

Introduction to NGS Variant-Calling

Thibault Latrille, Anouk Necsulea,
Annabelle Haudry, Maud Gautier



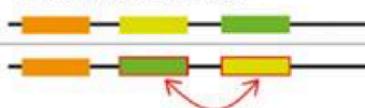
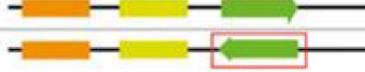
Plan

Part 1:
Overview of variant-calling

Part 2:
GATK workflow



Different types of variants

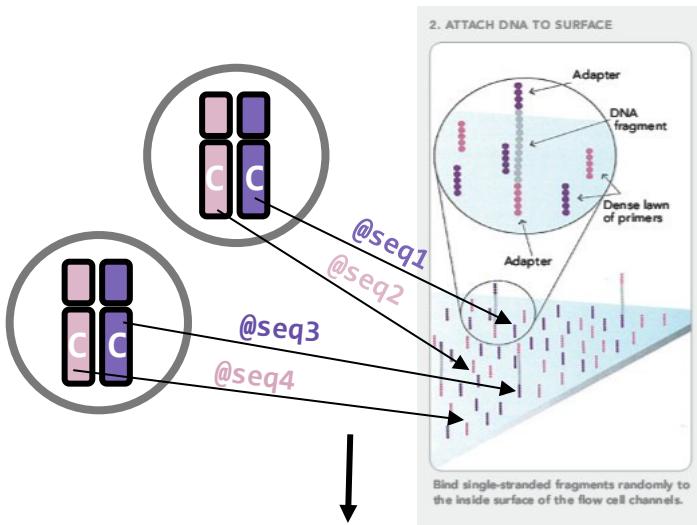
	Substitution	Insertion	Deletion	Indel
Wild-Type:	AACGGCC T GTAAC	AACGG C CTGTAAC	AACGGCC T GTAAC	AACGGCC T GTAAC
Mutant:	AACGG C AGTAAC	AACGG C AG T TAAC	AACGG C GTAA C	AACGG C GT T AA C
	Deletion	Insertion	Translocation	
Wild-Type:				
Mutant:				
	Duplication		Inversion	
Wild-Type:				
Mutant:				
Individual 1:	AACGG C CTGTAAC		Individual 7:	AACGG C CTGTAAC
Individual 2:	AACGG C CTGTAAC		Individual 8:	AACGG C CTGTAAC
Individual 3:	AACGG C CTGTAAC		Individual 9:	AACGG C CTGTAAC
Individual 4:	AACGG C AG T AA C		Individual 10:	AACGG C AG T AA C
Individual 5:	AACGG C CTGTAAC		Individual 11:	AACGG C CTGTAAC
Individual 6:	AACGG C AG T AA C		Individual 12:	AACGG C AG T AA C

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



Concrete view: Variants on a flowcell

- Scenario 1: Homozygote



Ref ATCGGG**T**ACCATCCAATCATTACC

GGCACC**AT**CCAAT

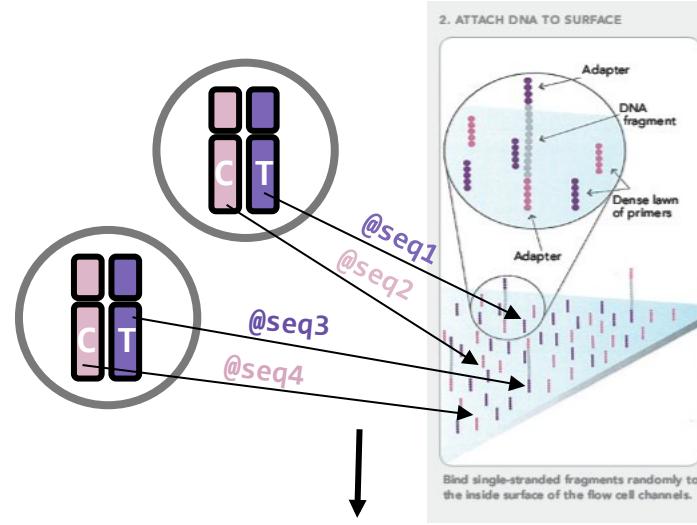
@seq2 ATCGGG**C**ACC**A**T

@seq3 GGGCACC**A**AT**CC**AA
TCGGGG**C**ACC**AT**C

@seq4 CGGG**C**ACC**A**AT**CC**

@seq1 CGGGCACC**A**AT**CC**AA

- Scenario 2: Heterozygote



Ref ATCGGG**T**ACCATCCAATCATTACC

GG**T**ACC**AT**CCAAT

@seq2 ATCGGG**C**ACC**A**T

@seq3 GGG**T**ACC**A**AT**CC**AA
TCGGGG**C**ACC**AT**C

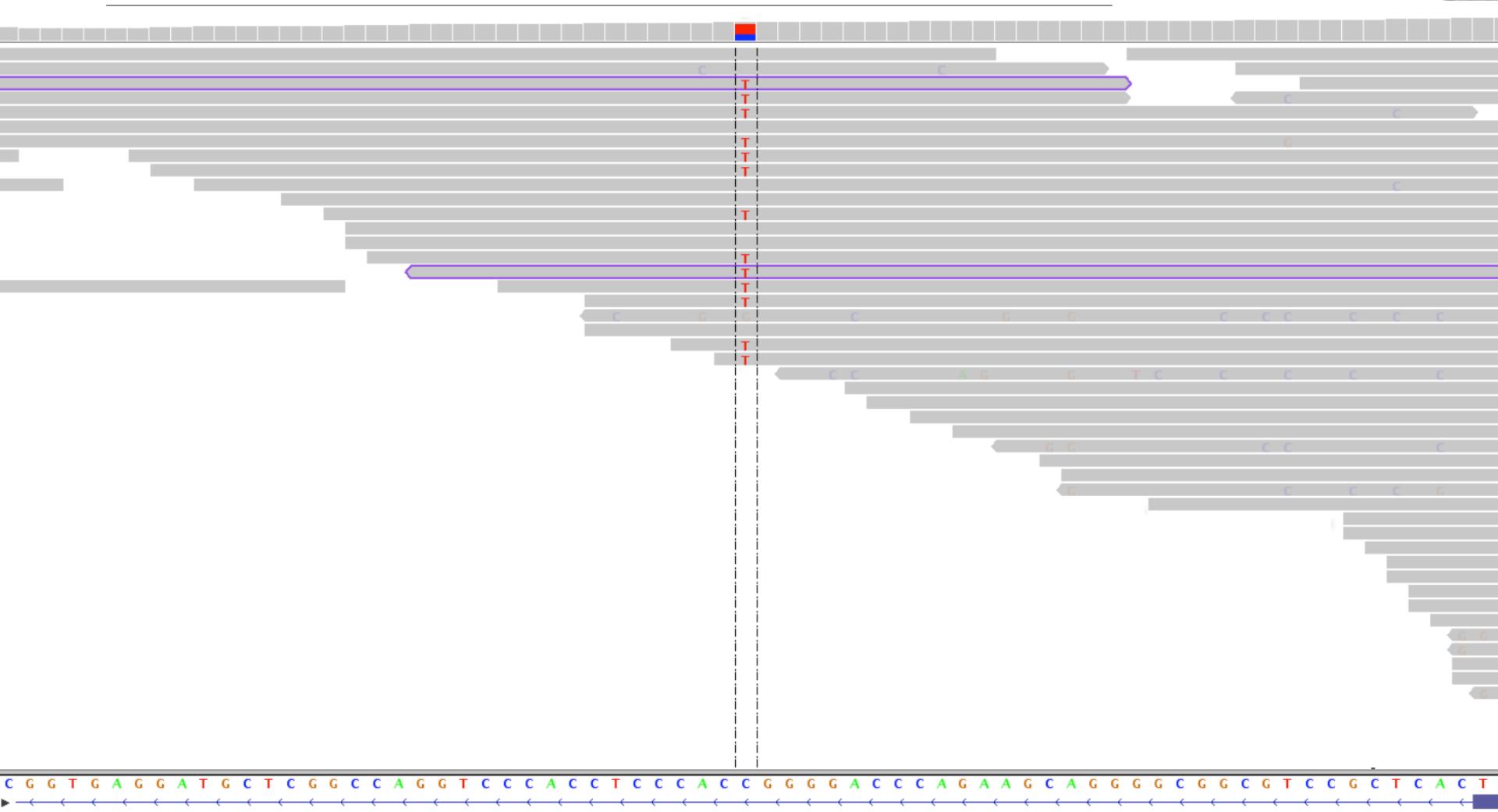
@seq4 CGGG**C**ACC**A**AT**CC**

@seq1 CGGG**T**ACCATCCA

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

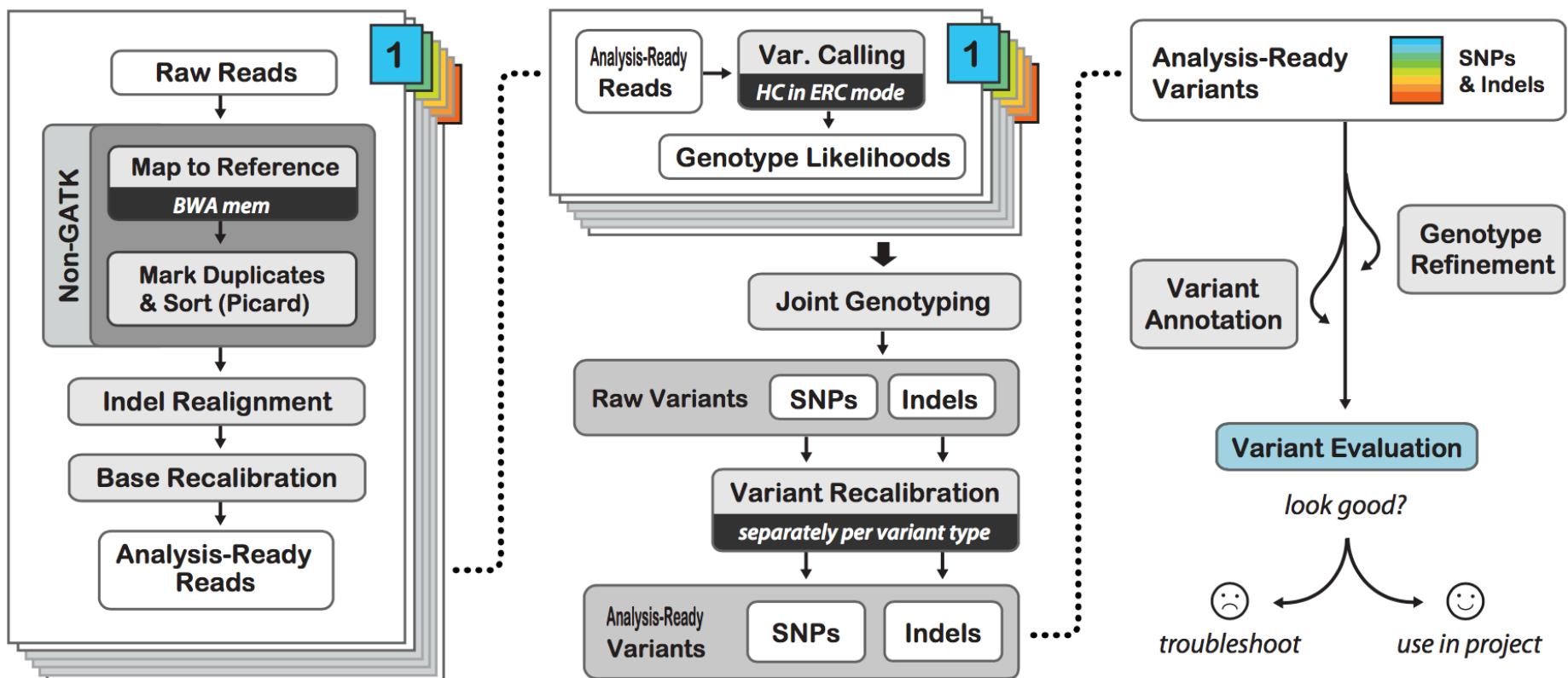
Data Pre-processing

>>

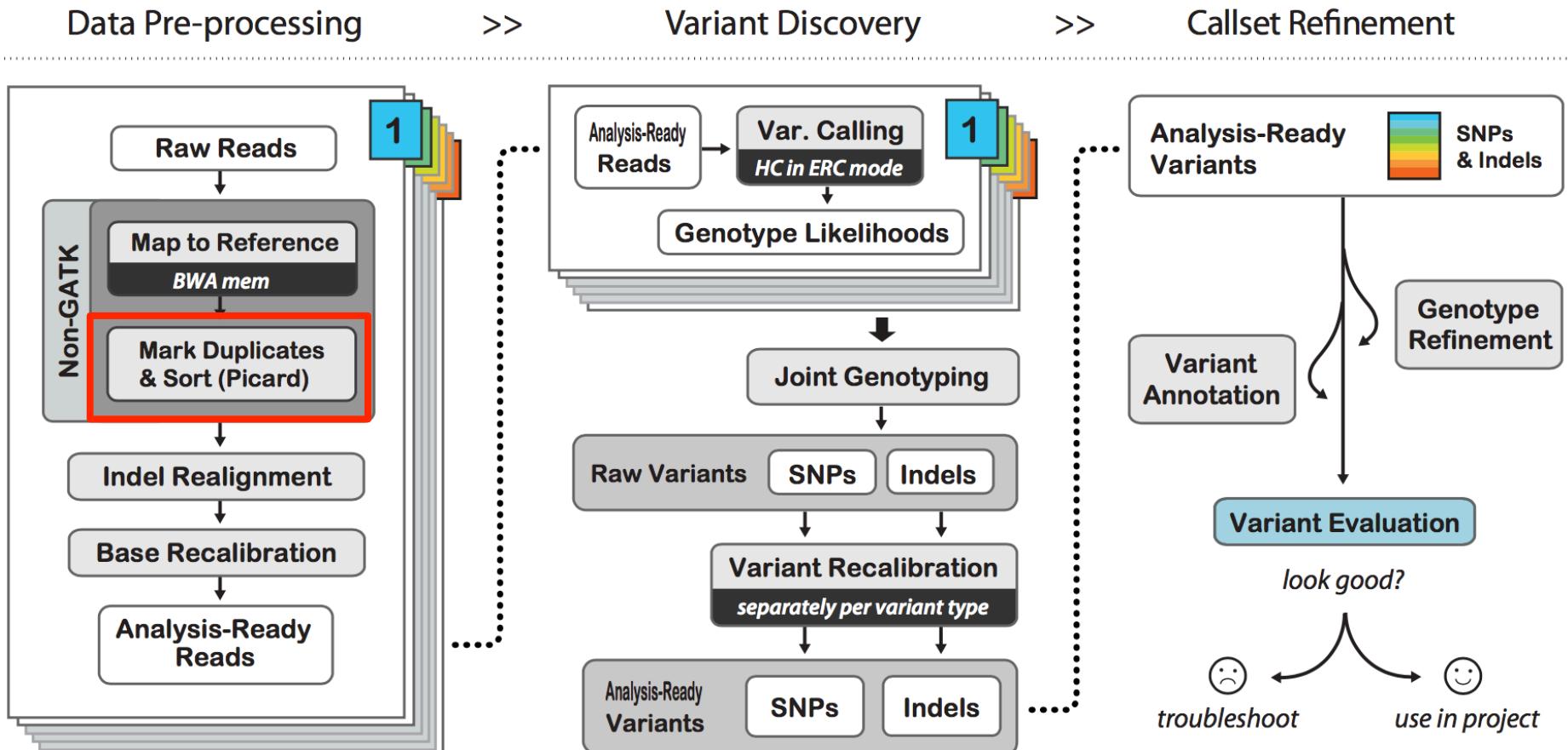
Variant Discovery

>>

Callset Refinement



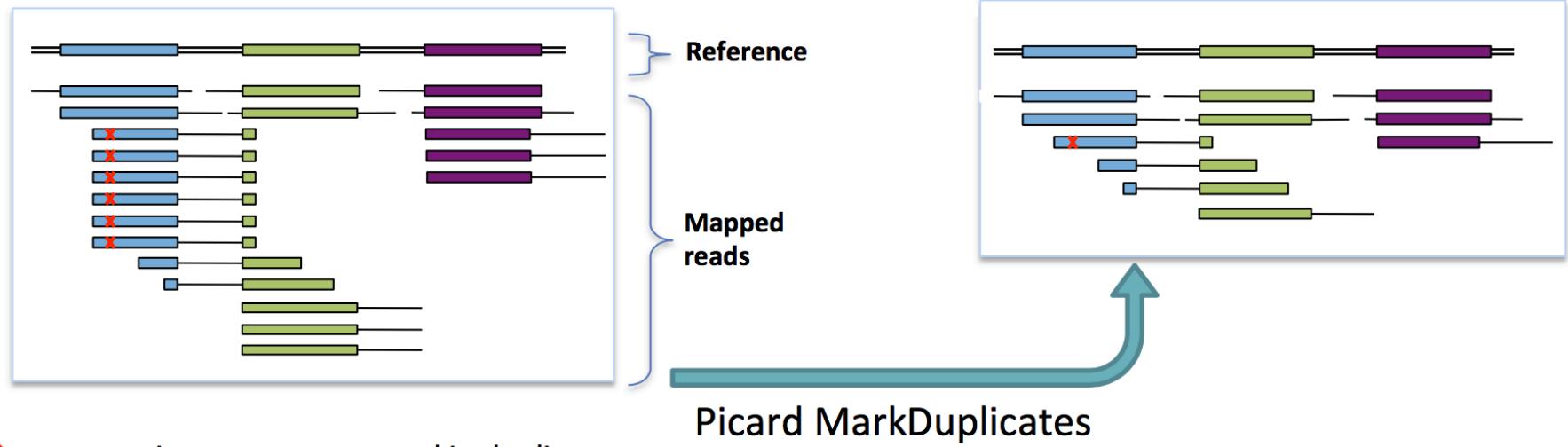
GATK workflow





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

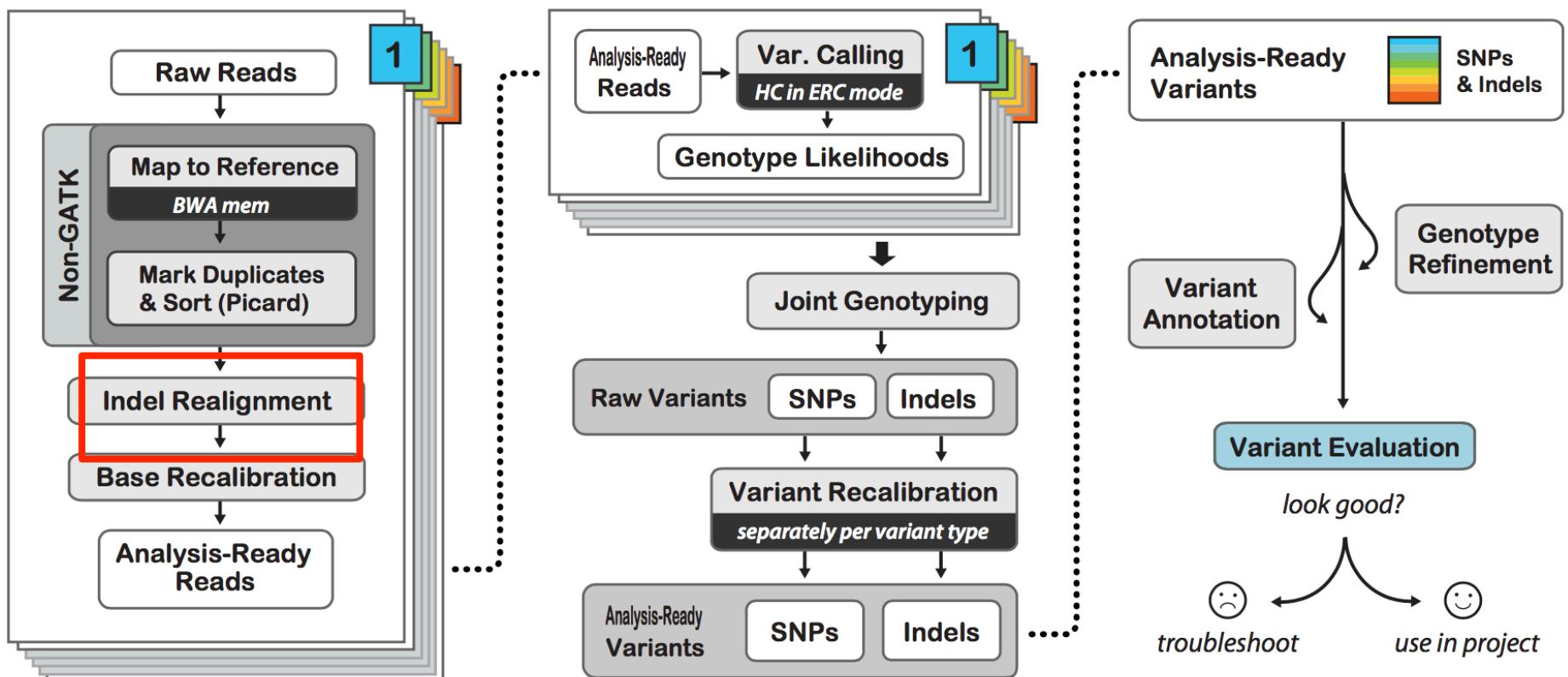
Data Pre-processing

>>

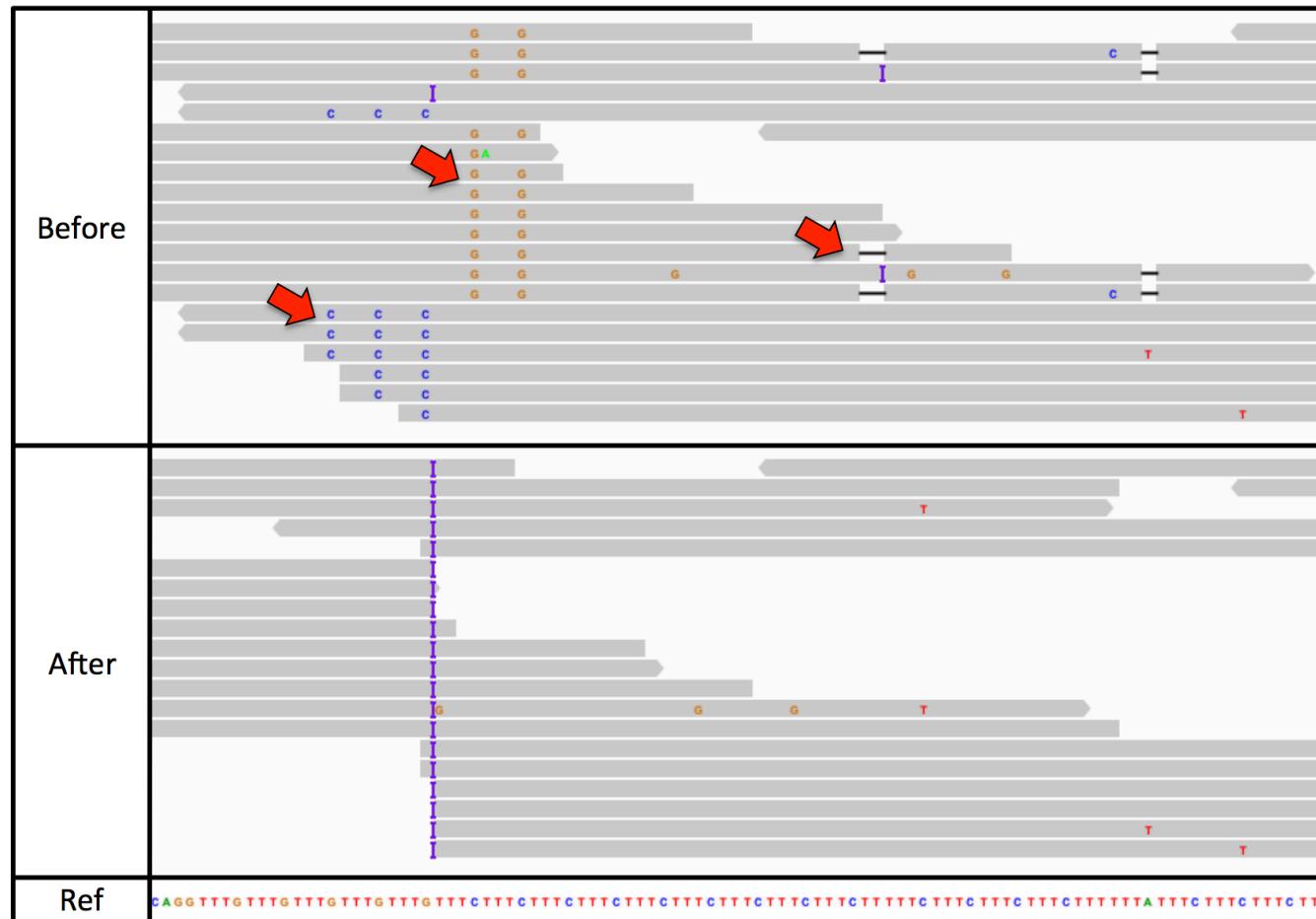
Variant Discovery

>>

Callset Refinement

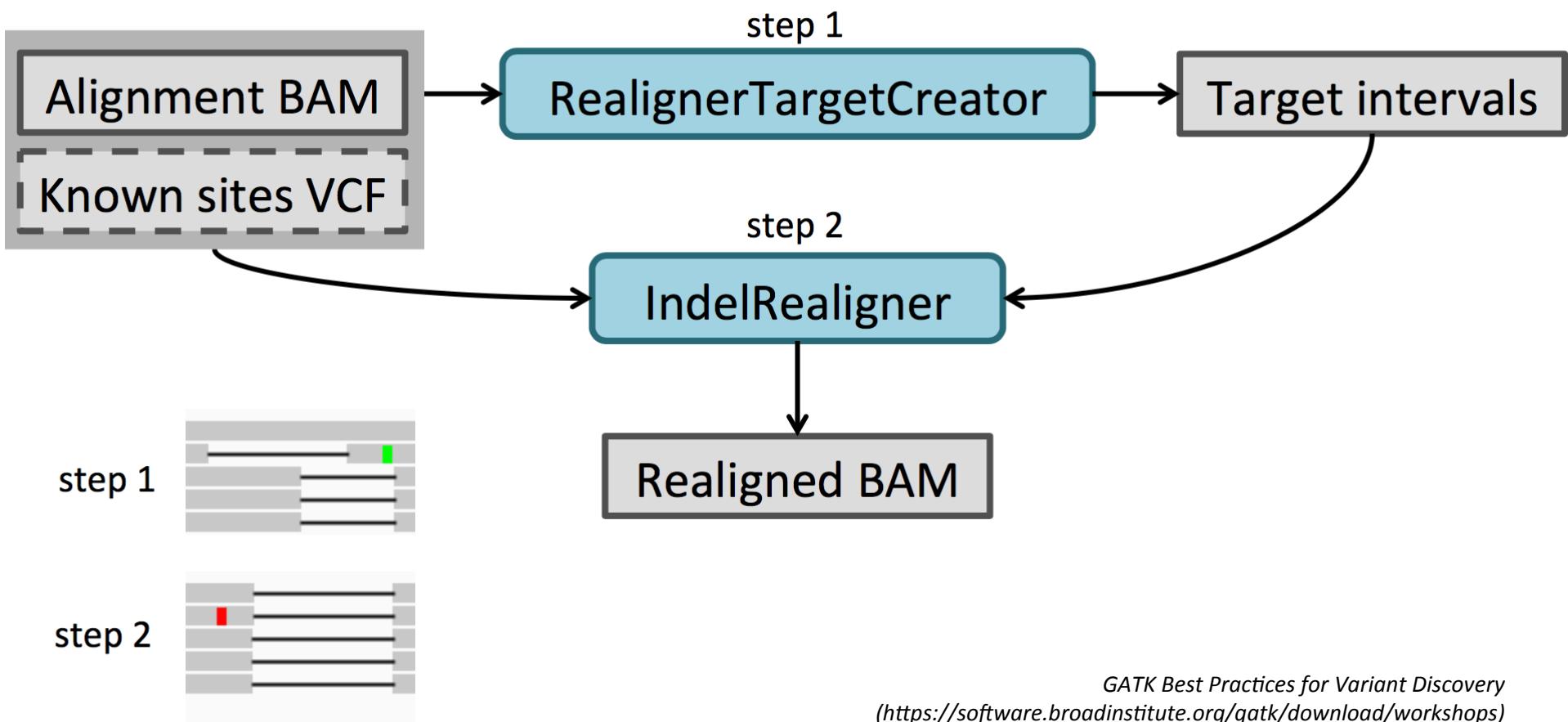


Indel realignment



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

Indel realignment workflow



GATK workflow

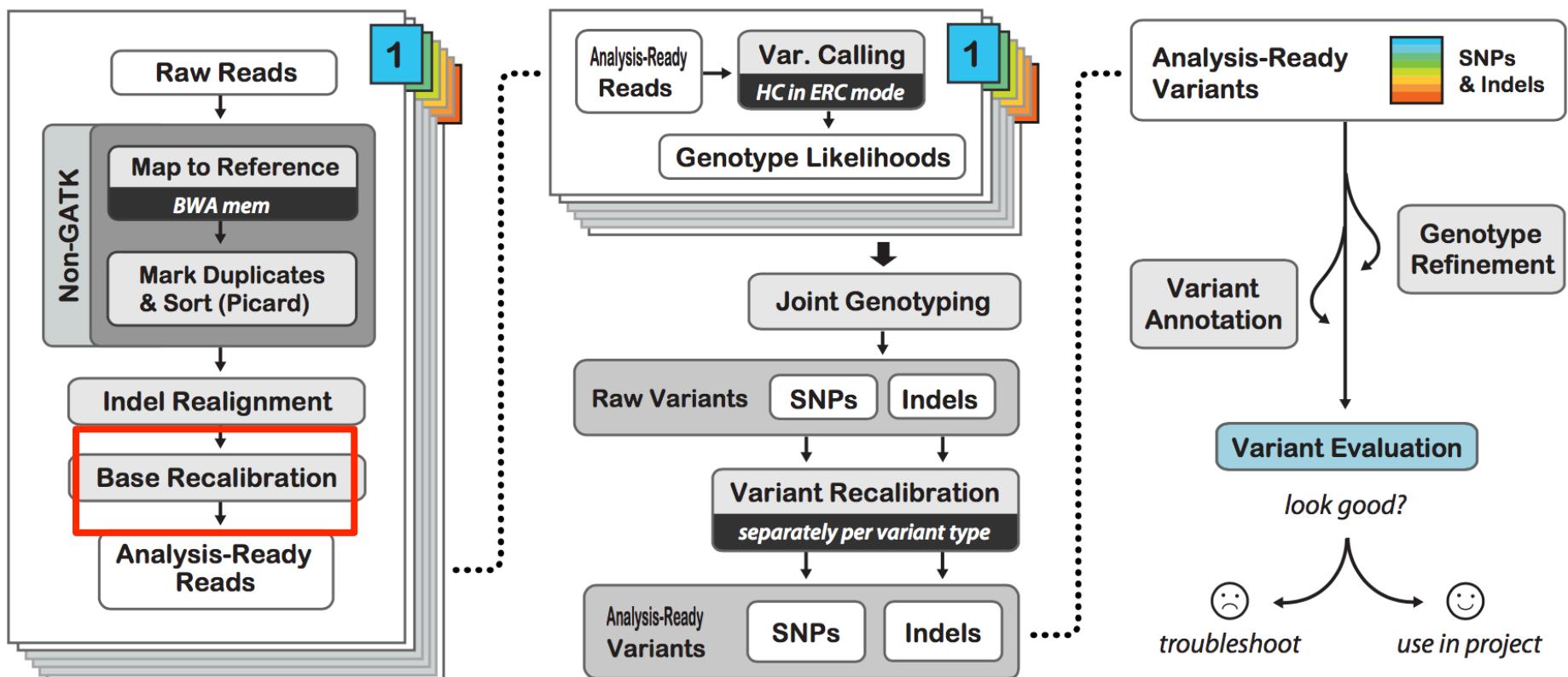
Data Pre-processing

>>

Variant Discovery

>>

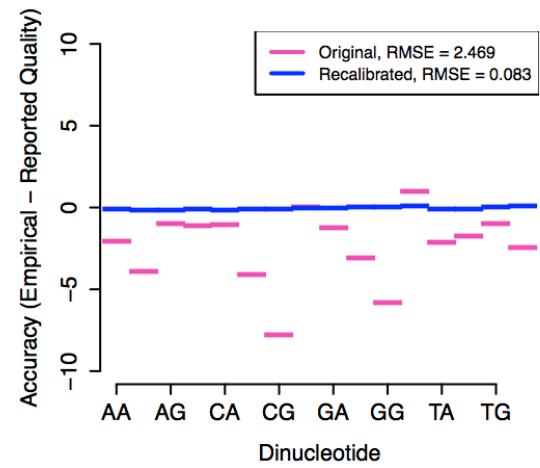
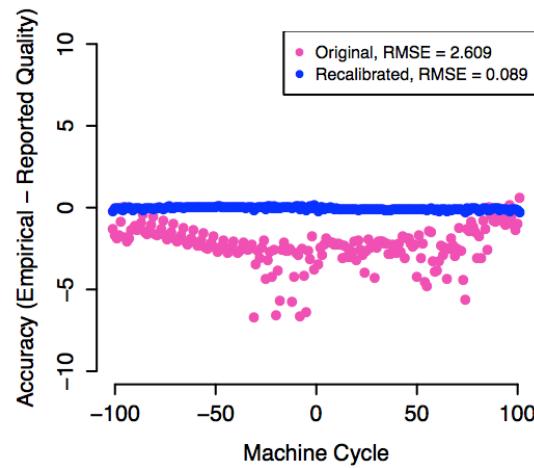
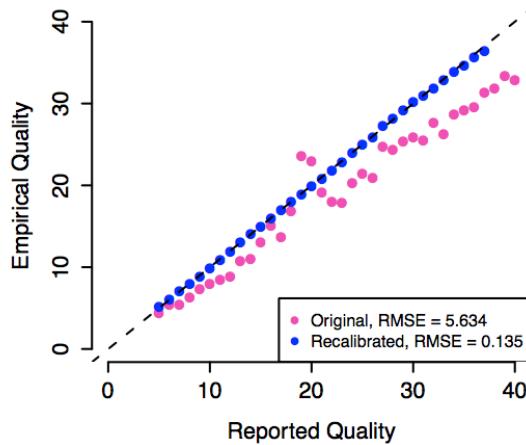
Callset Refinement





Base Quality Score Recalibration (BQSR)

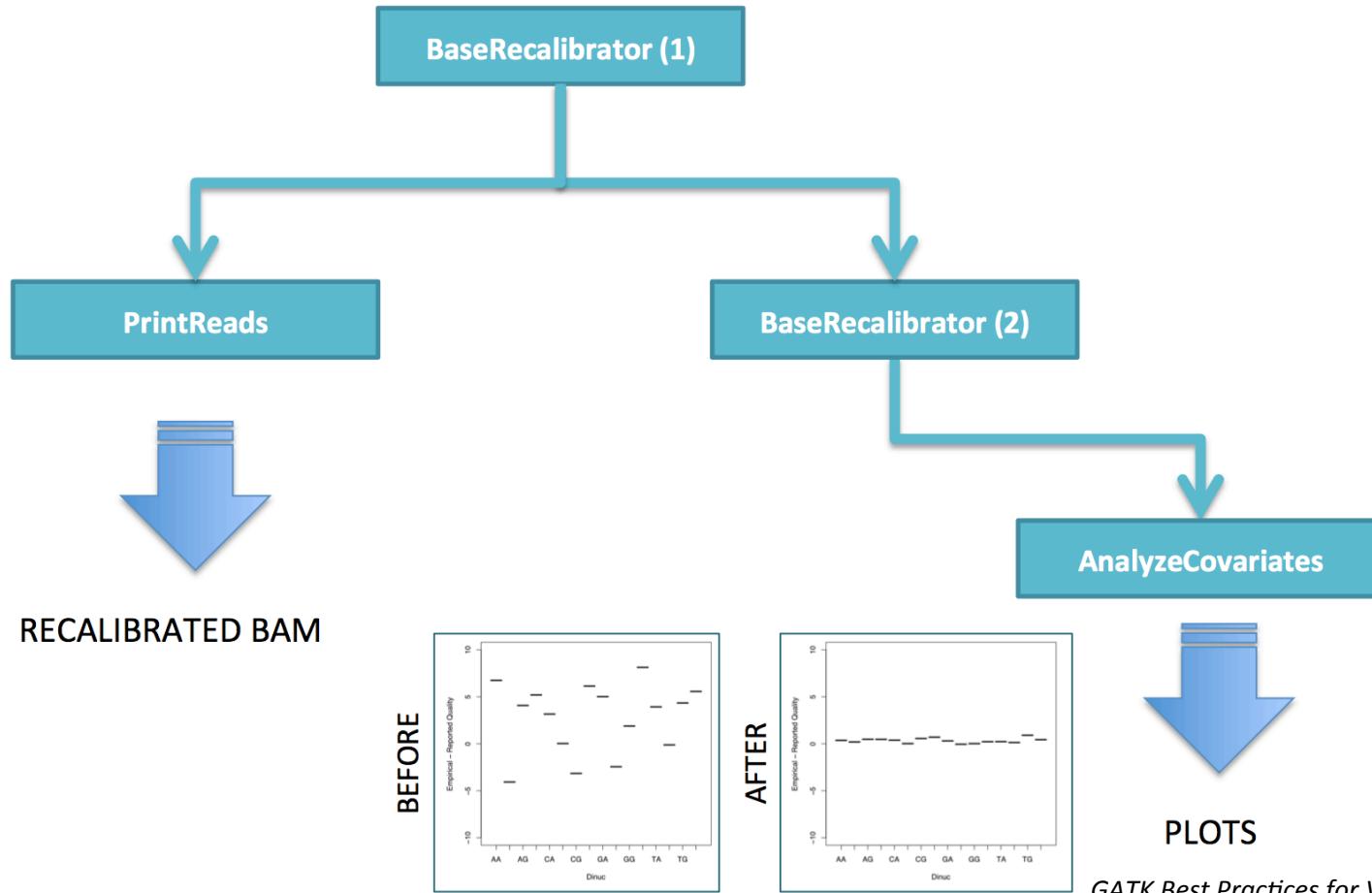
- Systematic biases: quality of base calls depend on nucleotidic context, machine cycle...
- → Machine learning algorithm to find how error varies with basecall features. Then, apply recalibration.



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

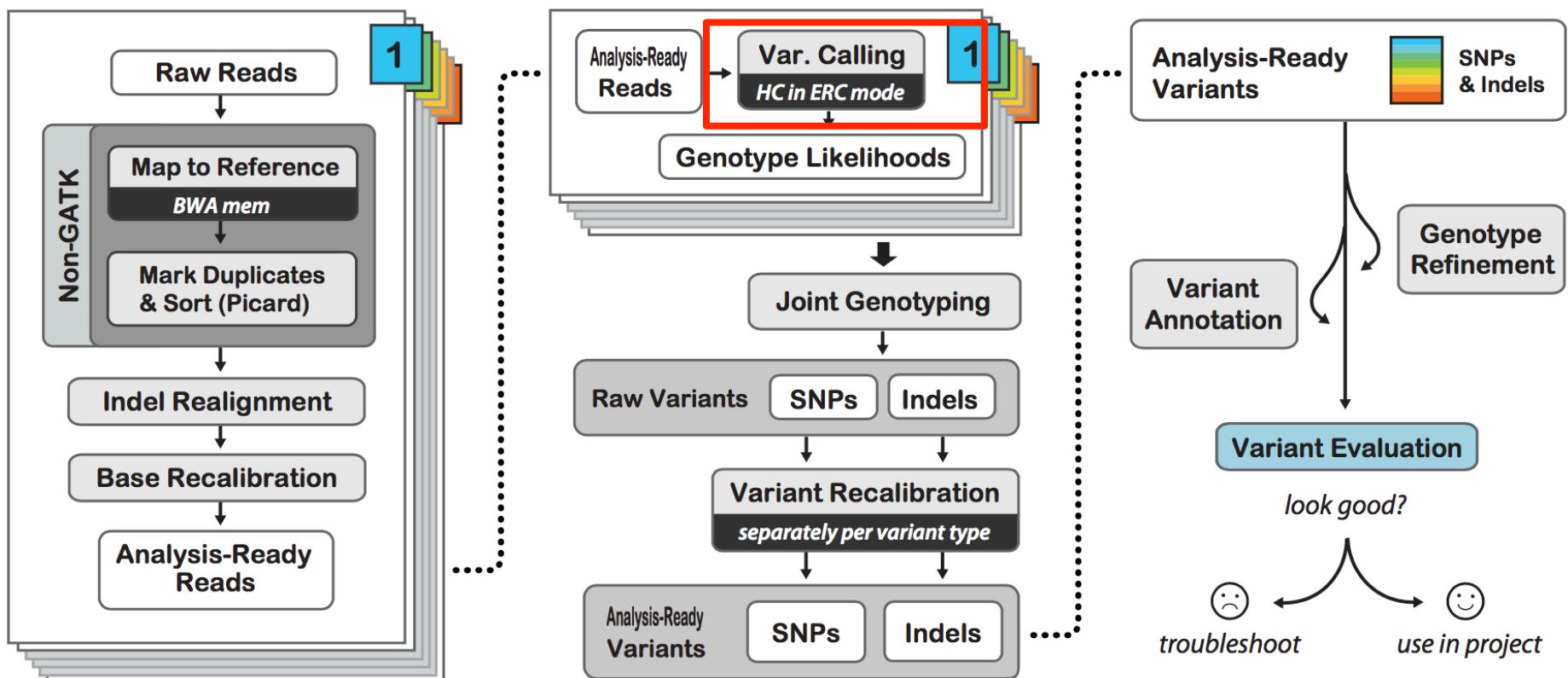
Data Pre-processing

>>

Variant Discovery

>>

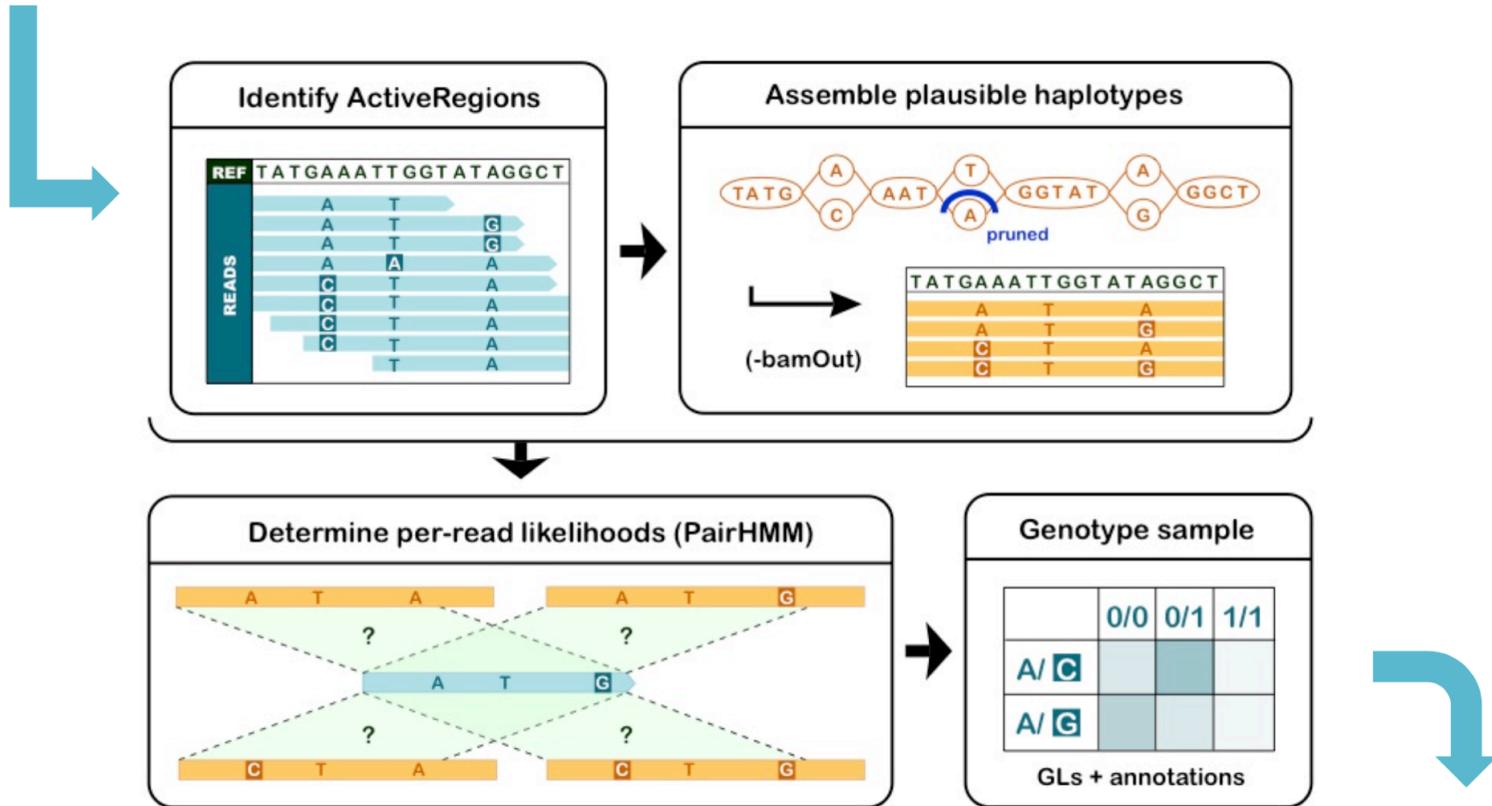
Callset Refinement





Variant-calling: Haplotype-Caller

BAM



VCF & index

GATK Best Practices for Variant Discovery
<https://software.broadinstitute.org/gatk/download/workshops>

GATK workflow

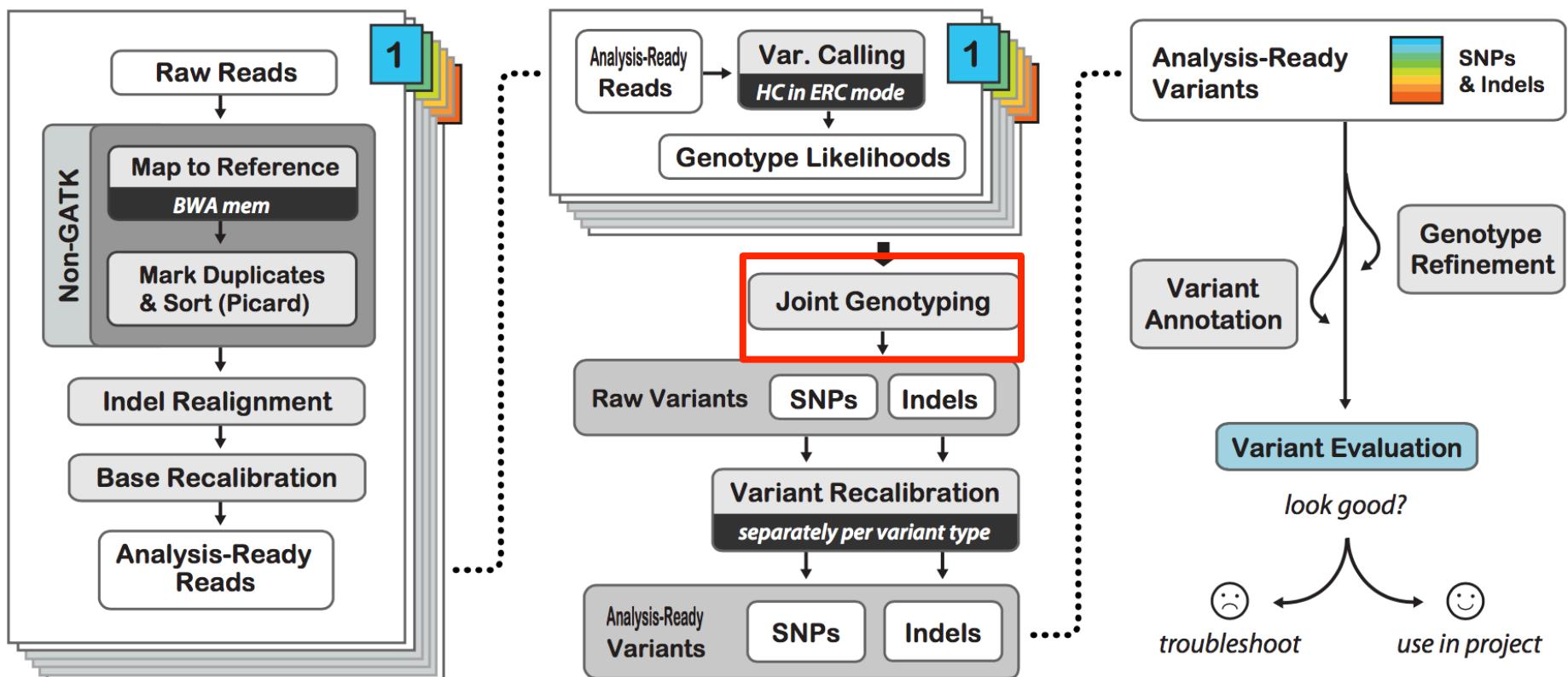
Data Pre-processing

>>

Variant Discovery

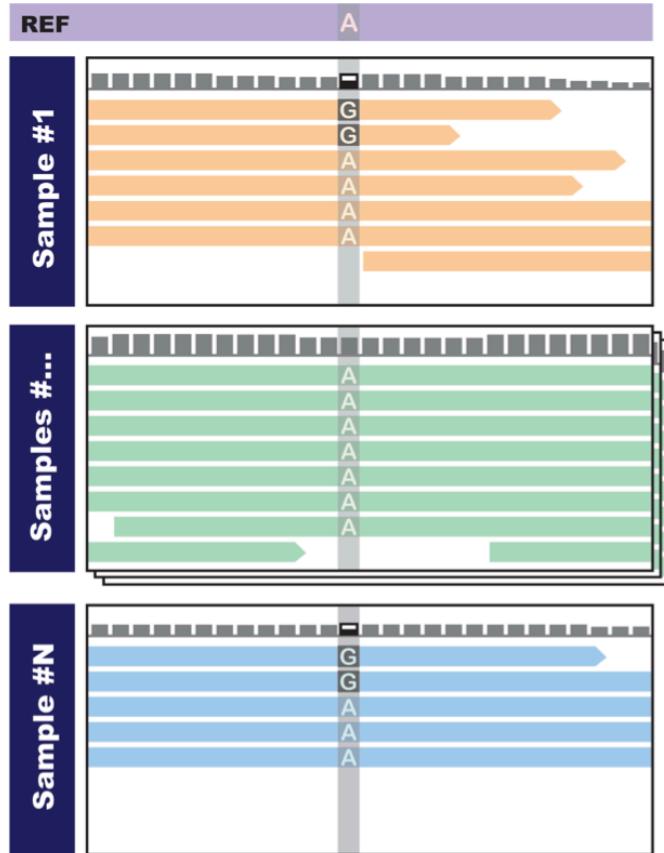
>>

Callset Refinement





Joint-genotyping



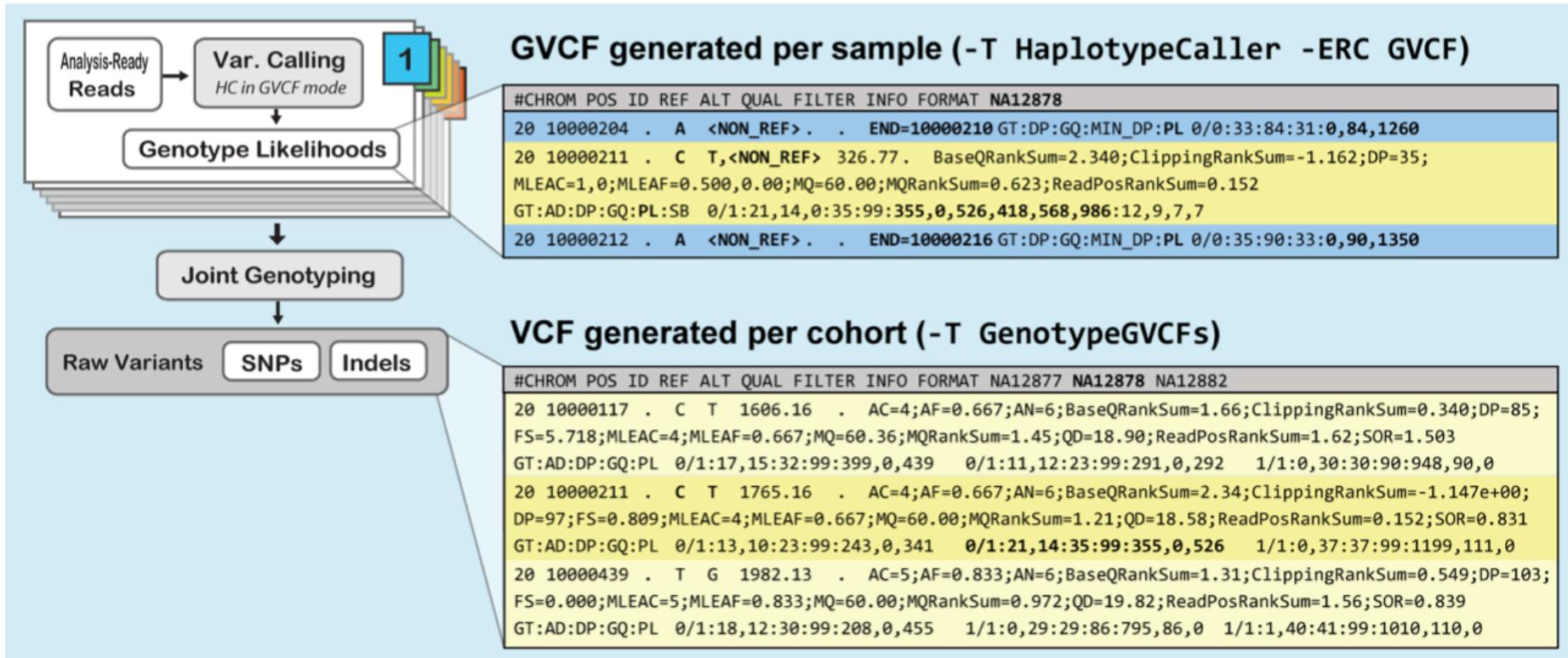
- When only 1 sample => "G" may be considered errors.
- When all samples => More confidence in the calling of A/G variant.

Joint callset → empowered analysis

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



gVCF VS regular VCF





Conclusion on the GATK workflow

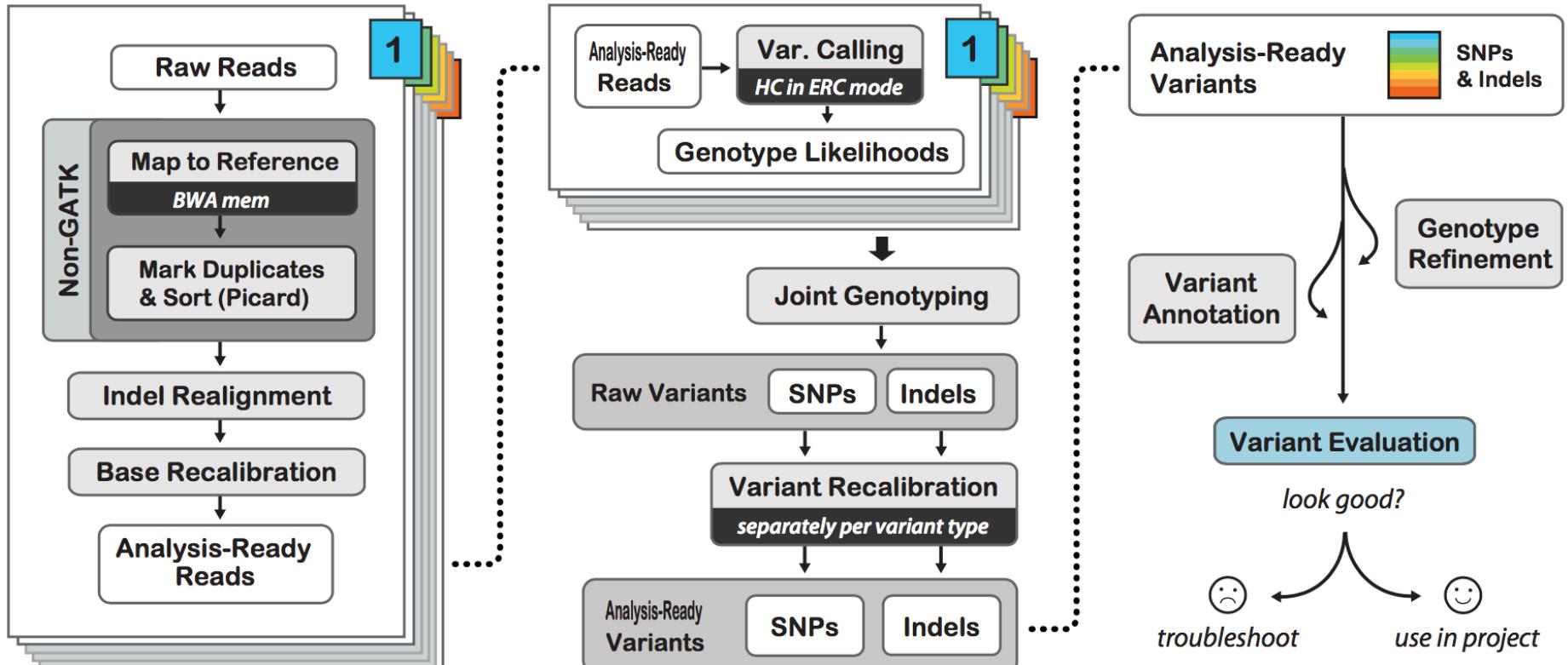
Data Pre-processing

>>

Variant Discovery

>>

Callset Refinement



Thanks for your attention