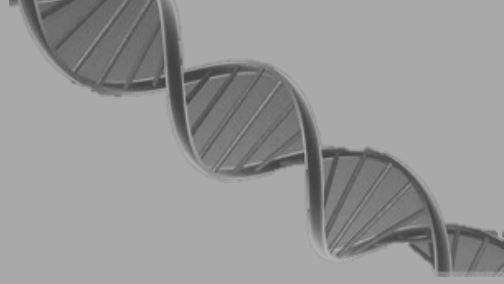




# Introduction to NGS Variant-Calling

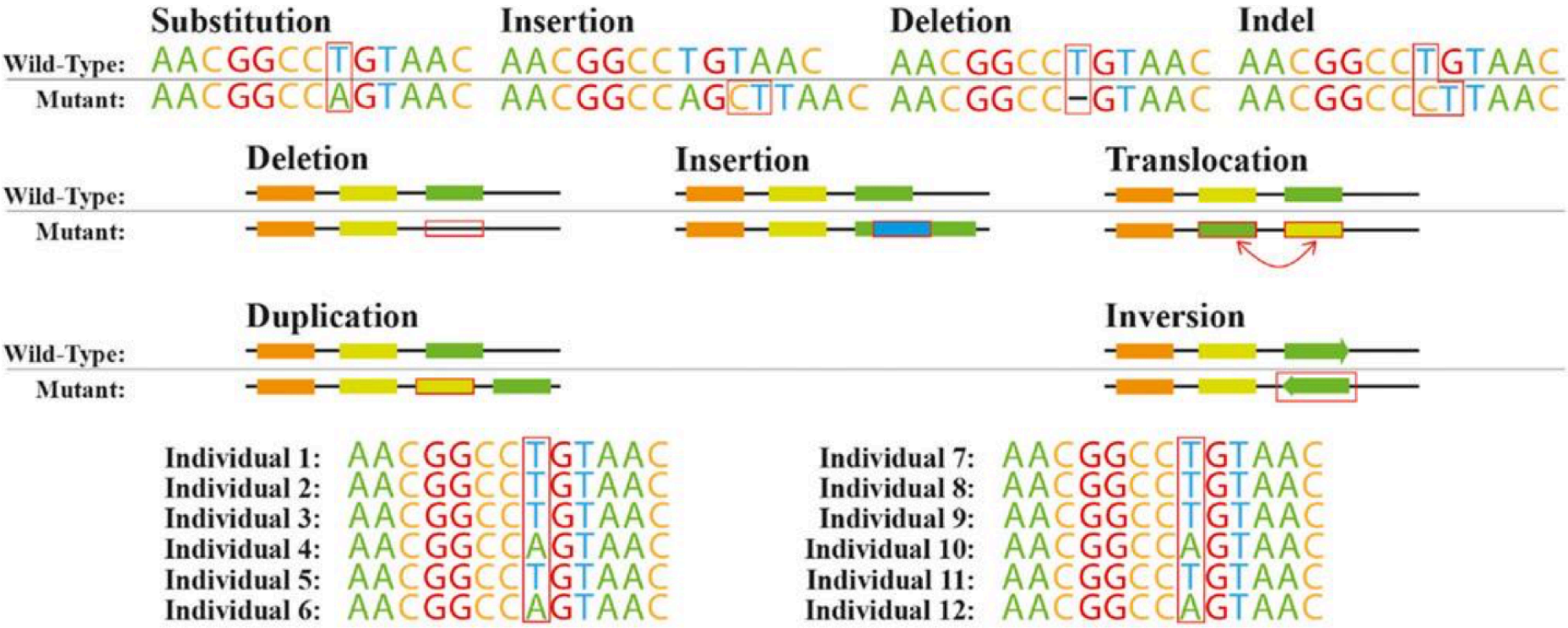
Thibault Latrille, Anouk Necsulea,
Annabelle Haudry, Maud Gautier

# Plan

**Part 1:**
Overview of variant-calling

**Part 2:**
GATK workflow

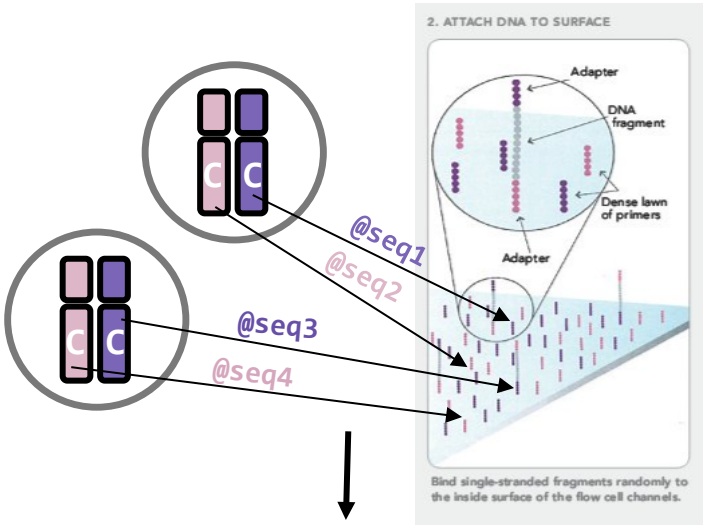
# Different types of variants




*Cardoso et al (2015)*
*DOI: 10.3389/fbioe.2015.00013*

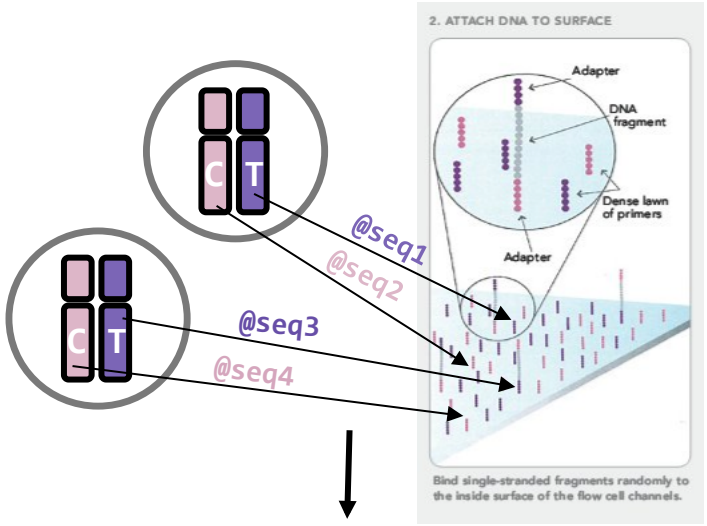
# Concrete view: Variants on a flowcell

- Scenario 1: Homozygote




```
Ref    ATCGGGTACCATCCAATCATTACC
          GGCACCATCCAAT
@seq2  ATCGGGCACCAT
@seq3     GGGCACCATCCAA
             TCGGGCACCATC
@seq4     CGGGCACCATCC
@seq1     CGGGCACCATCCA
```

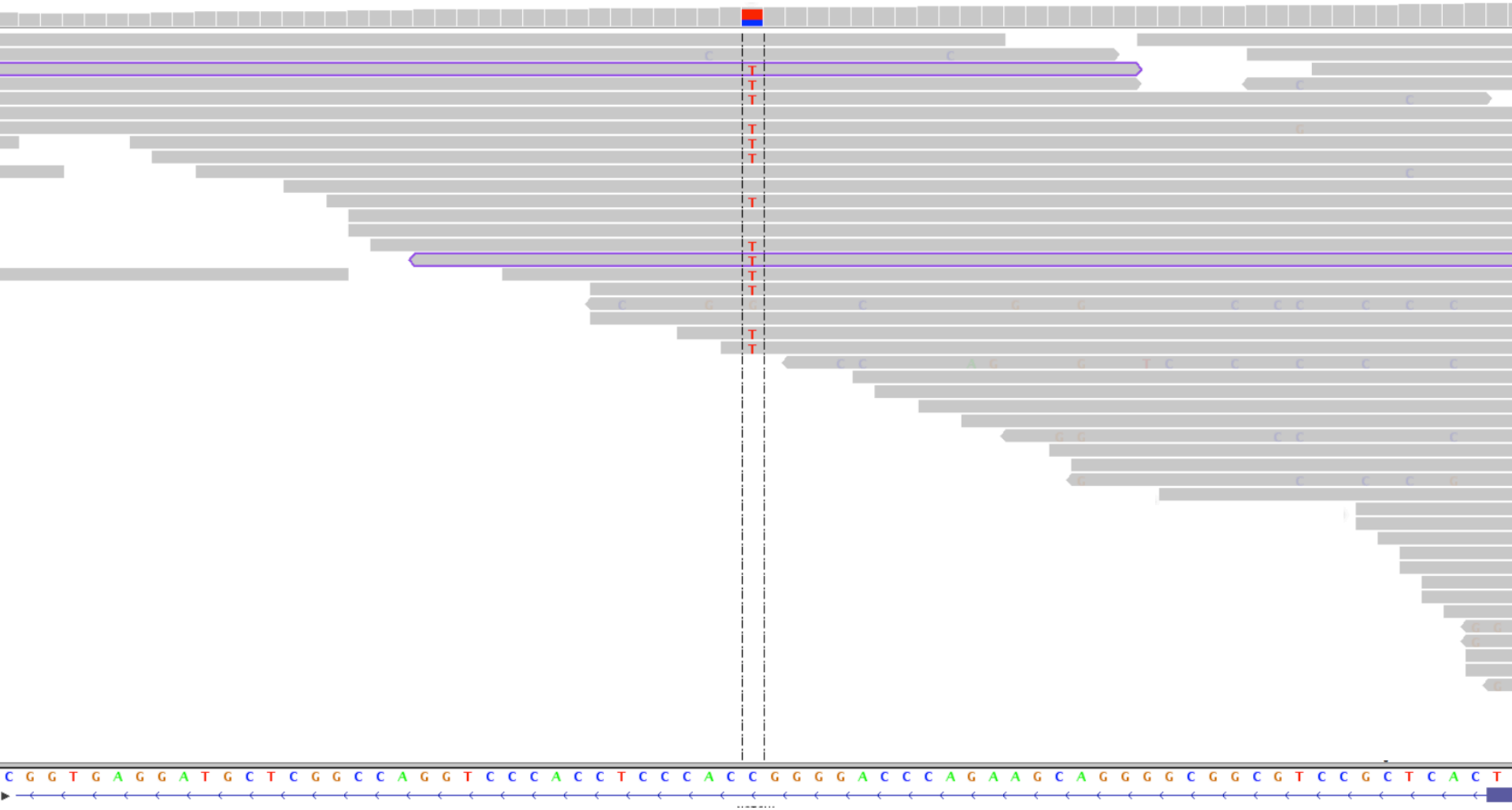
- Scenario 2: Heterozygote




```
Ref    ATCGGGTACCATCCAATCATTACC
          GGTACCATCCAAT
@seq2  ATCGGGCACCAT
@seq3     GGGTACCATCCAA
             TCGGGCACCATC
@seq4     CGGGCACCATCC
@seq1     CGGGTACCATCCA
```

*From Aaron Quinlab*

# Visualising variants in IGV



# Complexity of variant-calling

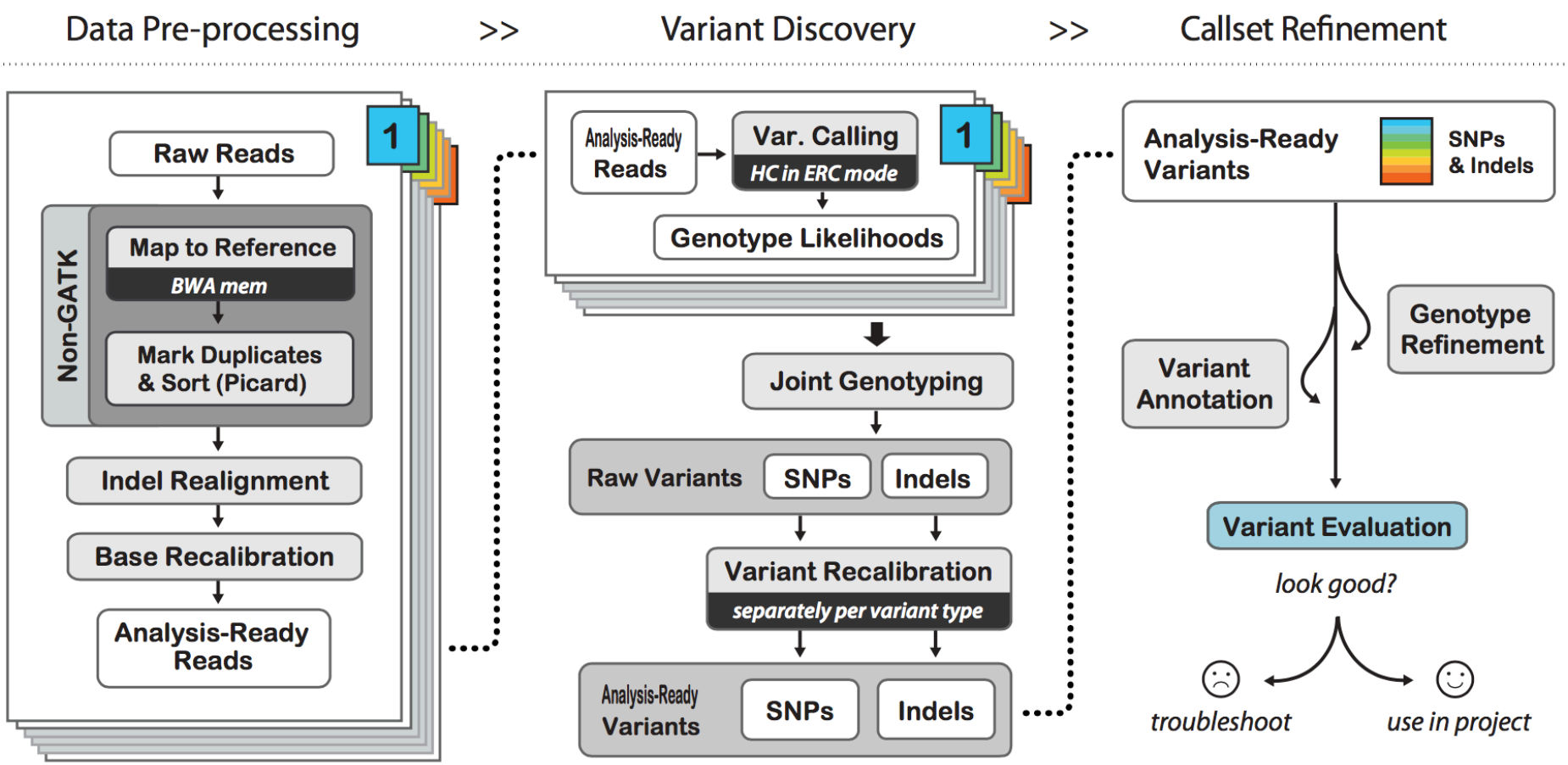
- Identify variants **relative to a reference** (hg38 for humans)
- Complexity of variant-calling: false positives => need to distinguish between true genetic variation and false positives
- False positives may come from:
  - PCR artifacts (→ MarkDuplicates)
  - Sequencing artifacts (→ Remove if low quality)
  - Alignment (→ Locally realign)
  - …

# Plan

**Part 1:**
Overview of variant-calling

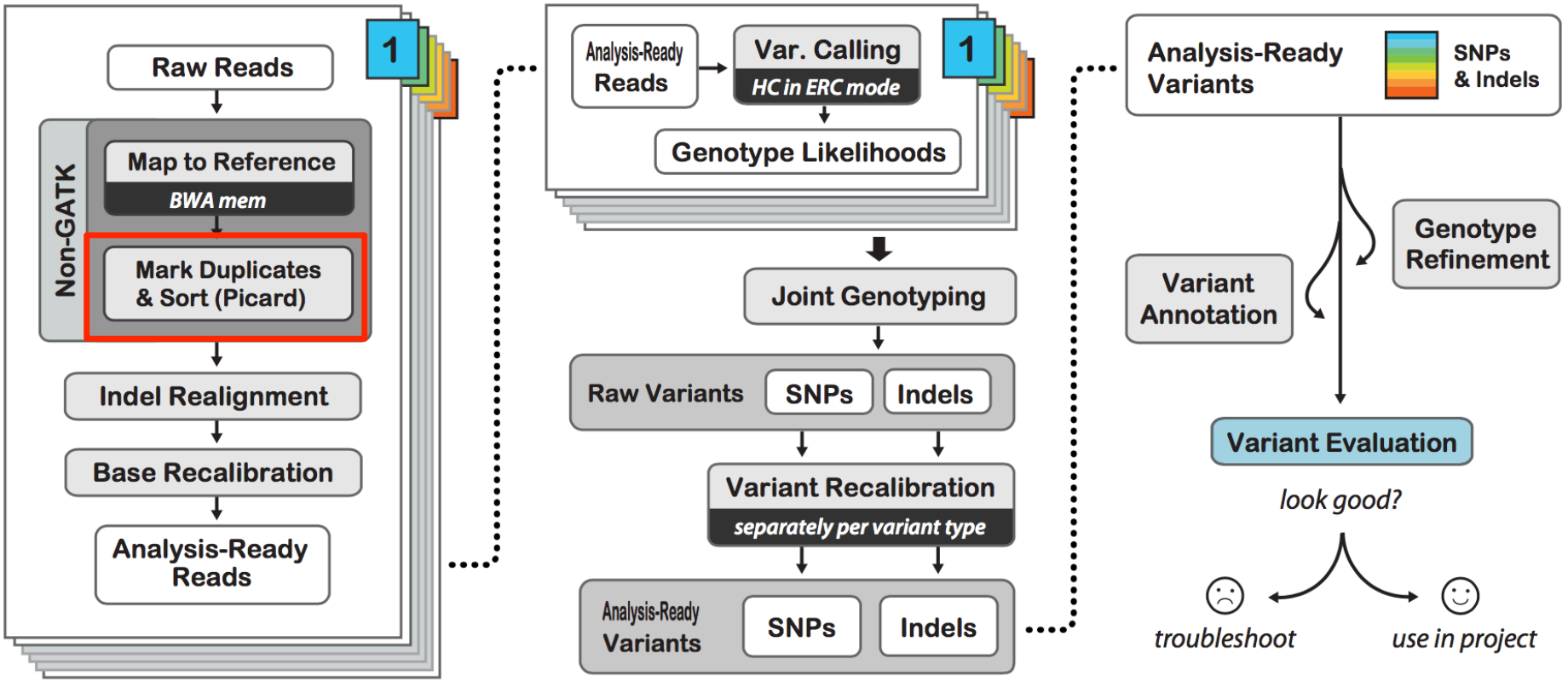
**Part 2:**
GATK workflow

# GATK workflow




*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# GATK workflow

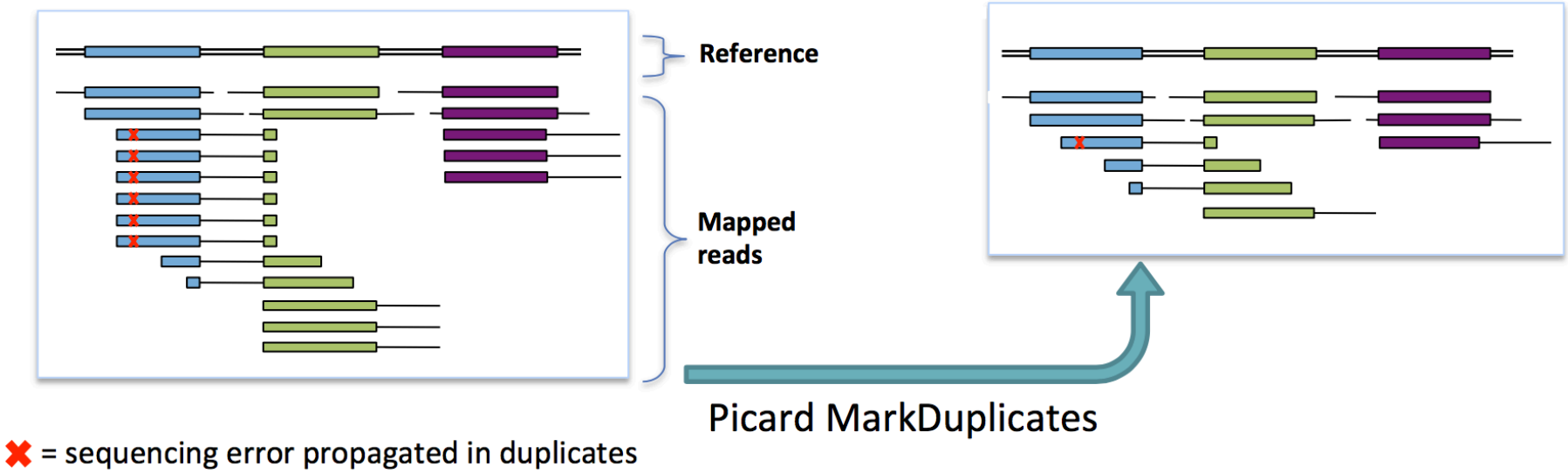



*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# Data pre-processing - MarkDuplicates

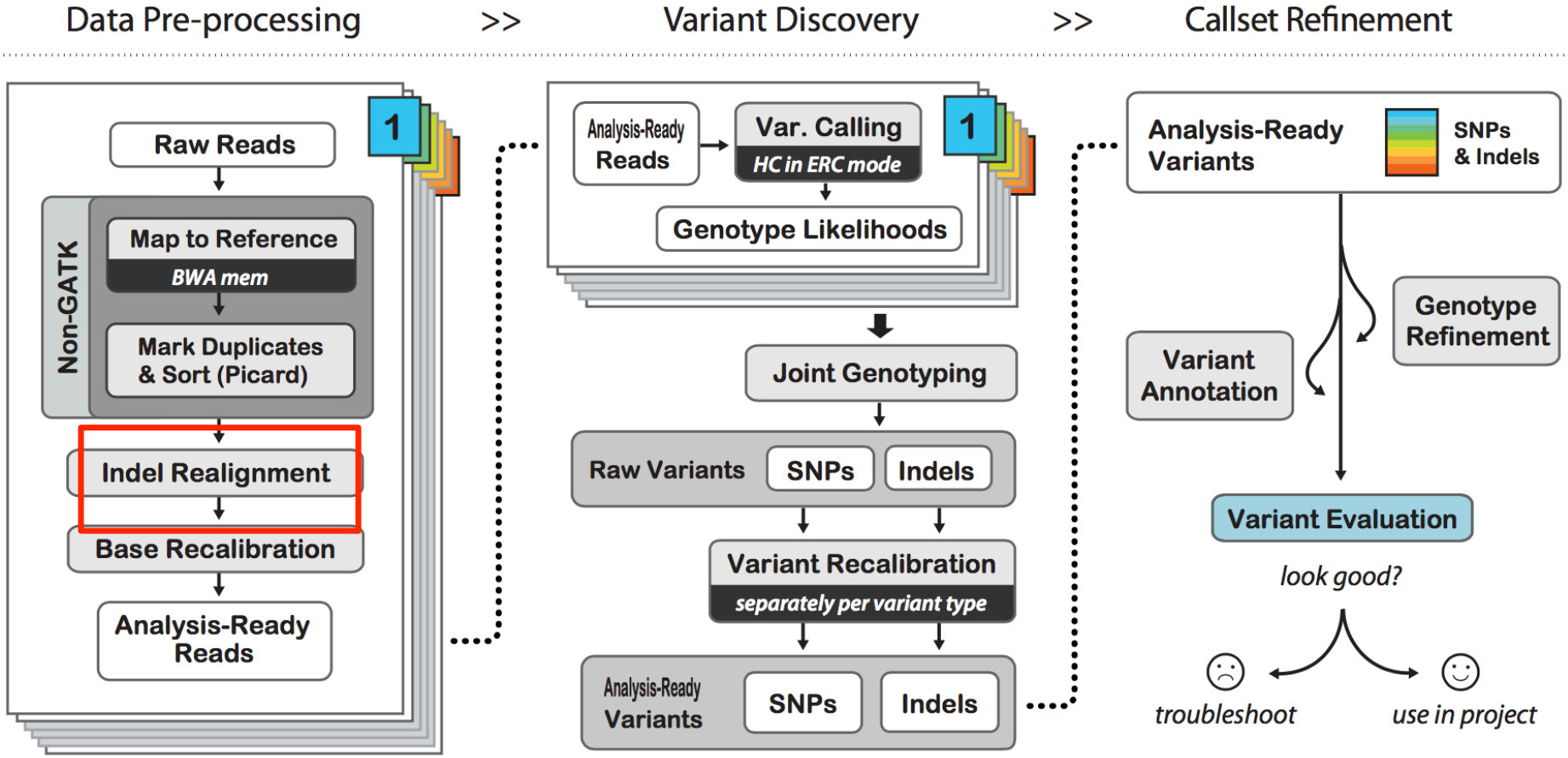
- Duplicates = non-independent measurements of a sequence

  → Must be removed



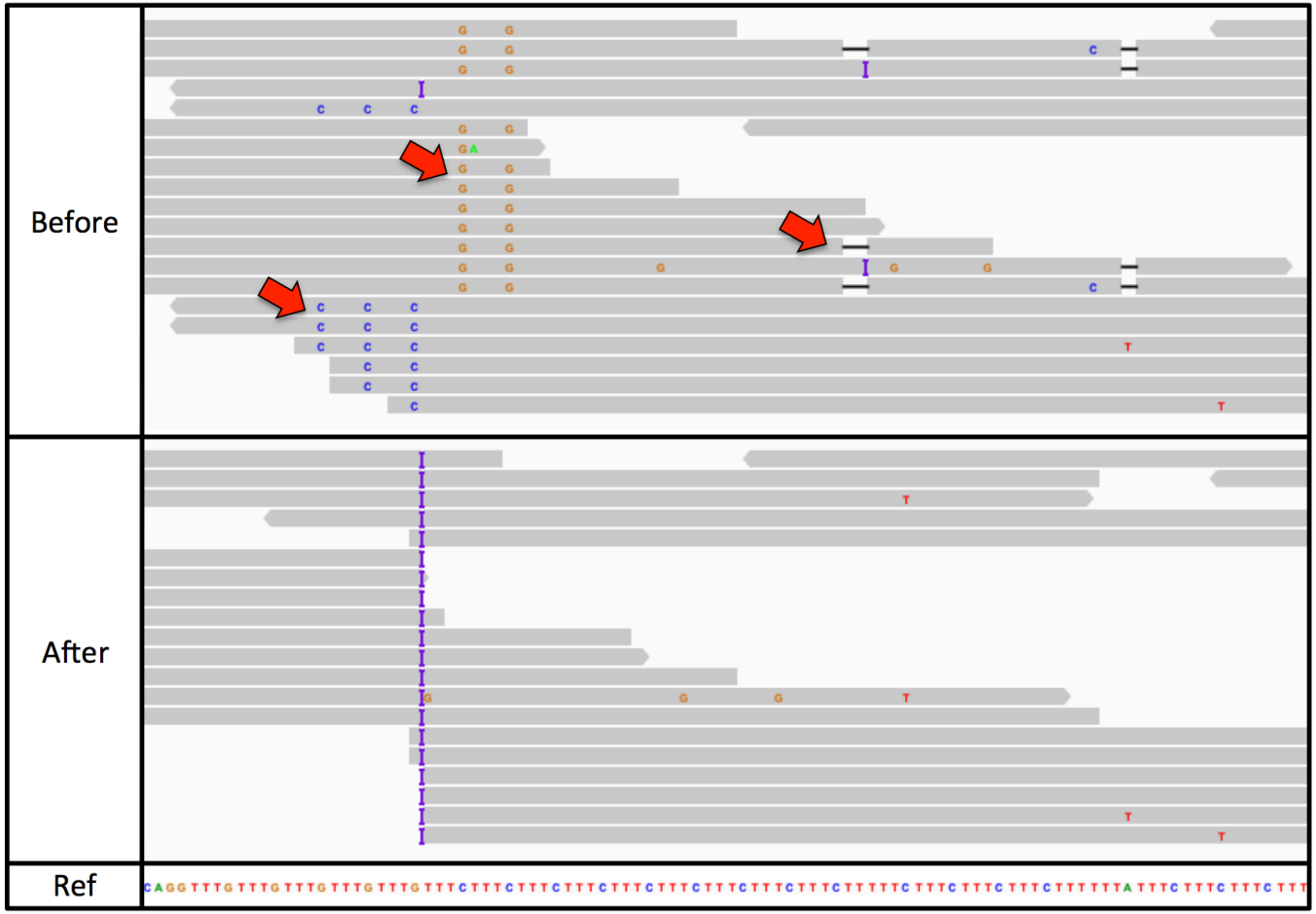

*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# GATK workflow



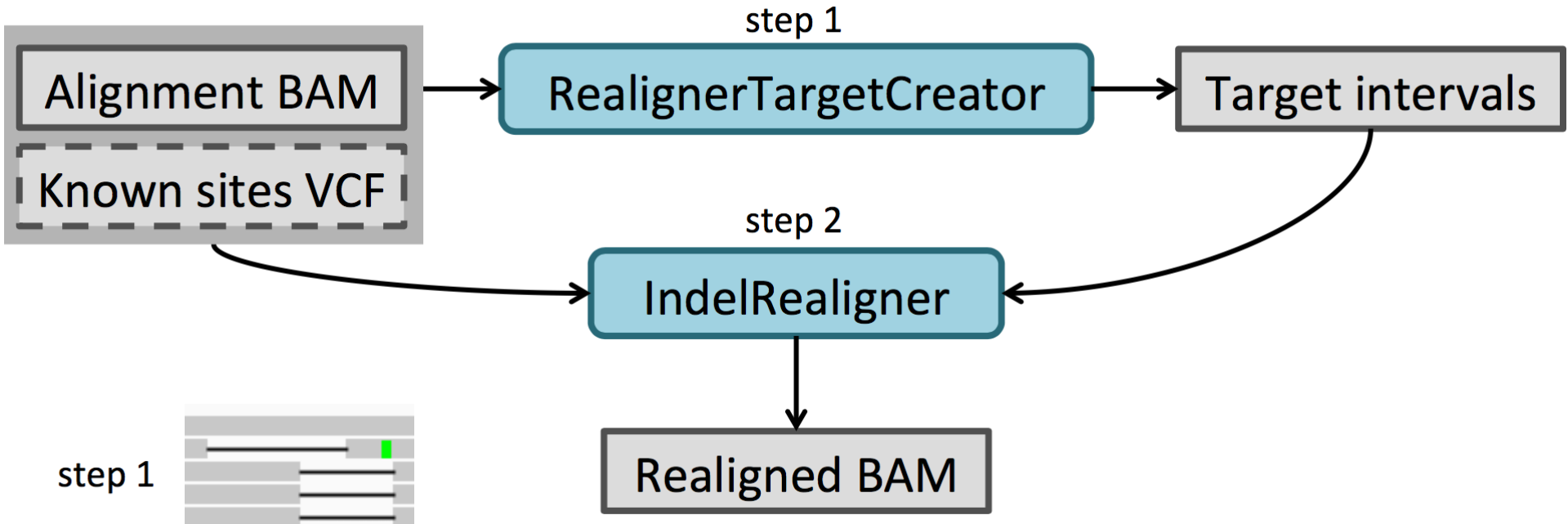

*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

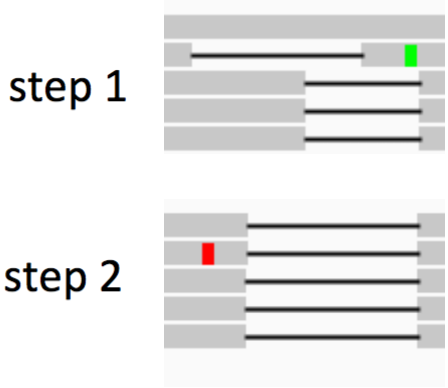
# Indel realignment




*GATK Best Practices for Variant Discovery*
*(https://software.broadinstitute.org/gatk/download/workshops)*

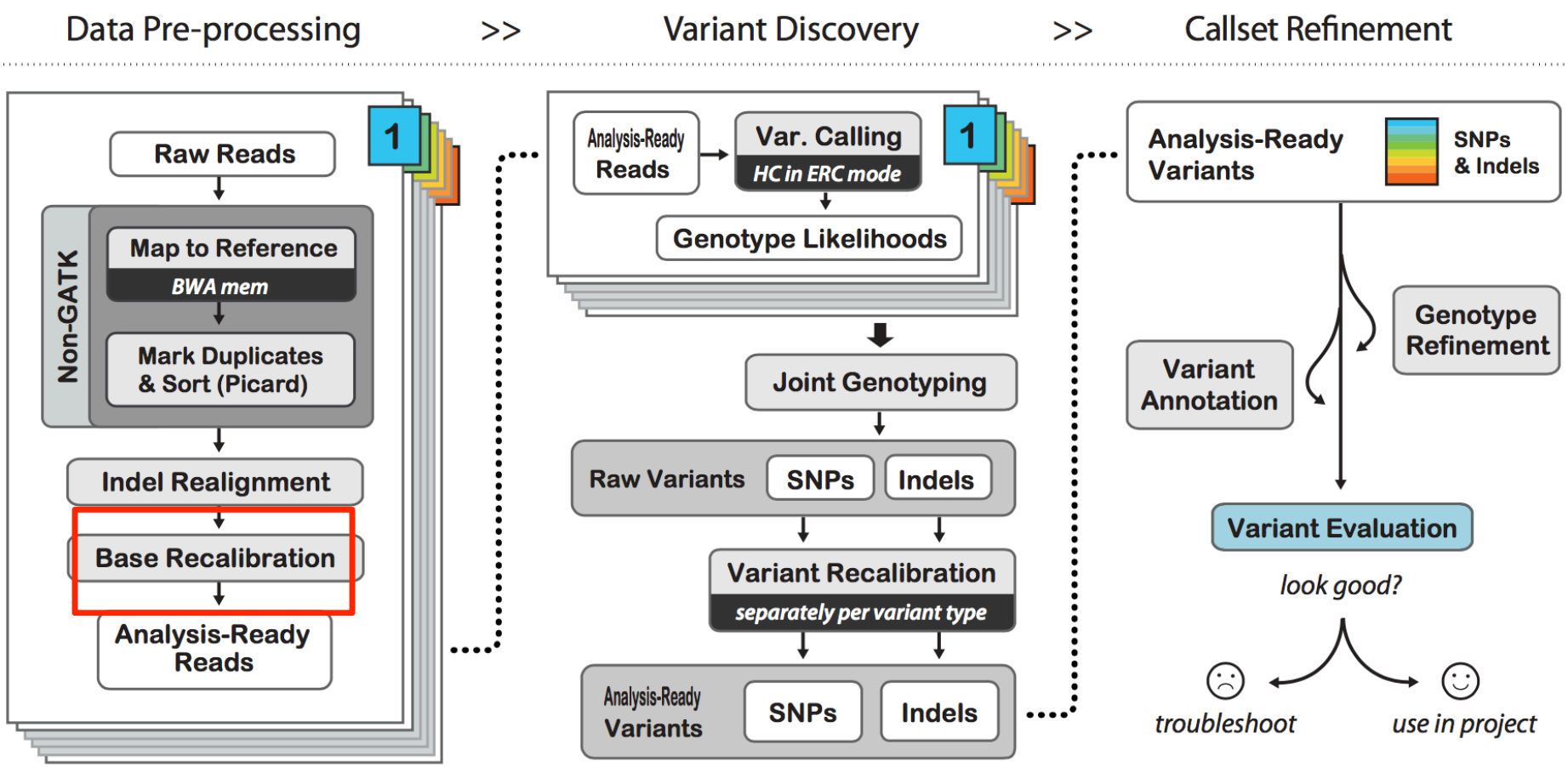
# Indel realignment workflow




*GATK Best Practices for Variant Discovery*
*(https://software.broadinstitute.org/gatk/download/workshops)*

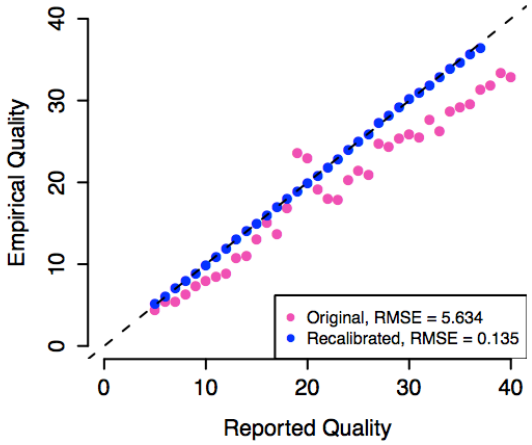
# GATK workflow

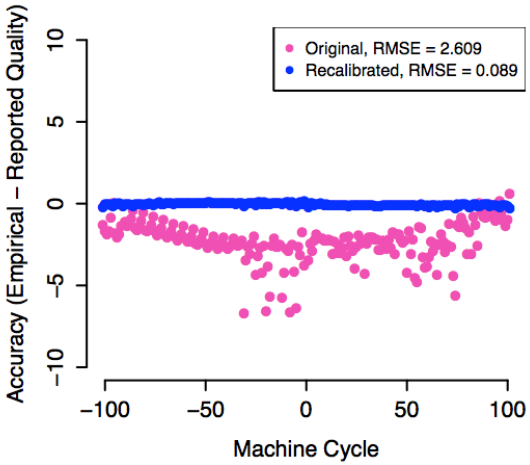


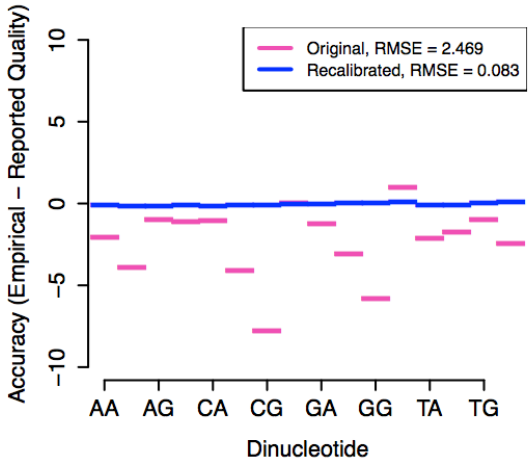

*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# Base Quality Score Recalibration (BQSR)

- Systematic biases: quality of base calls depend on nucleotidic context, machine cycle…
- → Machine learning algorithm to find how error varies with basecall features. Then, apply recalibration.

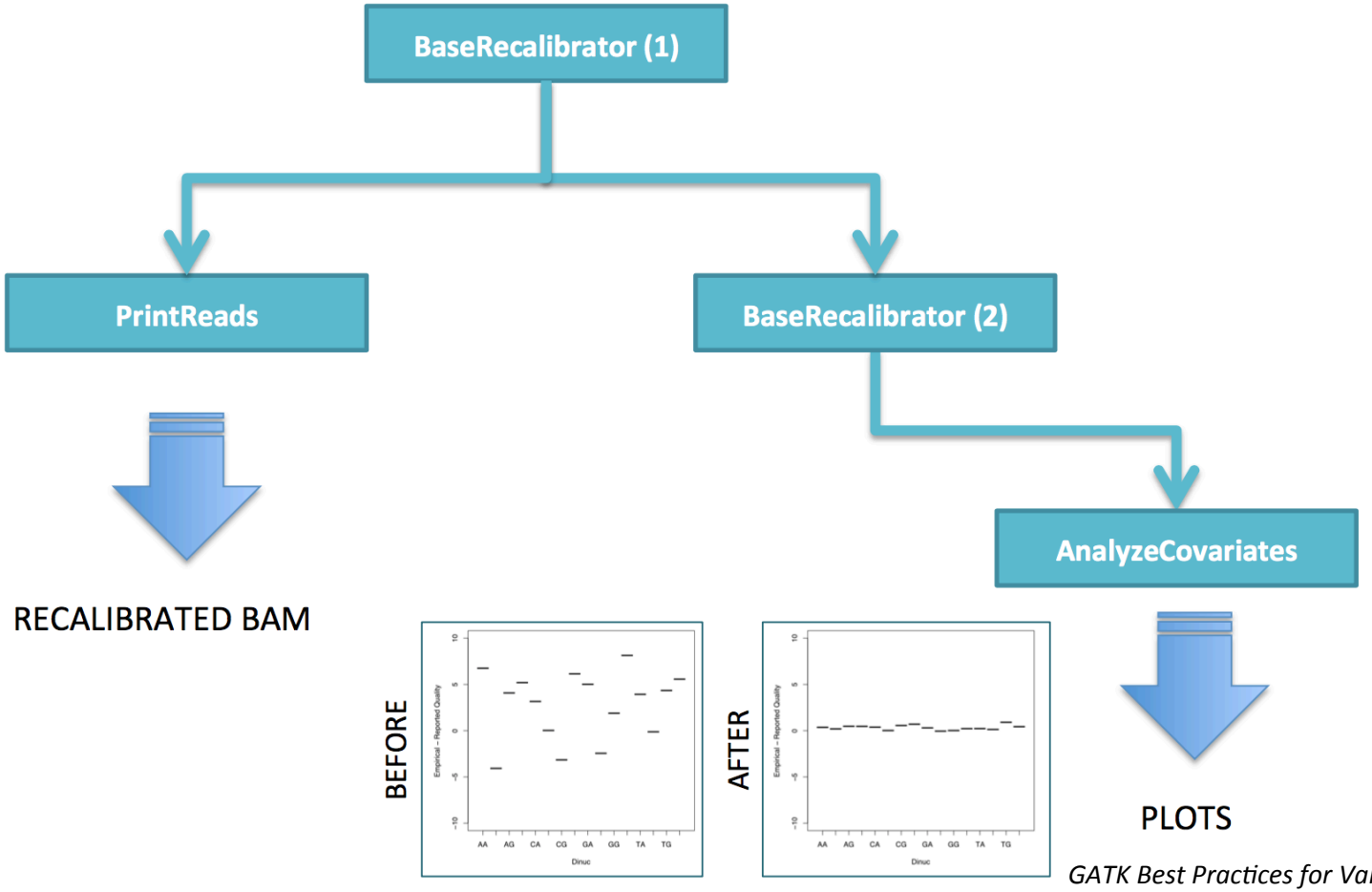


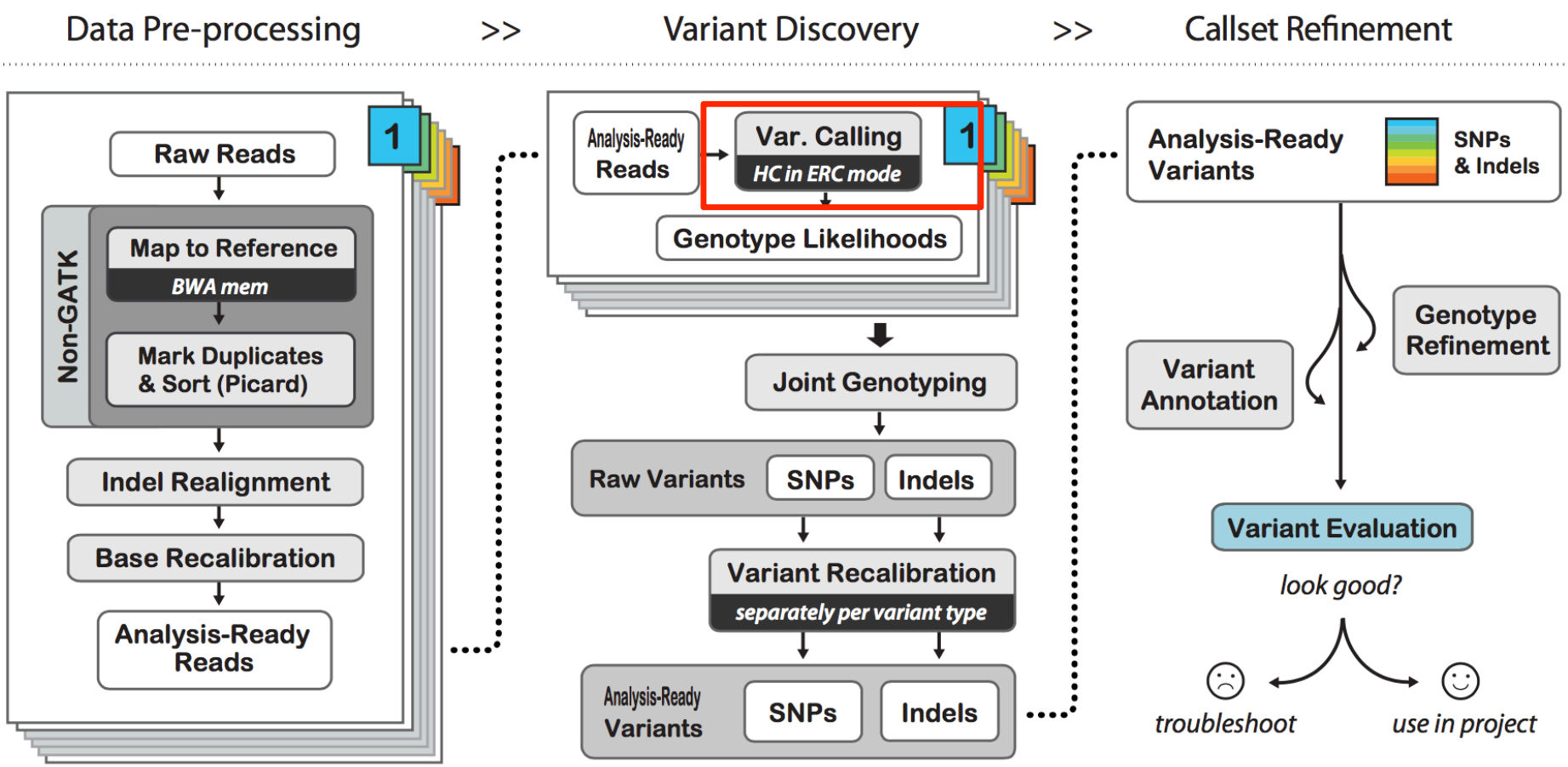

*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# BQSR workflow



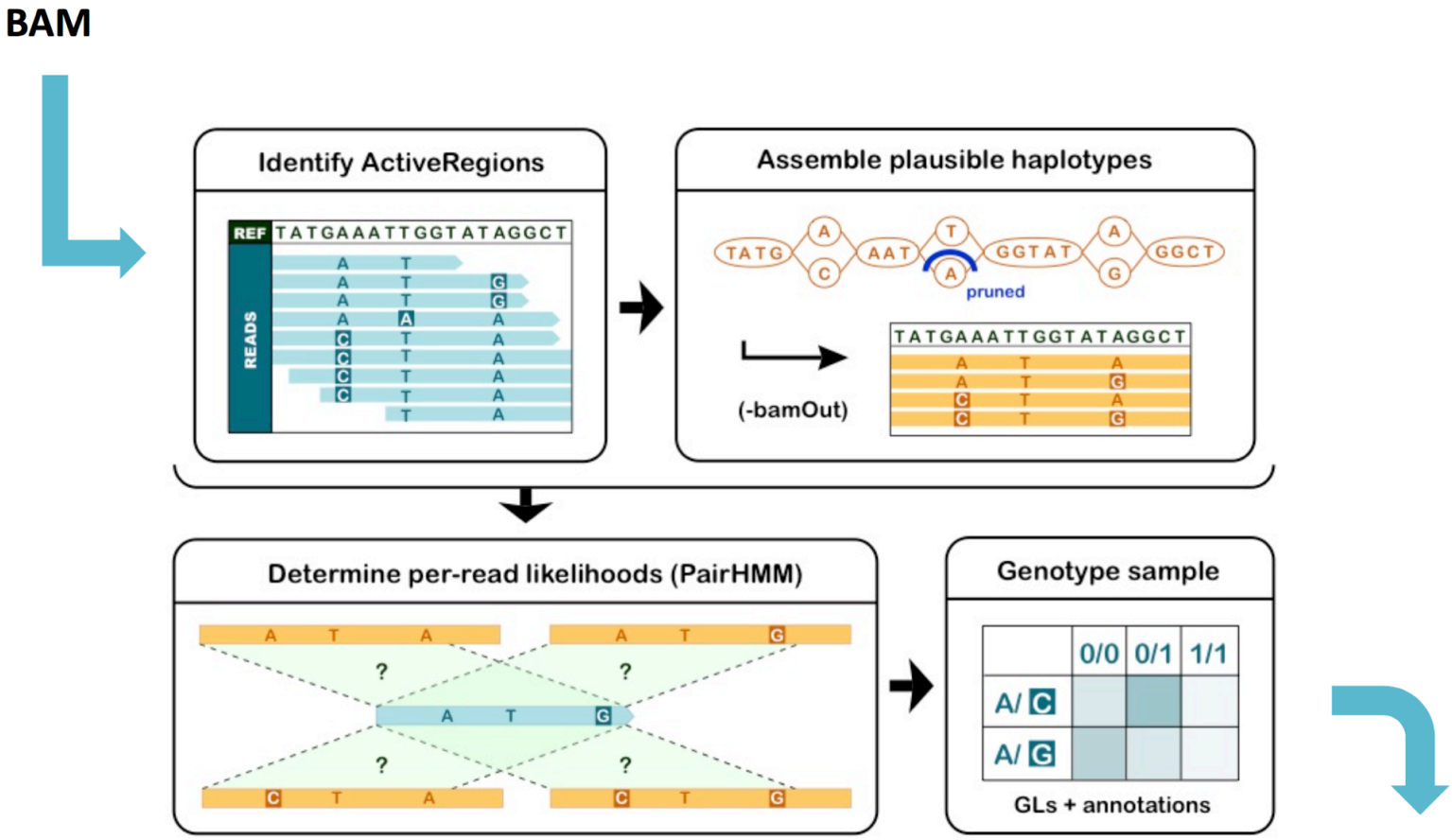

*GATK Best Practices for Variant Discovery*
*(https://software.broadinstitute.org/gatk/download/workshops)*

# GATK workflow



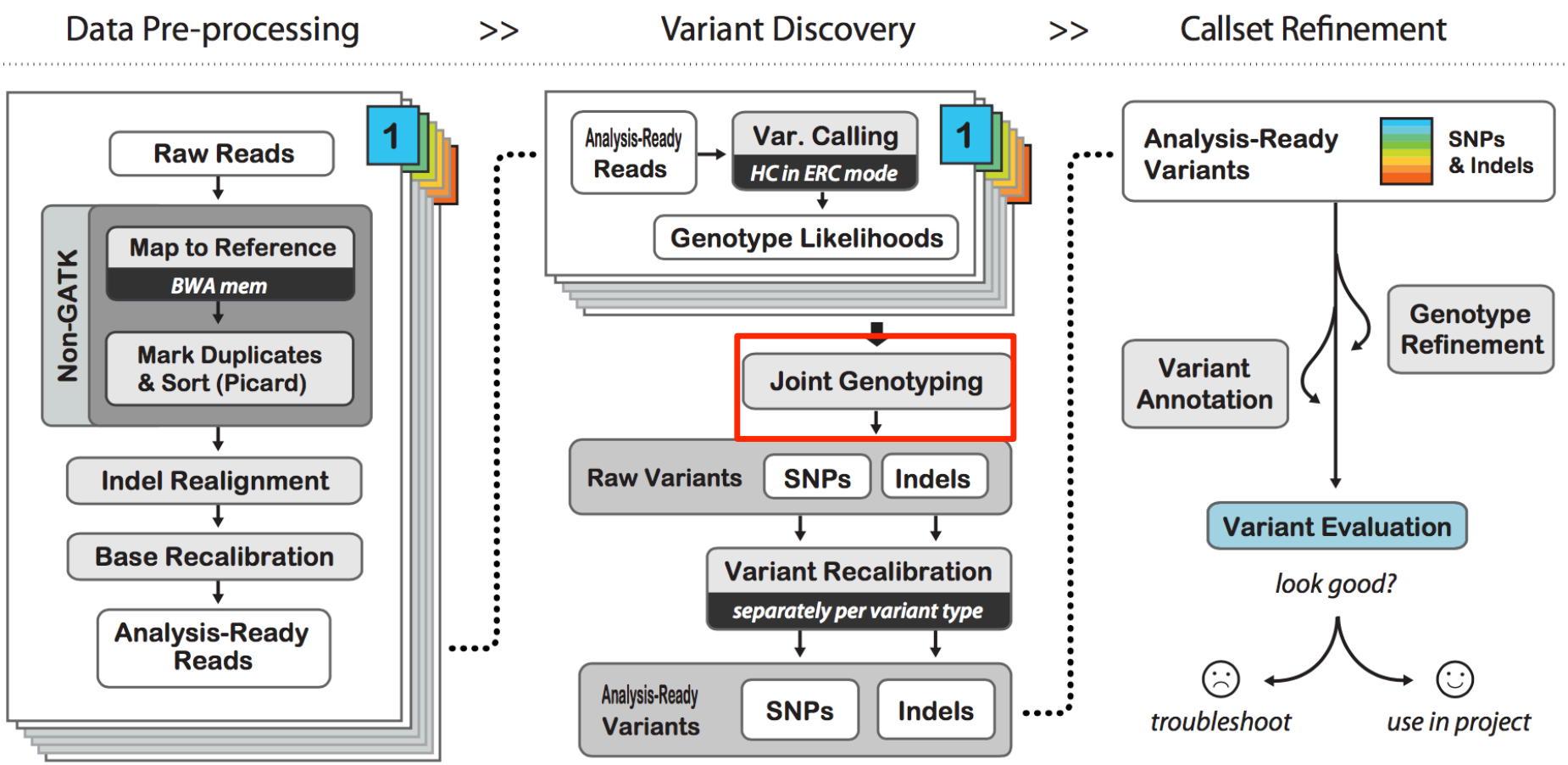

*GATK Best Practices for Variant Discovery*
*(https://software.broadinstitute.org/gatk/download/workshops)*

# Variant-calling: Haplotype-Caller



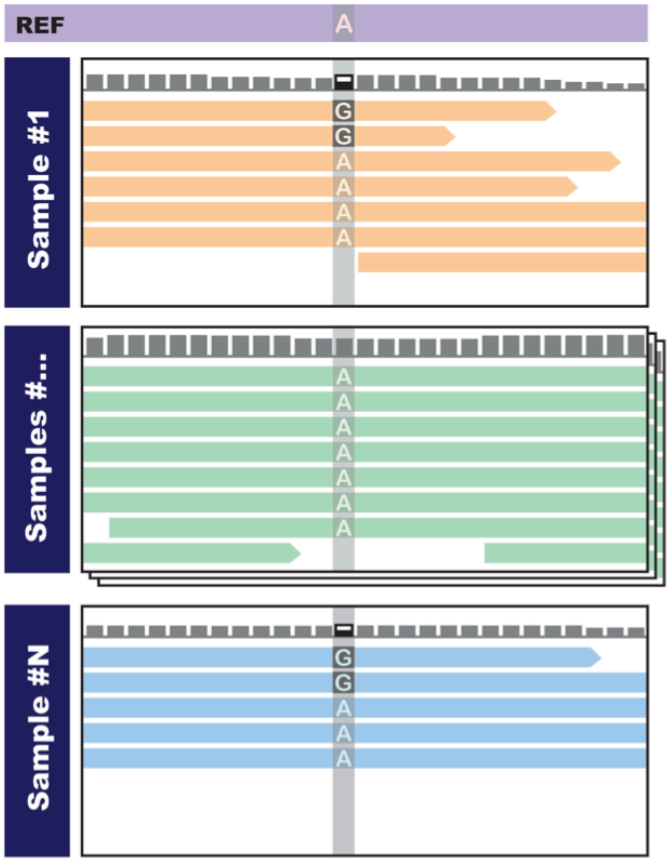

*GATK Best Practices for Variant Discovery*
*(https://software.broadinstitute.org/gatk/download/workshops)*

# GATK workflow




*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# Joint-genotyping

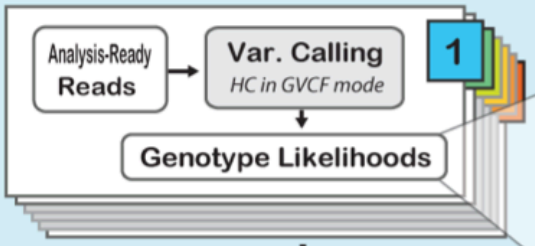



- When only 1 sample => "G" may be considered errors.
- When all samples => More confidence in the calling of A/G variant.

Joint callset → empowered analysis

*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# gVCF VS regular VCF




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

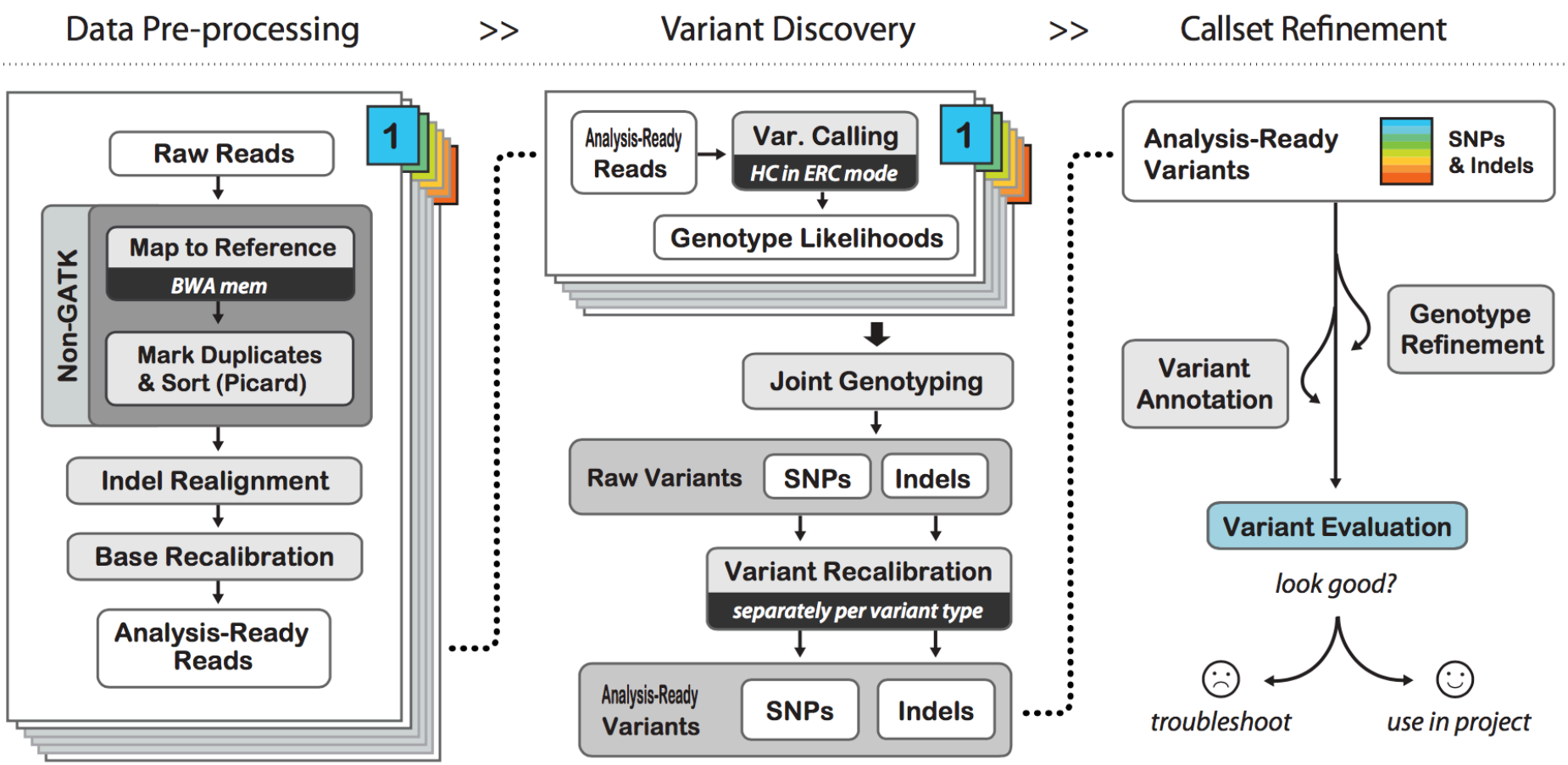
```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 . A <NON_REF> . . END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 . C T,<NON_REF> 326.77 . BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
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
```

*GATK Best Practices for Variant Discovery (https://software.broadinstitute.org/gatk/download/workshops)*

# Conclusion on the GATK workflow




Thanks for your attention