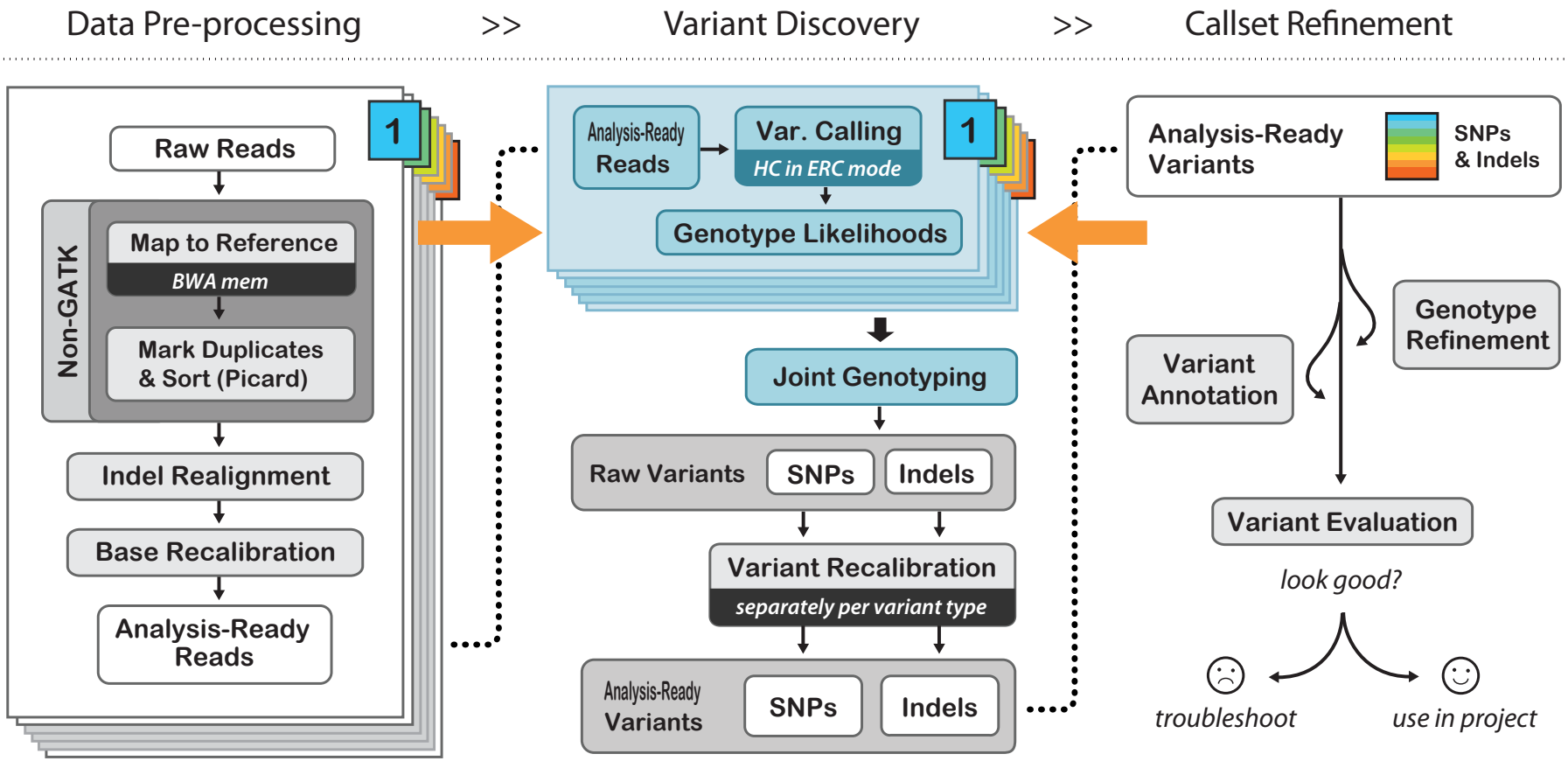
Germline variant calling and joint genotyping

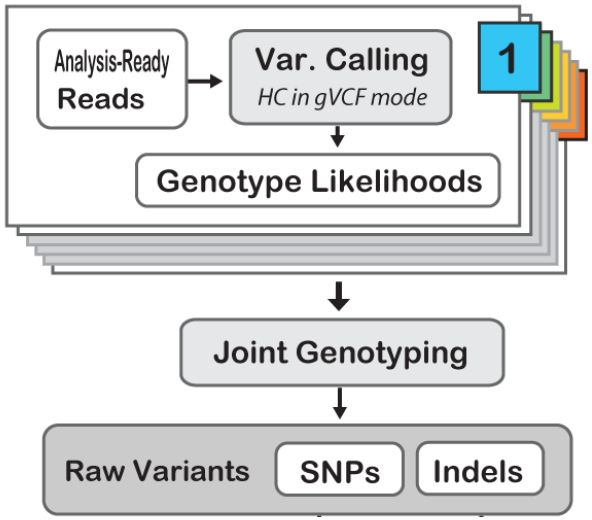
Applying the joint discovery workflow with HaplotypeCaller + GenotypeGVCFs

You are here in the GATK Best Practices workflow for **germline variant discovery**

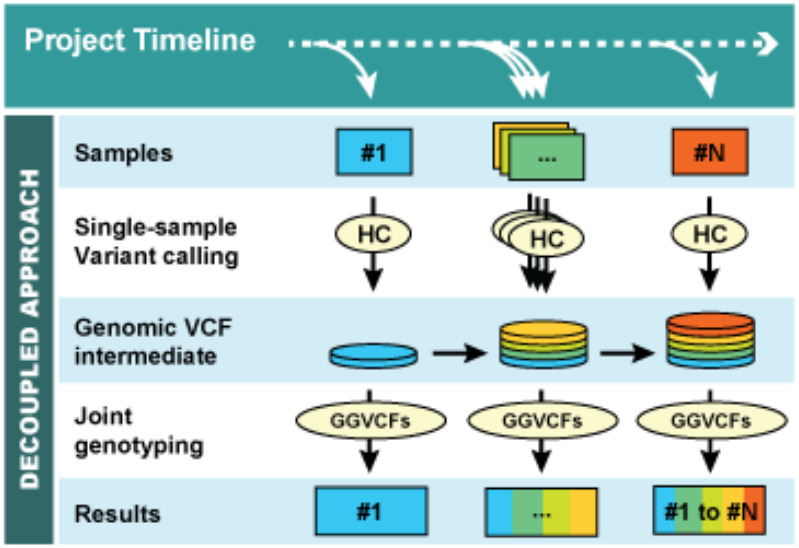


A scalable workflow for joint variant discovery

Scalable over sample size



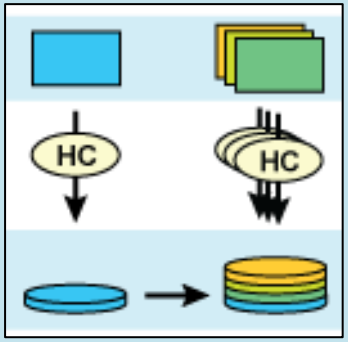
+ Incremental over time



Tools involved in the workflow

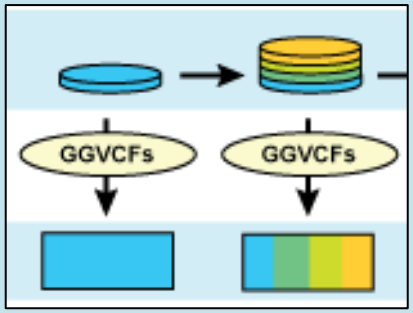
- Identify potential variants in each sample

  → **HaplotypeCaller**



- Perform joint genotyping on the cohort

  → **GenotypeGVCFs**



Key HaplotypeCaller features

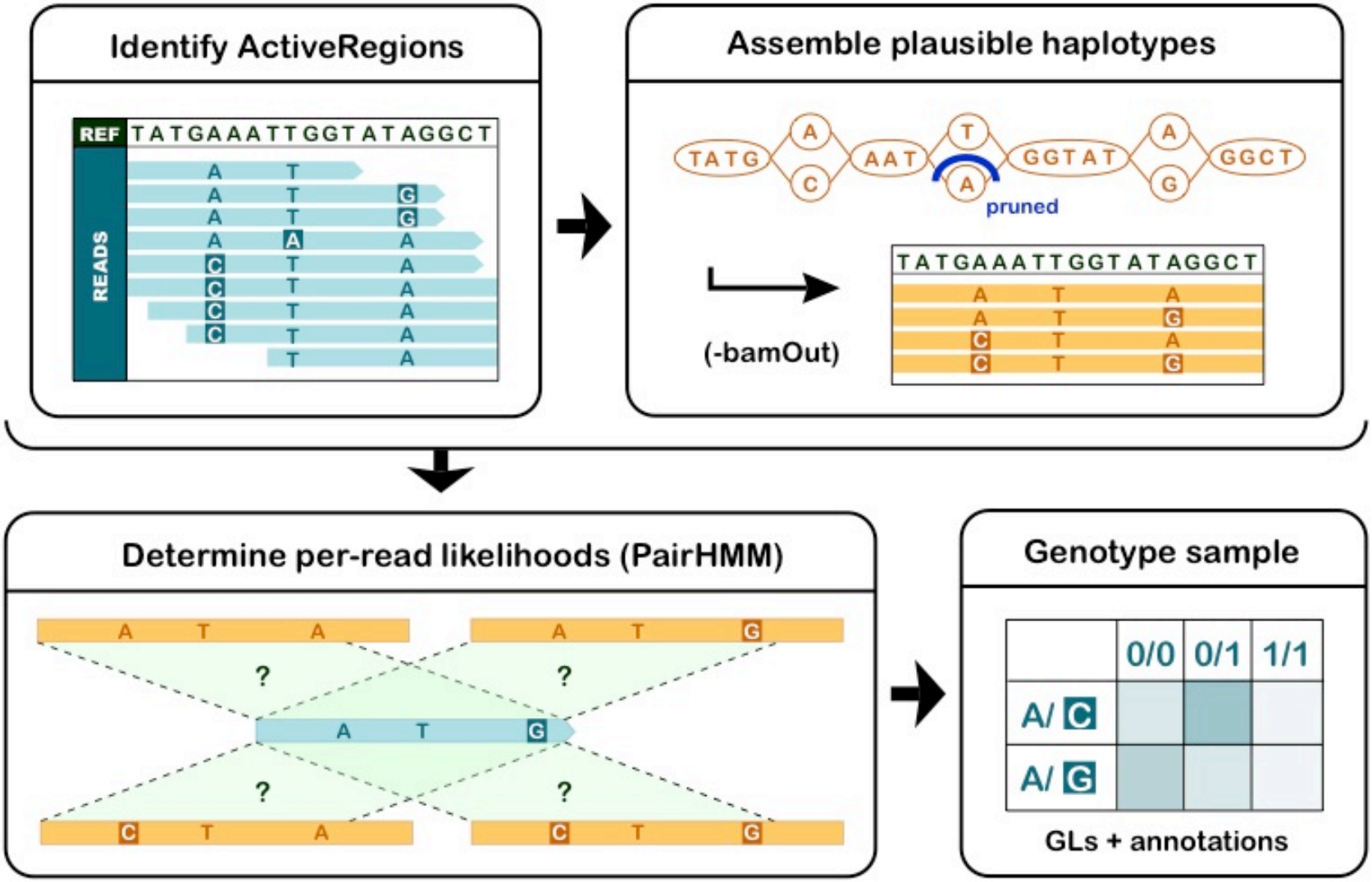
What it does:

- Calls SNP and indel variants simultaneously
- Performs local re-assembly to identify haplotypes
- Reference confidence model enables detection of low frequency variants
- Joint-discovery workflow (reference confidence model , GVCFs)
- Handles RNAseq natively
- Handles non-diploid organisms and pooled samples

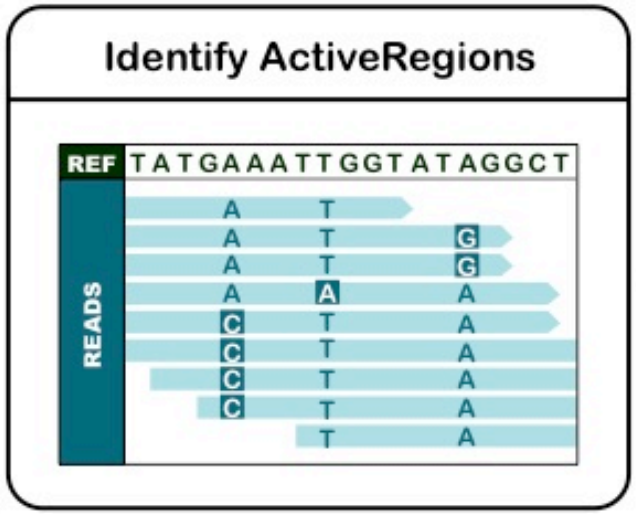
What it doesn't do

- Somatic variant calling (use MuTect2 instead!)

How HaplotypeCaller works in 4 simple steps

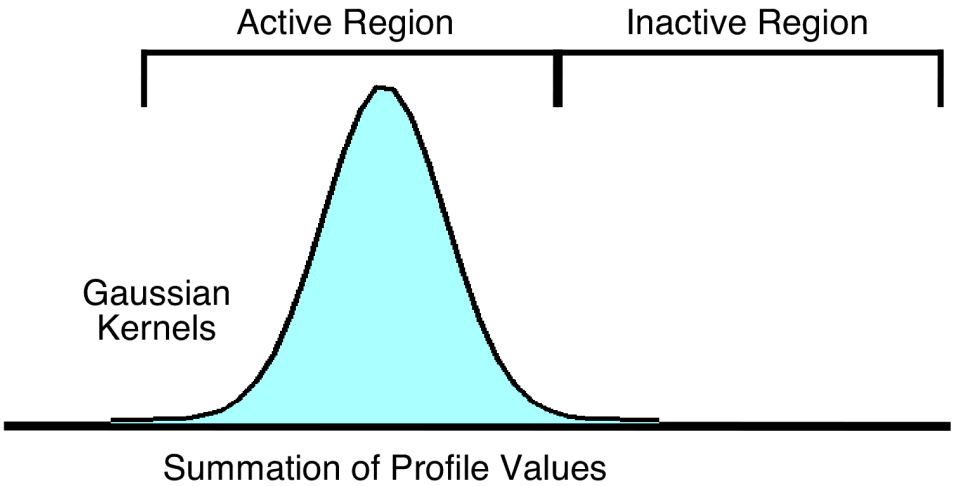


Step 1: Identify ActiveRegions



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

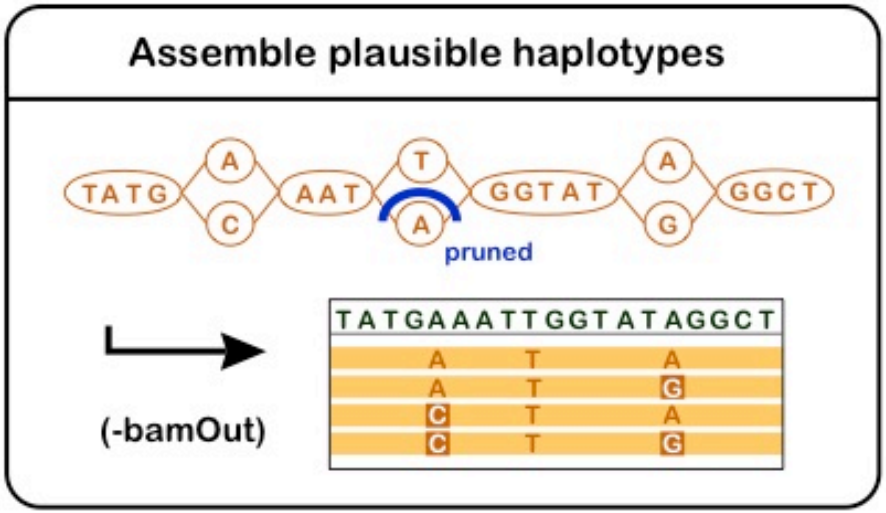
➢ **Measure of entropy**



Over threshold:
Trigger "ActiveRegion" to be processed

Step 2: Assemble plausible haplotypes

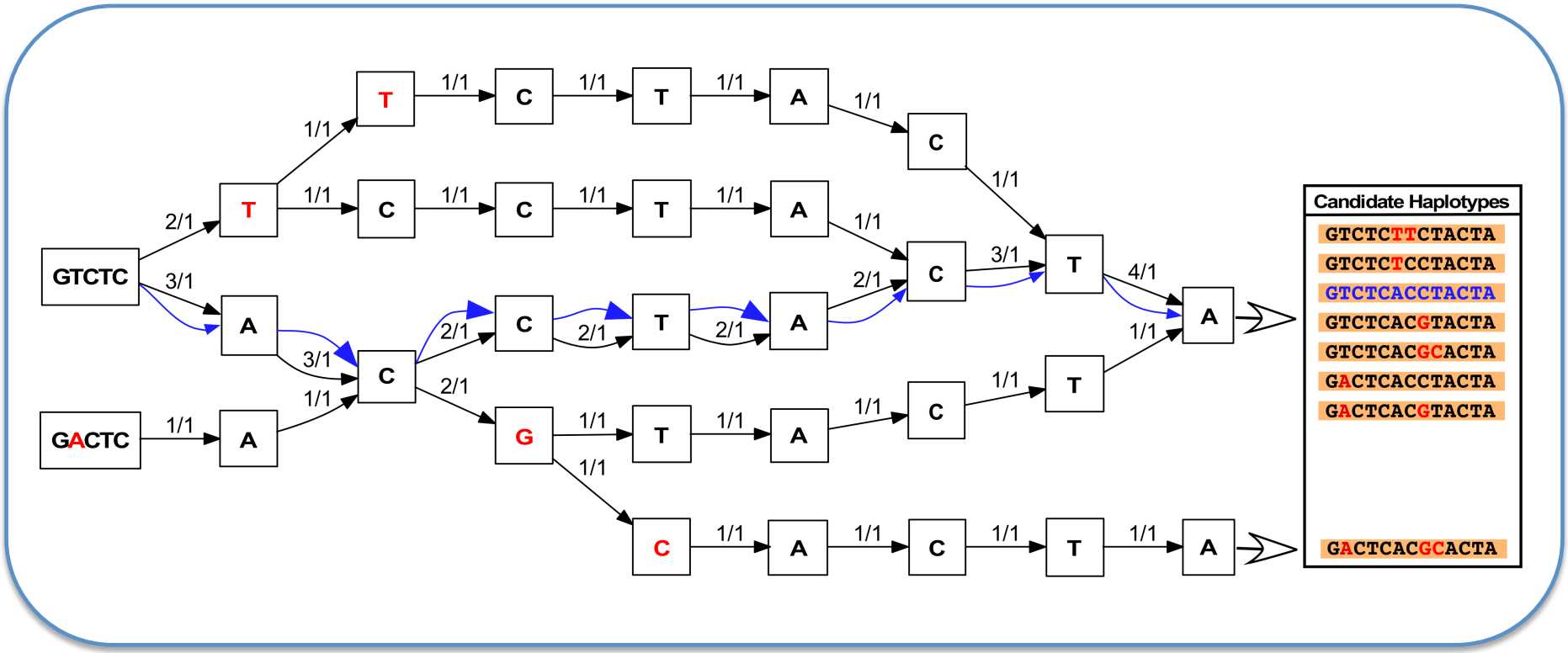
- Local realignment via graph assembly
- Traverse graph to collect most likely haplotypes
- Align haplotypes to ref using Smith-Waterman



Likely haplotypes + candidate variant sites

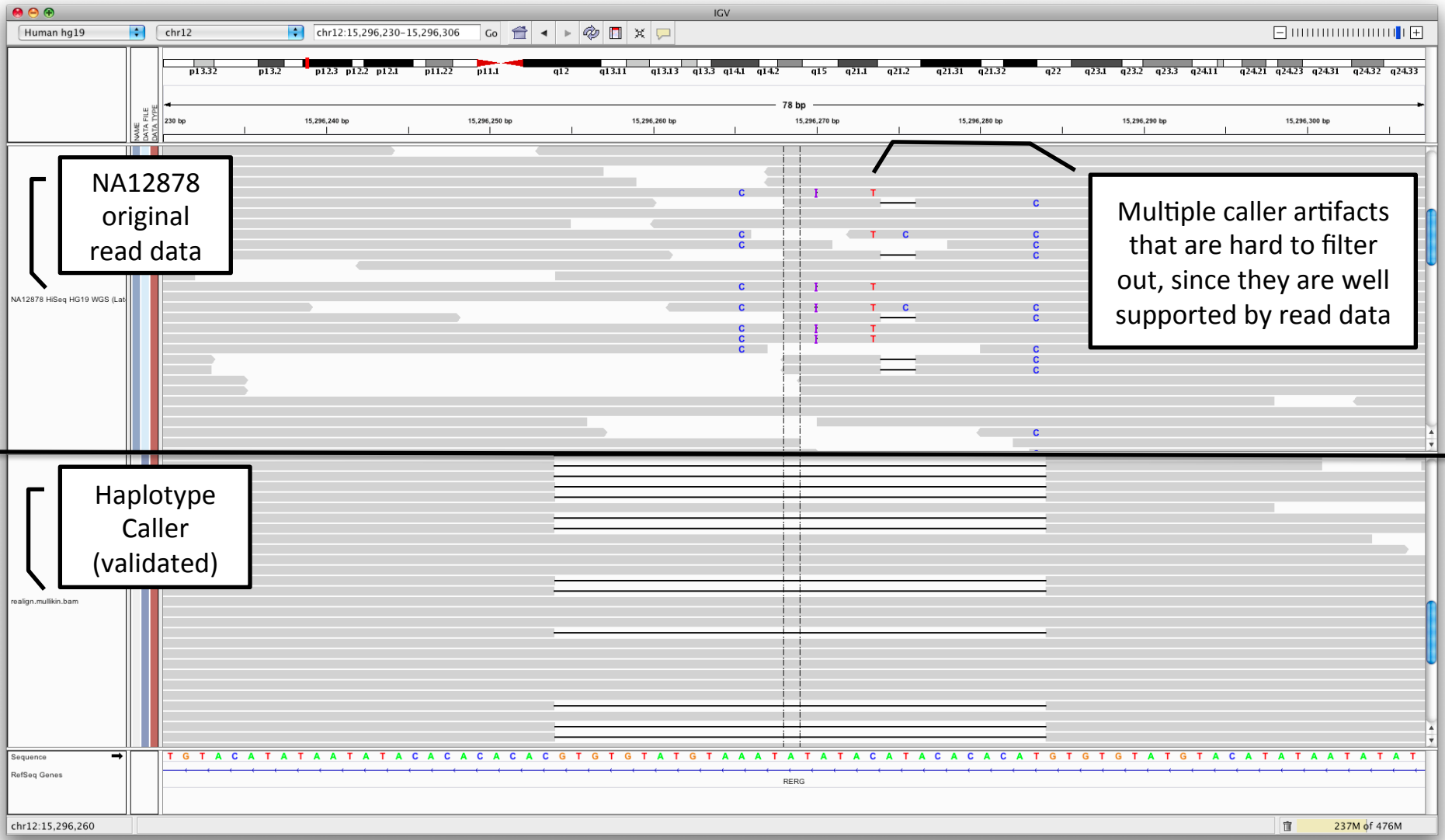
Can make HC output the reassembled reads and selected halpotypes using the –bamOut parameter

Example assembly graph produced by HaplotypeCaller



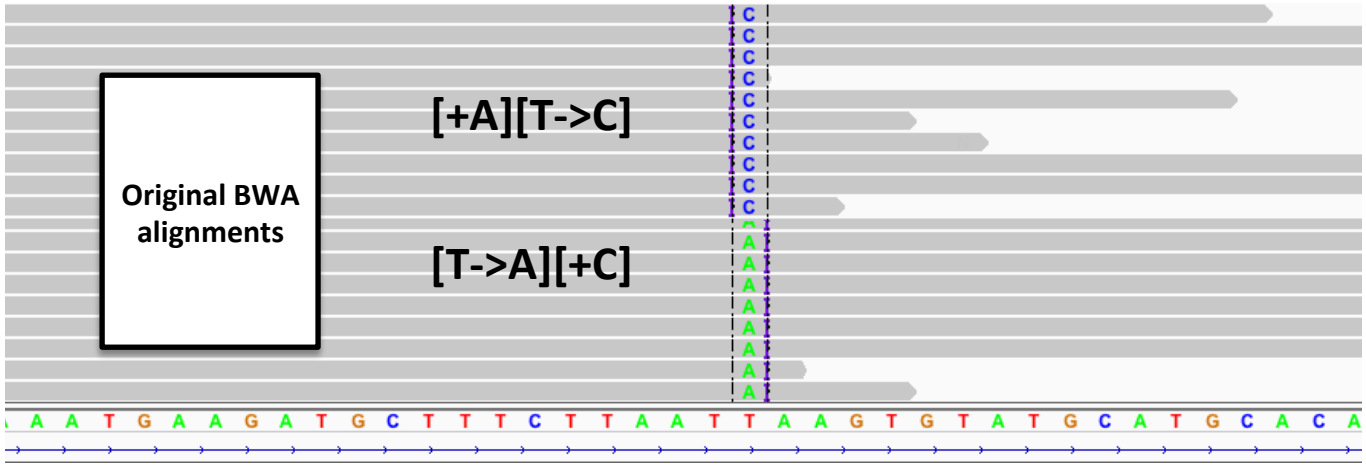
- Previous alignments are ignored
- K-mers consist of every possible sequence combination based on the reads
- Most likely paths through the graph are scored

Graph assembly recovers indels and removes artifacts



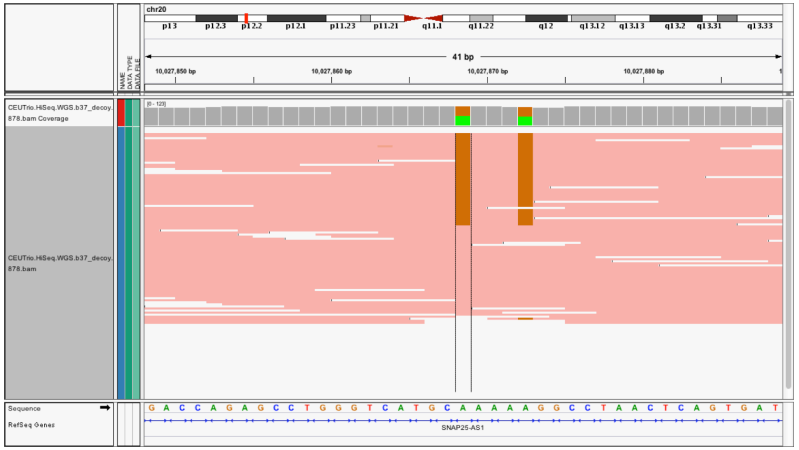
Graph assembly resolves complexity caused by mapper limitations

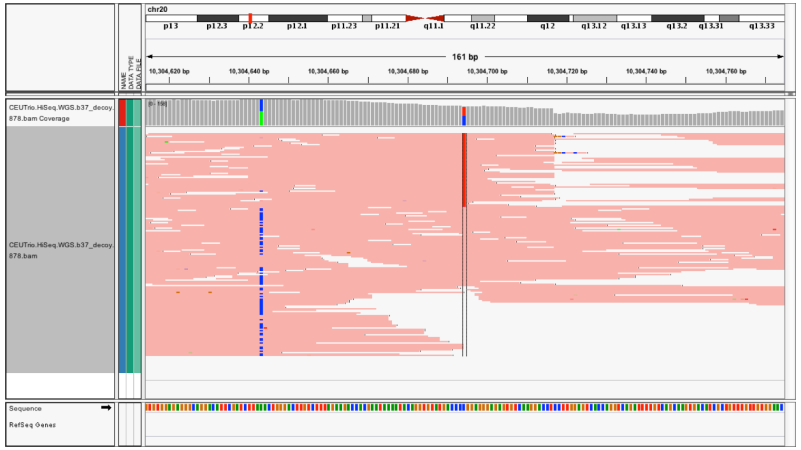
Can be represented by the mapper two different ways, at random:



HaplotypeCaller will settle on one representation -> cleaner output call

Bonus perk of haplotype calling: free physical phasing





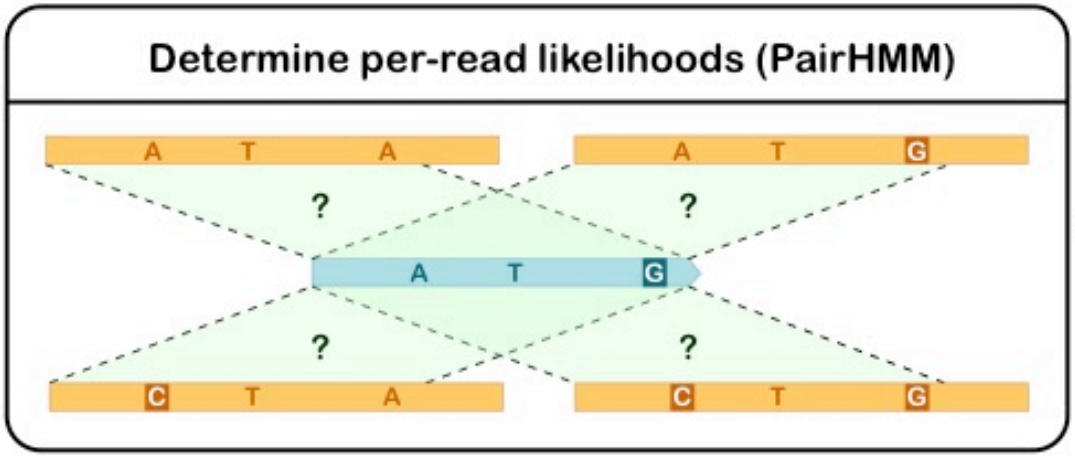
Two new sample-level annotations, PID (for phase identifier) and PGT (phased genotype)

```
1   1372243  .  T  <NON_REF>  .  .  END=1372267  <snip>  <snip>
1   1372268  .  G  A,<NON_REF>  .  .  <snip>  GT:AD:DP:GQ:PGT:PID:PL:SB 0/1:30,40,0:70:99:0|1:1372268_G_A:<sn
ip>
1   1372269  .  G  T,<NON_REF>  .  .  <snip>  GT:AD:DP:GQ:PGT:PID:PL:SB 0/1:30,41,0:71:99:0|1:1372268_G_A:<sn
ip>
1   1372270  .  C  <NON_REF>  .  .  END=1372299  <snip>  <snip>
```

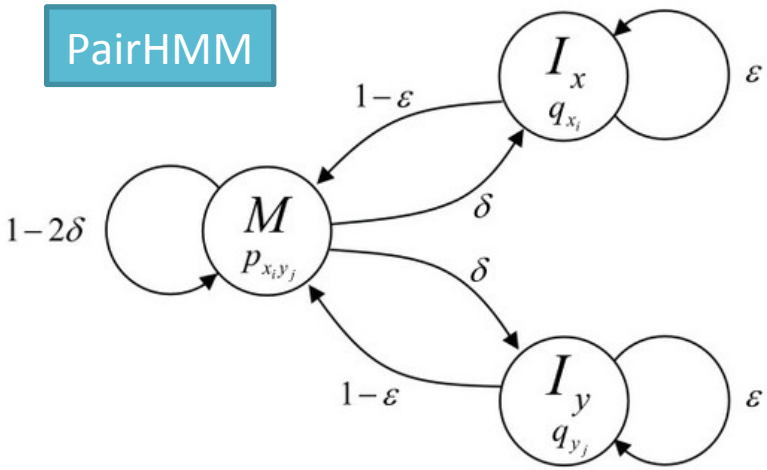
Step 3: Score haplotypes using PairHMM

- Calculate haplotype likelihoods given the read
  - PairHMM aligns each read to each haplotype

Likelihood of the haplotype given reads



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

Haplotypes

Reads

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

A_{ij} = probability of haplotype vs read

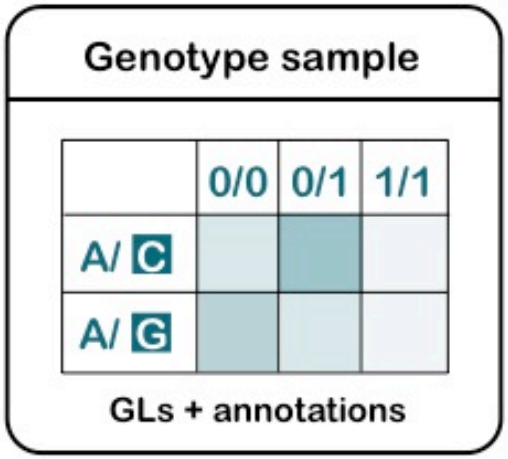
-> likelihoods of the haplotypes given the reads

-> store in matrix

Step 4 : Genotype each sample at each potential variant site

- Determine most likely alleles for each sample
- Based on support for haplotypes (from PairHMM)
- Evaluated over reads from each sample

Genotype calls for each sample



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

```
Reference:     ATCGATCATAGCTAGCTGCG
Haplotype 1:   ATCGA-CATAGCTAGCTGCG
Haplotype 2:   ATGGATCATAGCTTGCTGCG
Haplotype 3:   ATCGA-CATAGCTTGCTGCG
```

Haplotypes

Reads	* R	1	2	3
1	0.01	0.02	0.03	0.04
2	0.09	0.06	0.07	0.08
3	0.10	0.11	0.01	0.02

Alleles

Reads	-	T
1	0.04	0.03
2	0.08	0.09
3	0.11	0.10

Take highest probability of haplotypes given reads that contain the allele (for each variant position)

** These numbers are made up to give a sense of how the process works. In reality the numbers would be much smaller.*

And finally, a bit of Bayesian math

Just plug in the numbers!

Alleles: -	Alleles: T	Reads
0.04	0.03	1
0.08	0.09	2
0.11	0.10	3

Bayesian model

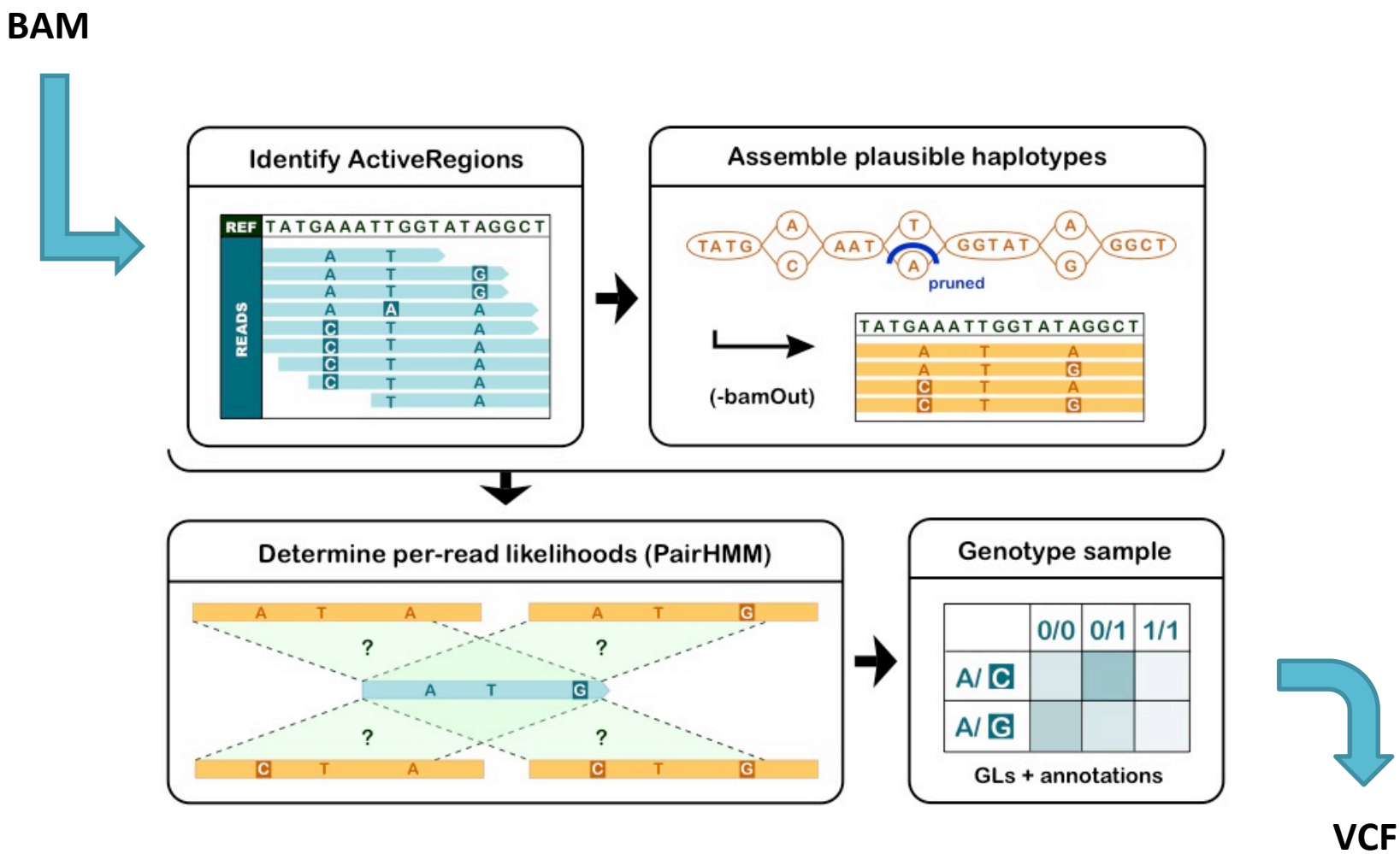
$$\Pr\{G|D\} = \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}}\ \overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2}\right) \text{ where } \overbrace{G = H_1 H_2}^{\text{Diploid assumption}}$$

$$\Pr\{D|H\} \text{ is the haploid likelihood function}$$

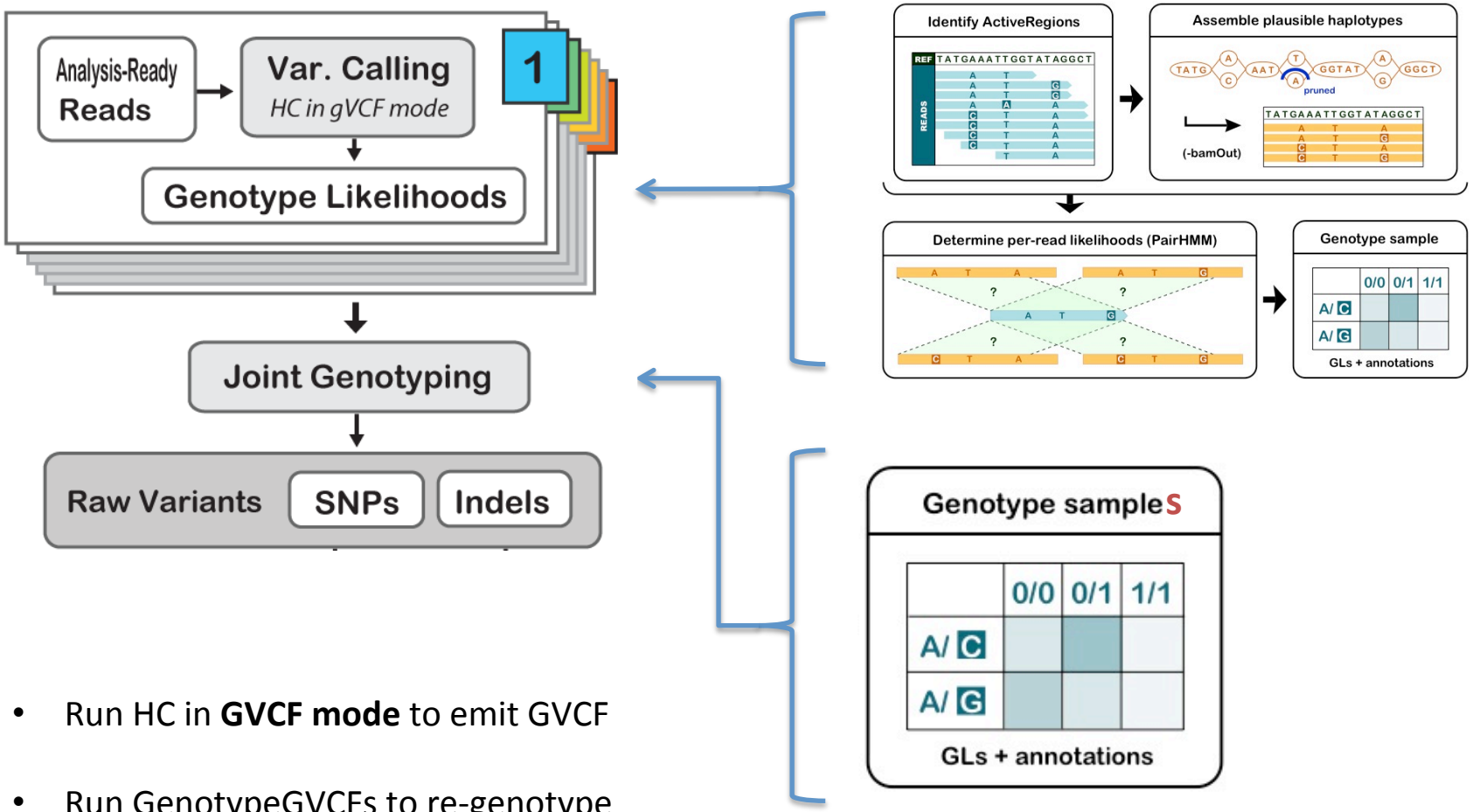
Determines the most likely genotype of the sample at each site where there is evidence of variation

HaplotypeCaller recap: reads in / variants out



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

For joint discovery: emit GVCF + add joint genotyping step



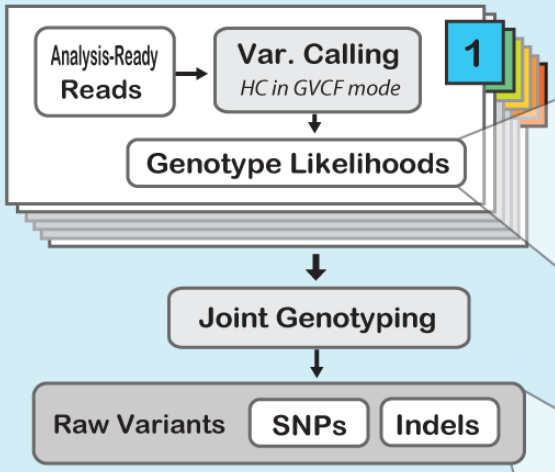
- Run HC in **GVCF mode** to emit GVCF
- Run GenotypeGVCFs to re-genotype samples with **multi-sample model**

GVCF includes **<NON-REF> allele + genotype likelihoods** for joint genotyping

Symbolic allele stands for all non-called but possible non-reference alleles

```
T    <NON_REF>    .    .    END=10000116
C    T,<NON_REF> 612.77   .    BaseQRankSu
```

end pos of hom-ref band



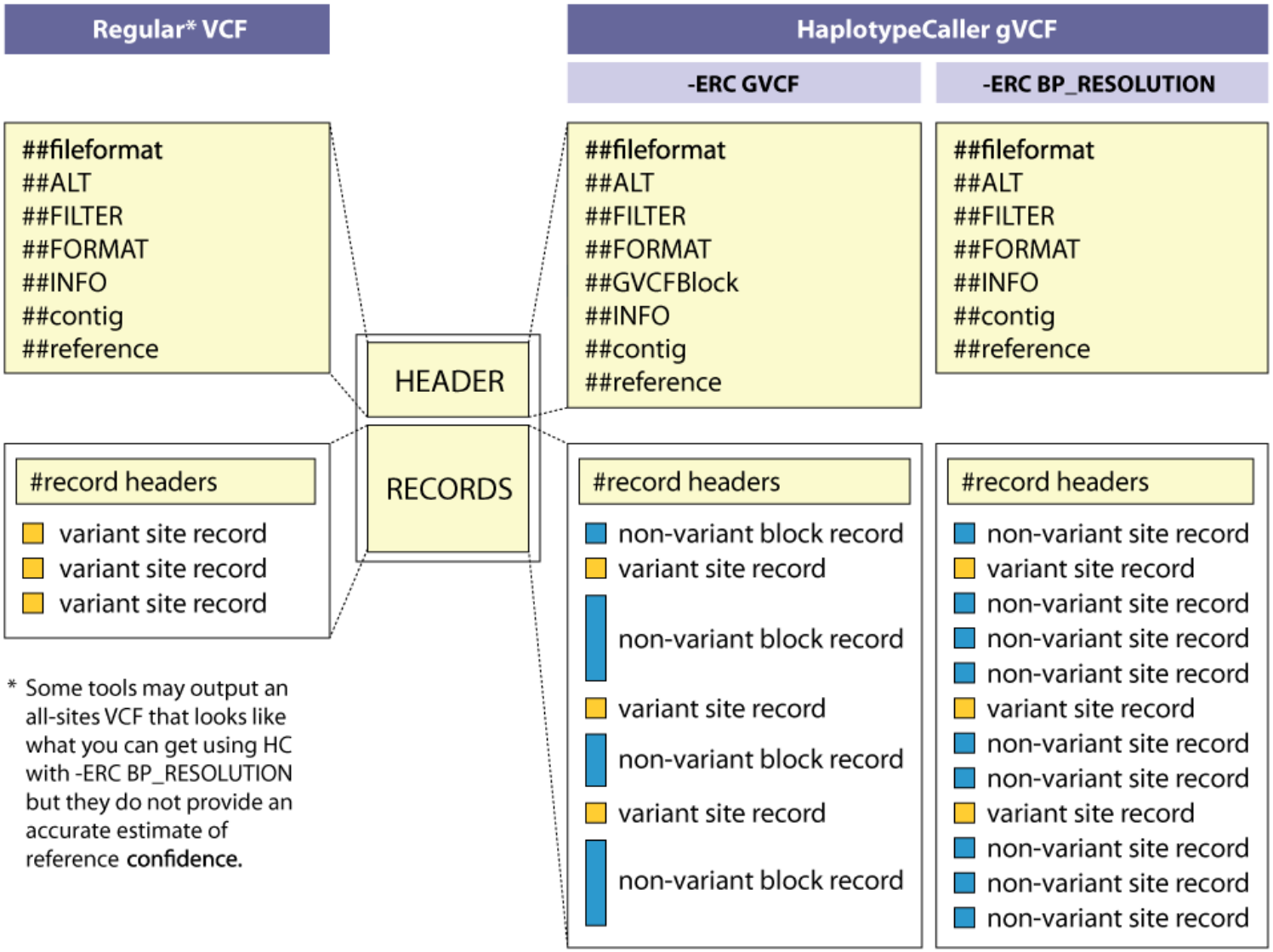
GVCF generated per sample (`-T HaplotypeCaller -ERC GVCF`)

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 . A <NON_REF> . . END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 . C T,<NON_REF> 326.77 . BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB 0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 . A <NON_REF> . . END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

VCF generated per cohort (`-T GenotypeGVCFs`)

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12877 NA12878 NA12882
20 10000117 . C T 1606.16 . AC=4;AF=0.667;AN=6;BaseQRankSum=1.66;ClippingRankSum=0.340;DP=85;
FS=5.718;MLEAC=4;MLEAF=0.667;MQ=60.36;MQRankSum=1.45;QD=18.90;ReadPosRankSum=1.62;SOR=1.503
GT:AD:DP:GQ:PL 0/1:17,15:32:99:399,0,439 0/1:11,12:23:99:291,0,292 1/1:0,30:30:90:948,90,0
20 10000211 . C T 1765.16 . AC=4;AF=0.667;AN=6;BaseQRankSum=2.34;ClippingRankSum=-1.147e+00;
DP=97;FS=0.809;MLEAC=4;MLEAF=0.667;MQ=60.00;MQRankSum=1.21;QD=18.58;ReadPosRankSum=0.152;SOR=0.831
GT:AD:DP:GQ:PL 0/1:13,10:23:99:243,0,341 0/1:21,14:35:99:355,0,526 1/1:0,37:37:99:1199,111,0
20 10000439 . T G 1982.13 . AC=5;AF=0.833;AN=6;BaseQRankSum=1.31;ClippingRankSum=0.549;DP=103;
FS=0.000;MLEAC=5;MLEAF=0.833;MQ=60.00;MQRankSum=0.972;QD=19.82;ReadPosRankSum=1.56;SOR=0.839
GT:AD:DP:GQ:PL 0/1:18,12:30:99:208,0,455 1/1:0,29:29:86:795,86,0 1/1:1,40:41:99:1010,110,0
```

GVCFs are valid VCFs with extra information



Multiple GVCFs combined form a **squared-off matrix** of genotypes

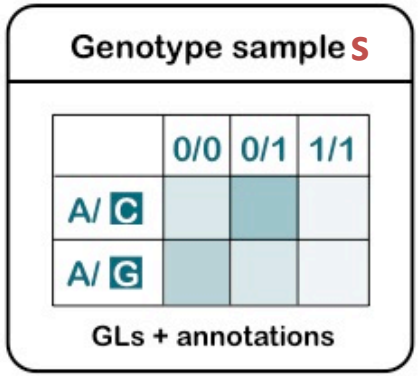
All case and control samples

~3M variants

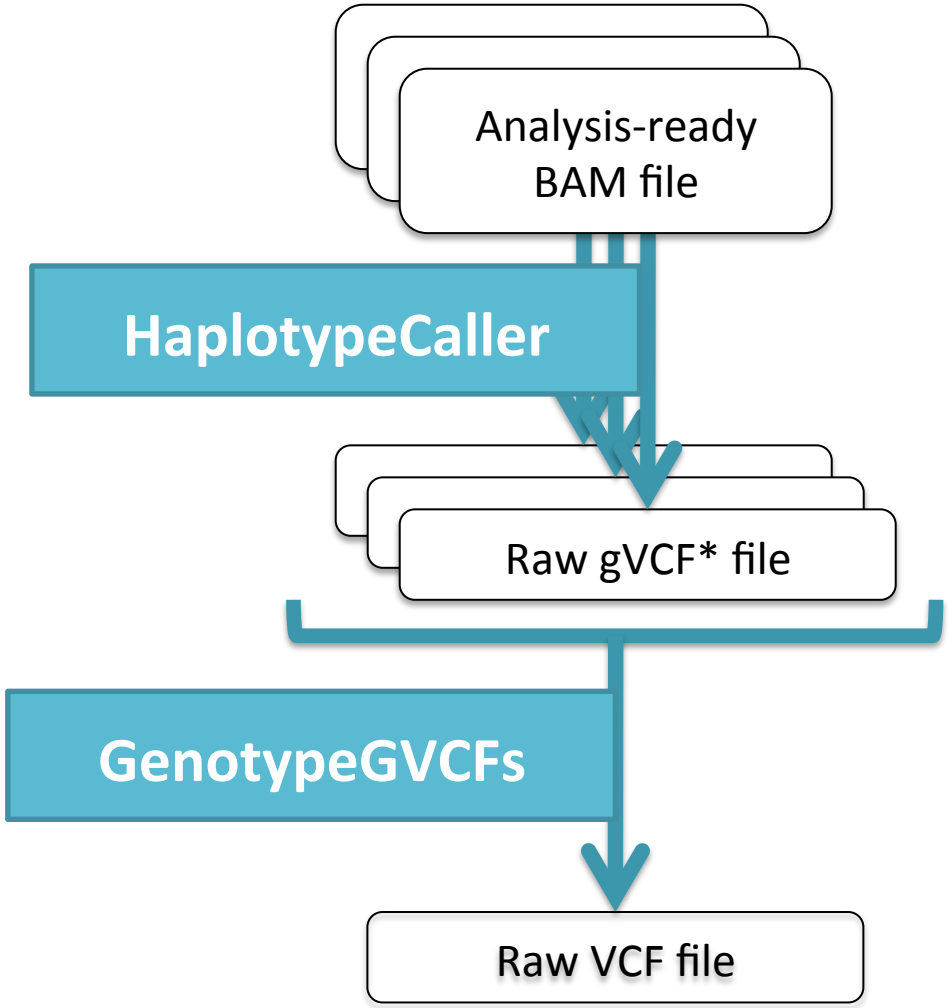
	Site	Variant	Sample 1	Sample 2	...	Sample N
SNP	***1:1000***	***A/C***	***0/0 0,10,100***	***0/1 20,0,200***	...	***0/0 0,100,255***
Indel	1:1050	T/TC	0/0 0,10,100	0/0 0,20,200	...	1/0 255,0,255
SNP	***1:1100***	***T/G***	***0/0 0,10,100***	***0/1 20,0,200***	...	***0/0 0,100,255***
	...	...	...	...	...	...
SNP	***X:1234***	***G/T***	***0/1 10,0,100***	***0/1 20,0,200***	...	***1/1 255,100,0***

Genotypes:
0/0 ref
0/1 het
1/1 hom-alt

Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data



The joint discovery workflow in practice




```
java –jar GenomeAnalysisTK.jar
        –T HaplotypeCaller \
        –R human.fasta \
        –I sample1.bam \
        –o sample1.g.vcf \
      [ –L exome_targets.intervals \ ]
        –ERC GVCF
```

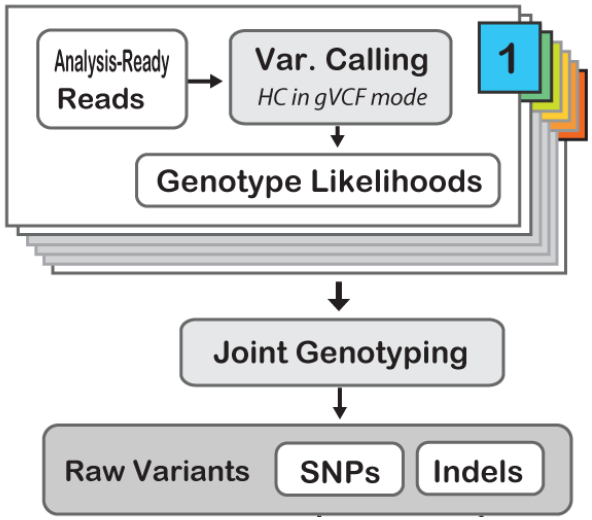
```
java –jar GenomeAnalysisTK.jar
        –T GenotypeGVCFs \
        –R human.fasta \
        –V sample1.g.vcf \
        –V sample2.g.vcf \
        –V sampleN.g.vcf \
        –o output.vcf
```

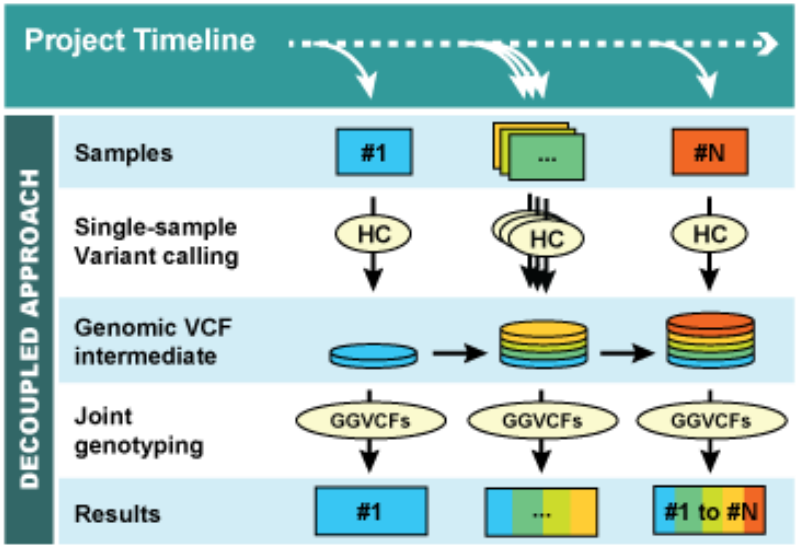
If >200 samples, combine in batches first using CombineGVCFs

And that is how we can scale joint discovery to eleventy thousand samples

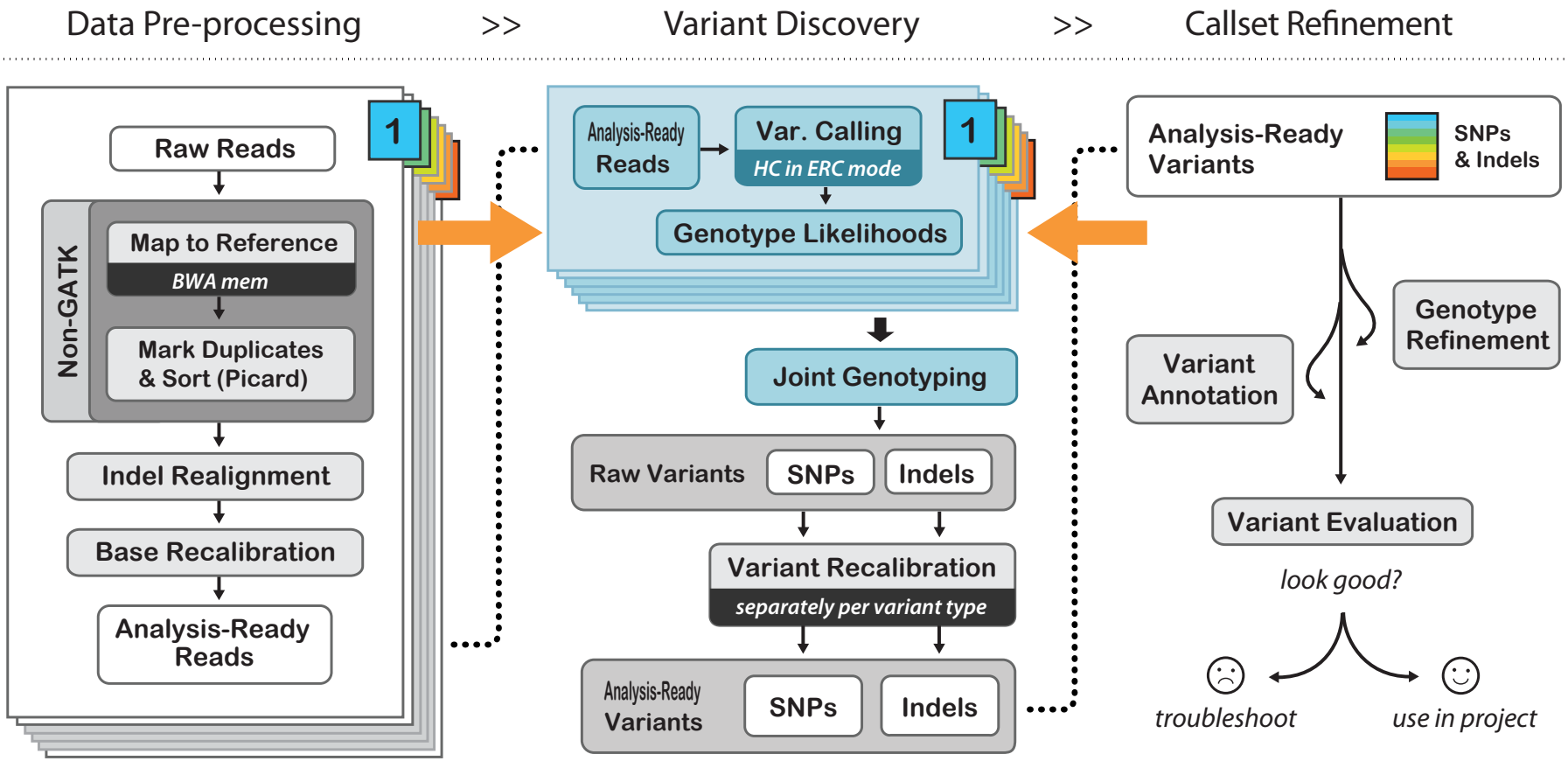
Scalable over sample size

+ Incremental over time





You are here in the GATK Best Practices workflow for **germline variant discovery**



Further reading

http://www.broadinstitute.org/gatk/guide/best-practices

http://www.broadinstitute.org/gatk/guide/article?id=1237

https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php

https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_variantutils_GenotypeGVCFs.php