL is the normalized Phred-scaled probability of each genotype

	A/A	A/C	C/C
P(G|R)	0.128	0.305	0.567
Raw PL	8.94	5.15	2.46
Normalized **PL**	6	3	0

$(-10) * \log_{10}\{P(G|R)\}$

subtract smallest PL

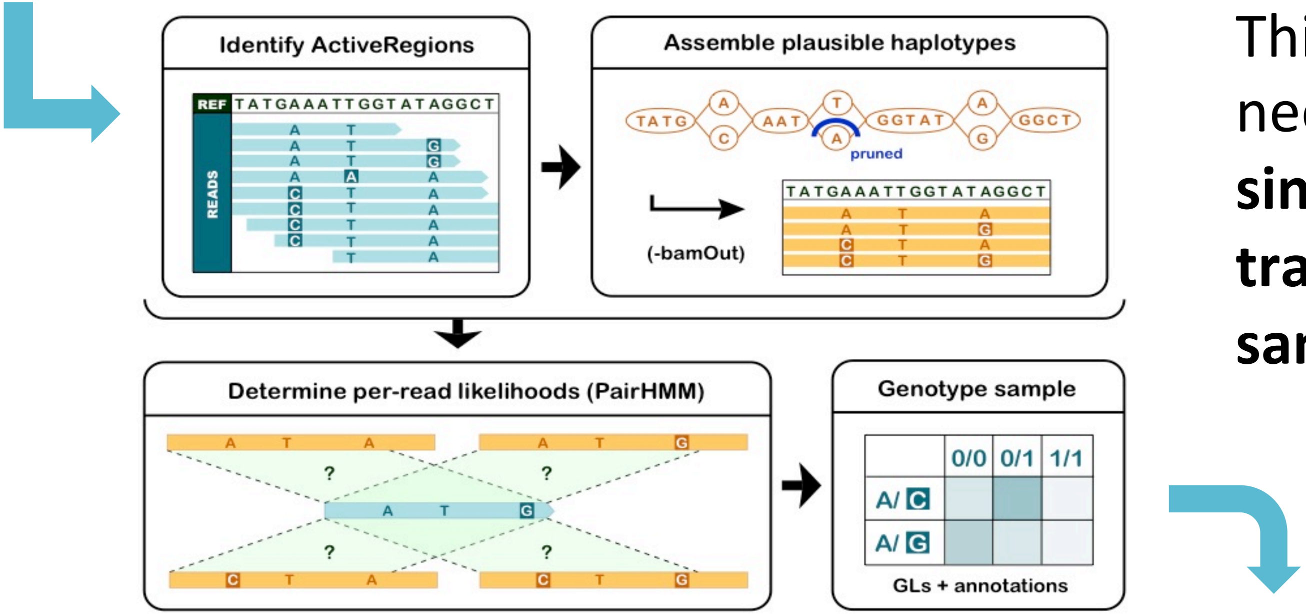
GQ

- GQ is the genotype quality and is the smaller of the 2nd PL or 99
- PLs are in increasing order of possible genotypes, e.g. 0/0, 0/1 and 1/1.

```
#…  REF    ALT  …    FORMAT        SAMPLE
…   A      C    …    GT…:GQ:PL…    1/1…:3:6,3,0…
```

HaplotypeCaller recap: reads in / variants out

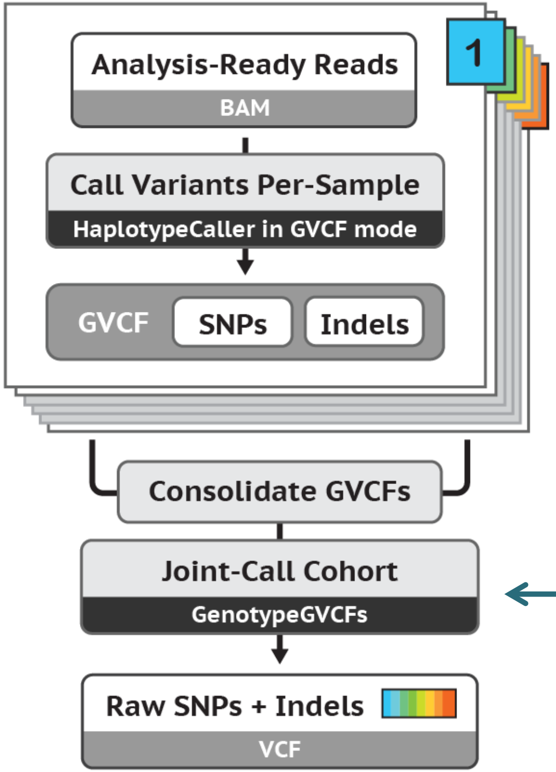
BAM



VCF & index

This is all you need for a **single sample** or **traditional multi-sample** analysis

For scalable analysis: emit **GVCF** + add **joint calling** step



Analysis-Ready Reads

BAM

Call Variants Per-Sample

HaplotypeCaller in GVCF mode

GVCF SNPs Indels

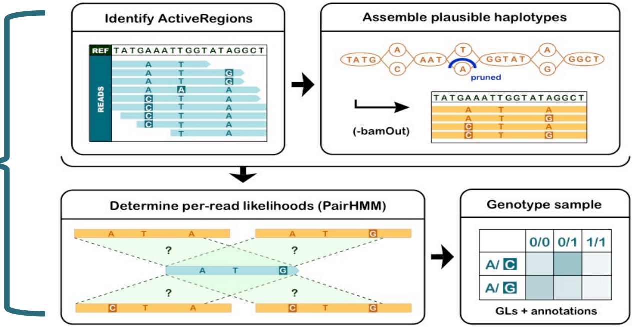
Consolidate GVCFs

Joint-Call Cohort

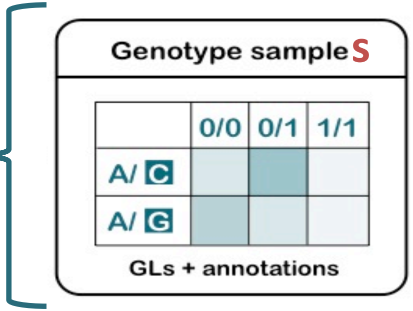
GenotypeGVCFs

Raw SNPs + Indels

VCF



- Run HC in **GVCF mode** to emit GVCF



- Run GenotypeGVCFs to re-genotype samples with **multi-sample model**

Running HaplotypeCaller

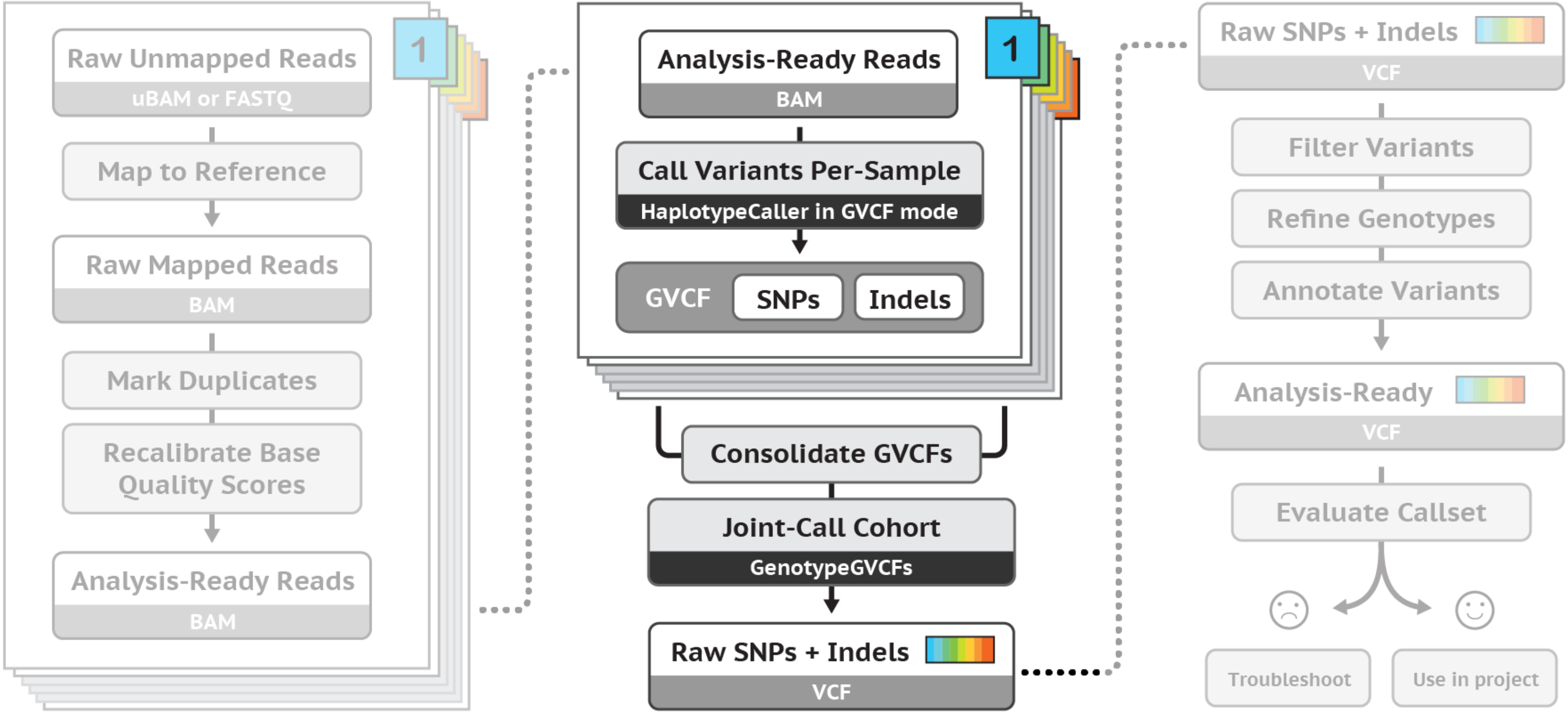
Basic mode (no GVCF):

```
gatk HaplotypeCaller \
    -R reference.fasta \
    -I preprocessed_reads.bam \
    -O germline_variants.vcf
```

To produce a block-compressed GVCF, substitute output filename and add:

```
    -O germline_variants.g.vcf \
    -ERC GVCF
```

Next step: joint calling


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1
Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project