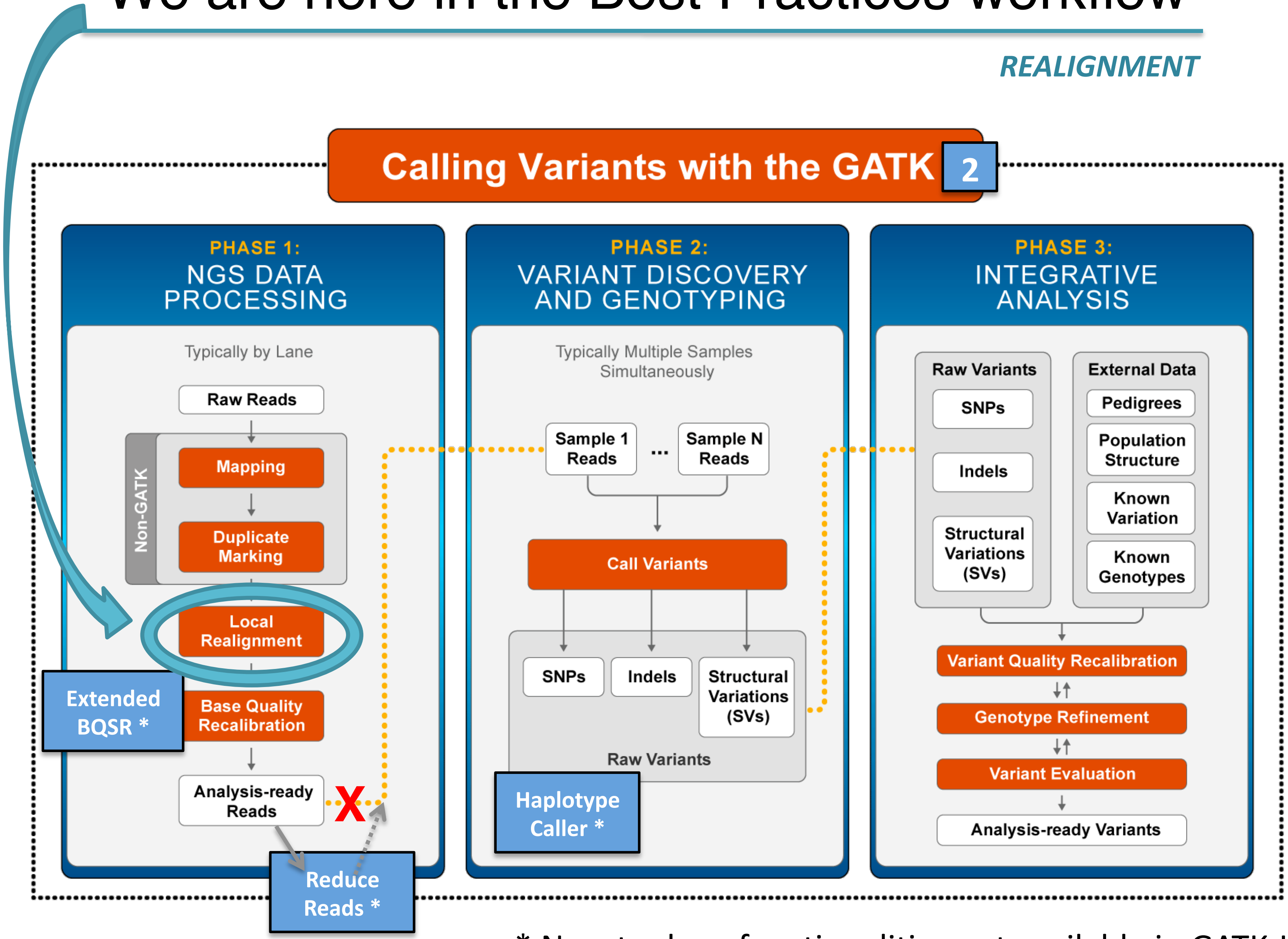
Indel-based Realignment

Improving the original alignments of the reads based on multiple sequence (re-)alignment

We are here in the Best Practices workflow

REALIGNMENT



* New tools or functionalities not available in GATK-Lite

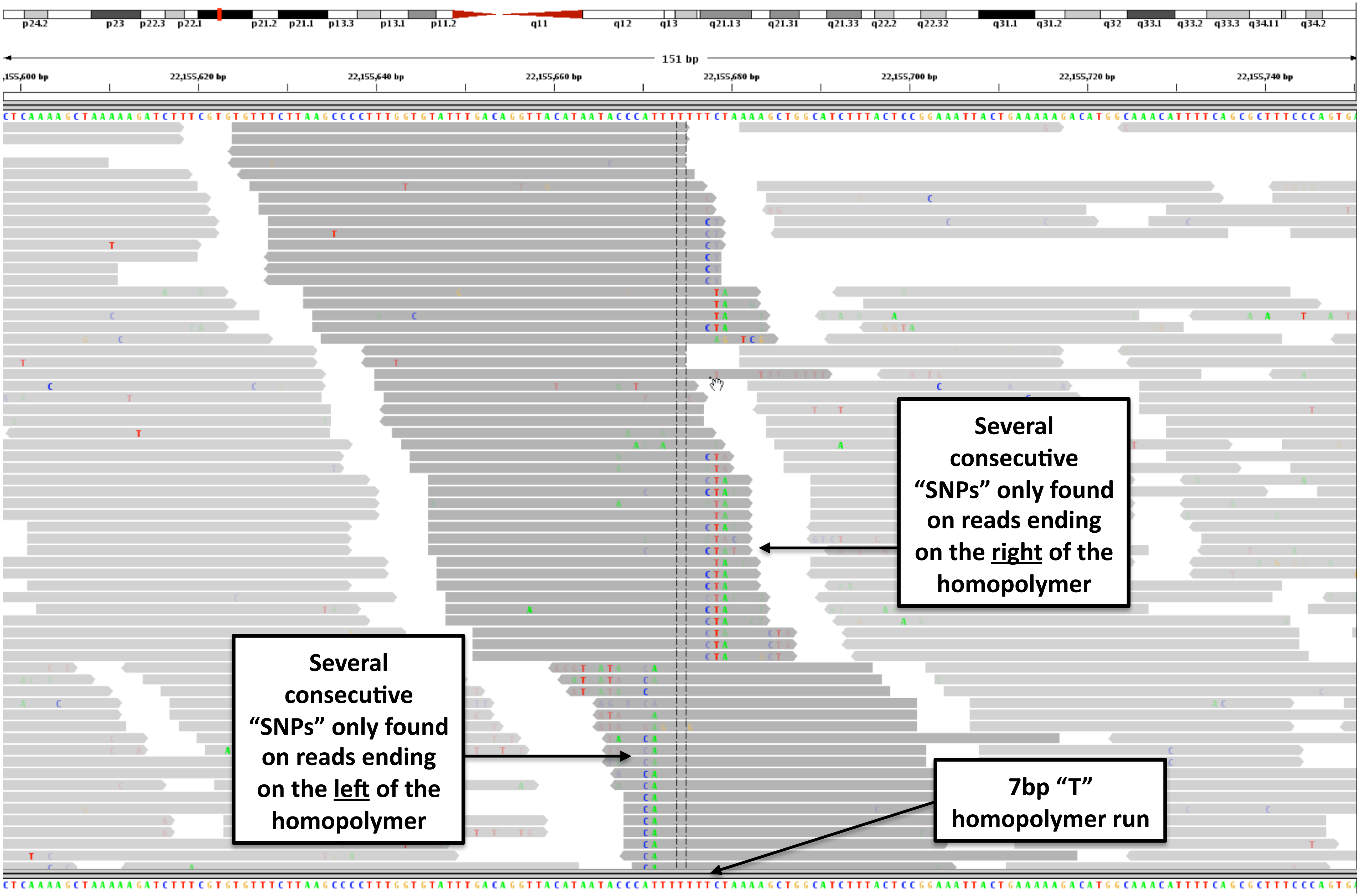
PURPOSE

Why realign around indels?

- InDels in reads (especially near the ends) can trick the mappers into mis-aligning with mismatches
- These artifactual mismatches can harm base quality recalibration and variant detection (unless a sophisticated caller like the Haplotype Caller is used)

☑ **Realignment around indels helps improve the accuracy of several of the downstream processing steps.**

An example of a strand-discordant locus

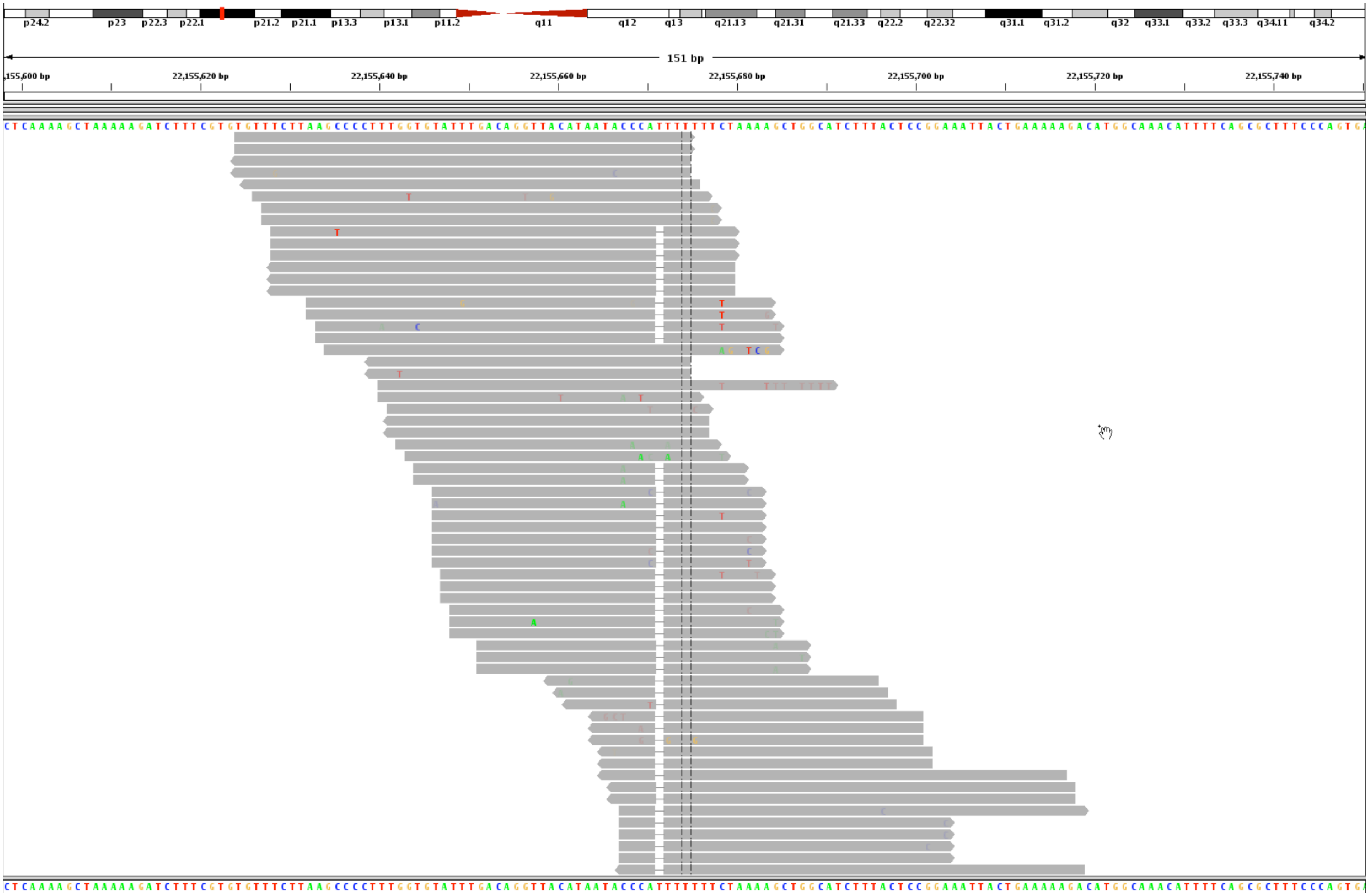

p24.2 p23 p22.3 p22.1 p21.2 p21.1 p13.3 p13.1 p11.2 q11 q12 q13 q21.13 q21.31 q21.33 q22.2 q22.32 q31.1 q31.2 q32 q33.1 q33.2 q33.3 q34.11 q34.2
151 bp
,155,600 bp
22,155,620 bp
22,155,640 bp
22,155,660 bp
22,155,680 bp
22,155,700 bp
22,155,720 bp
22,155,740 bp
CTCAAAAGCTAAAAAGATCTTTCGTGTGTTTCTTAAGCCCCTTTGGTGTATTTGACAGGTTACATAATACCCATTTTTTTCTAAAAGCTGGCATCTTTACTCCGGAAATTACTGAAAAAGACATGGCAAACATTTTCAGCGCTTTCCCAGTGA

Several consecutive "SNPs" only found on reads ending on the right of the homopolymer

Several consecutive "SNPs" only found on reads ending on the left of the homopolymer

7bp "T" homopolymer run

Local realignment uncovers the hidden indel in these reads and eliminates all the potential FP SNPs



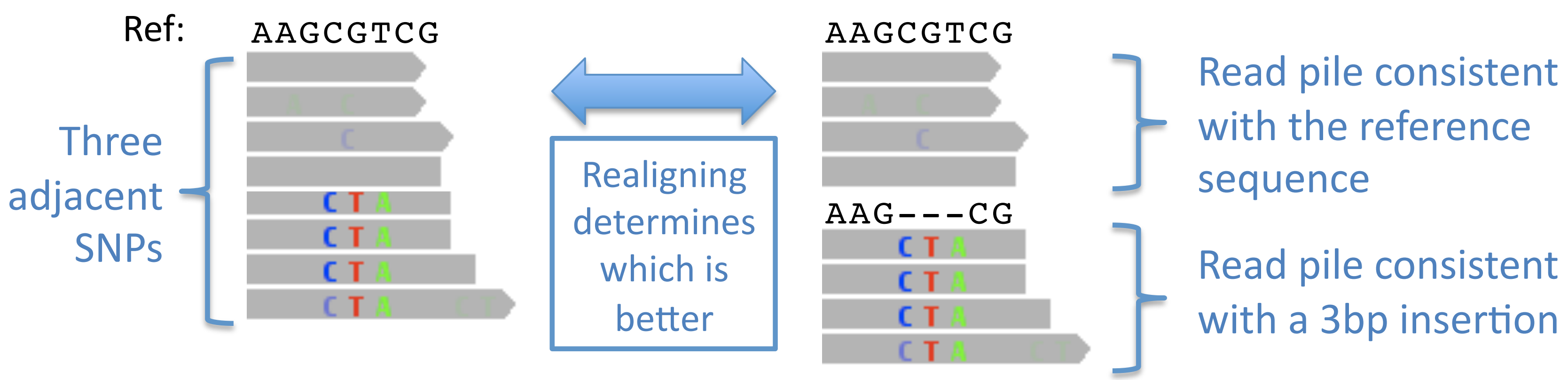
PRINCIPLES

Three types of realignment targets

- Known sites (e.g. dbSNP, 1000 Genomes)
- Indels seen in original alignments (in CIGARs)
- Sites where evidence suggests a hidden indel

Local realignment identifies most parsimonious alignment along all reads at a problematic locus

1. Find the best alternate consensus sequence that, together with the reference, best fits the reads in a pile (maximum of 1 indel)



Three adjacent SNPs

Realigning determines which is better

Read pile consistent with the reference sequence

Read pile consistent with a 3bp insertion

2. The score for an alternate consensus is the total sum of the quality scores of mismatching bases

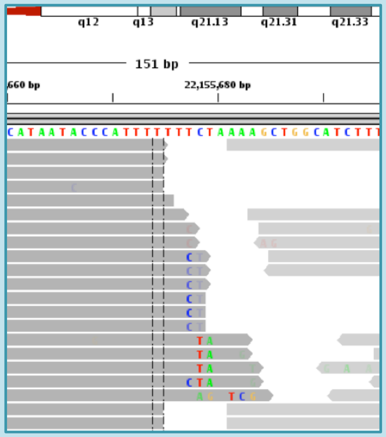
3. If the score of the best alternate consensus is sufficiently better than the original alignments (using a LOD score), then we accept the proposed realignment of the reads

PROTOCOL

Indel Realignment steps/tools

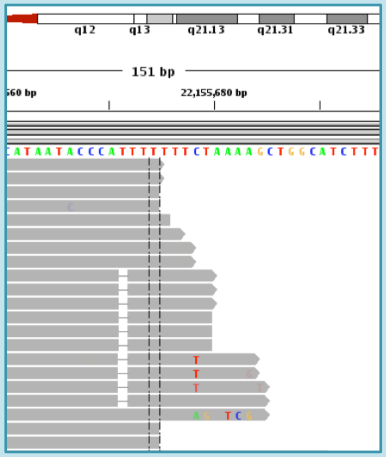
- Identify what regions need to be realigned

  → **RealignerTargetCreator**

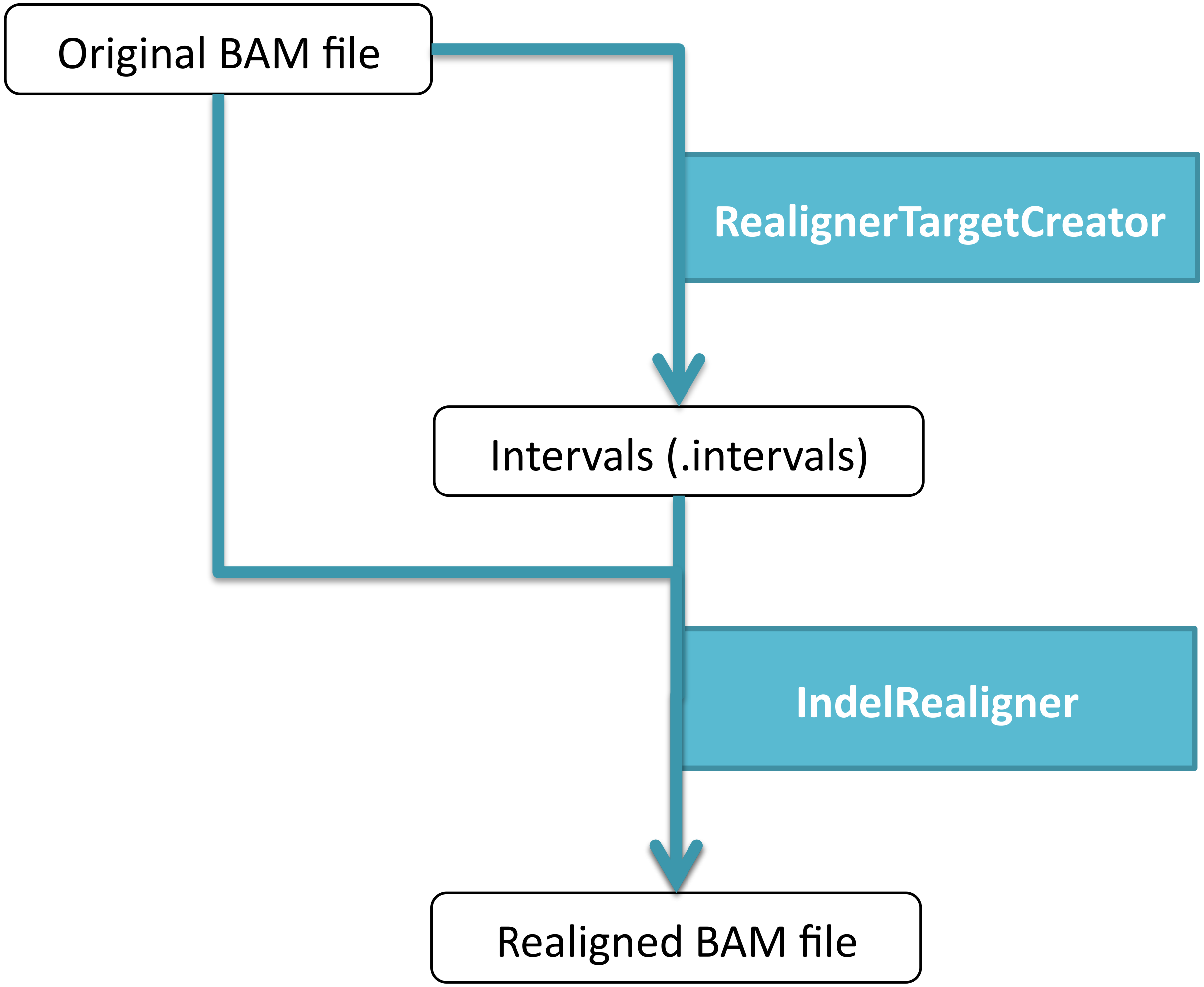


- Perform the actual realignment

  → **IndelRealigner**



Indel Realignment workflow



Original BAM file

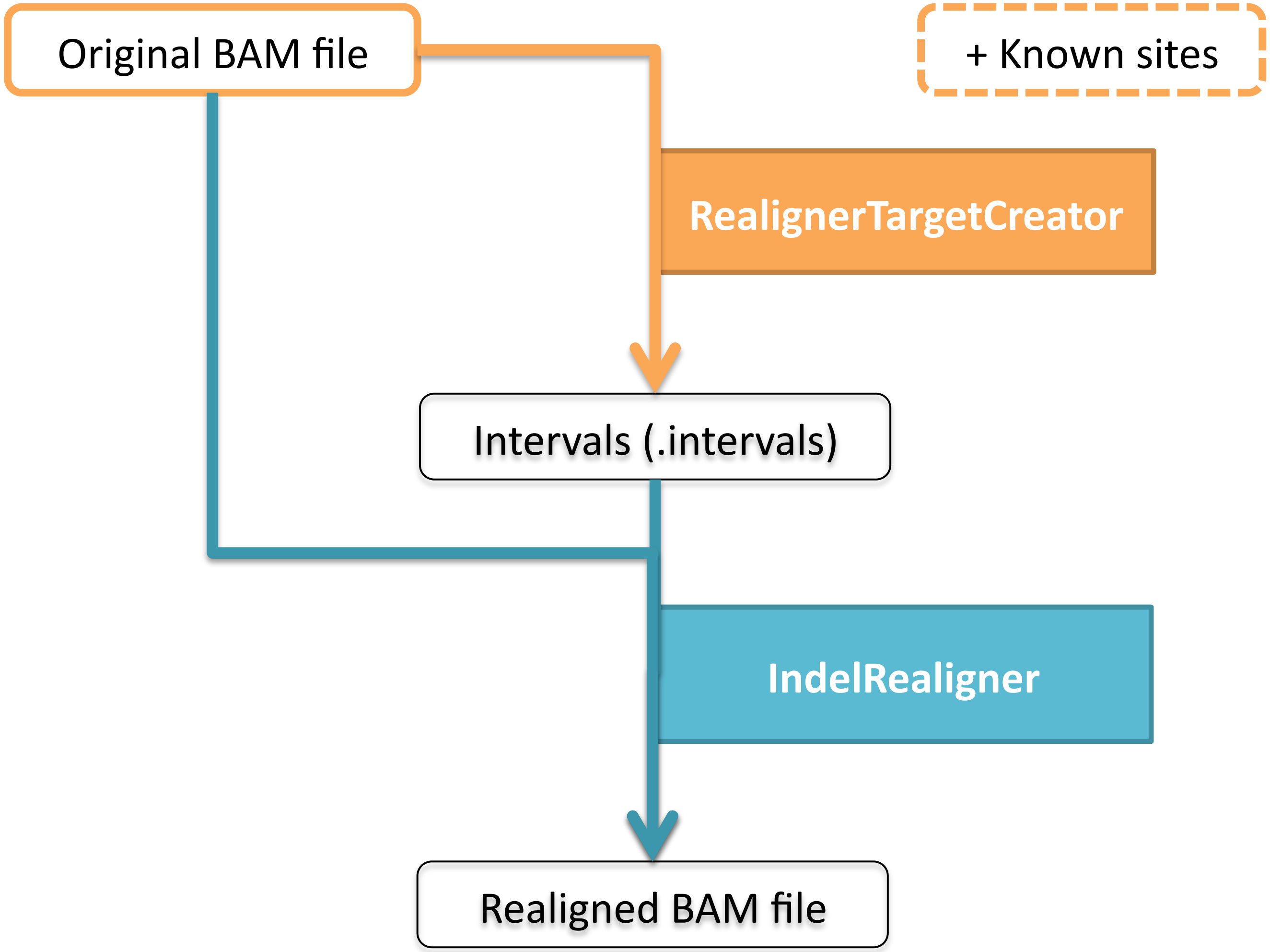
RealignerTargetCreator

Intervals (.intervals)

IndelRealigner

Realigned BAM file

Indel Realignment workflow



Original BAM file

+ Known sites

RealignerTargetCreator

Intervals (.intervals)

IndelRealigner

Realigned BAM file

TOOL TIPS

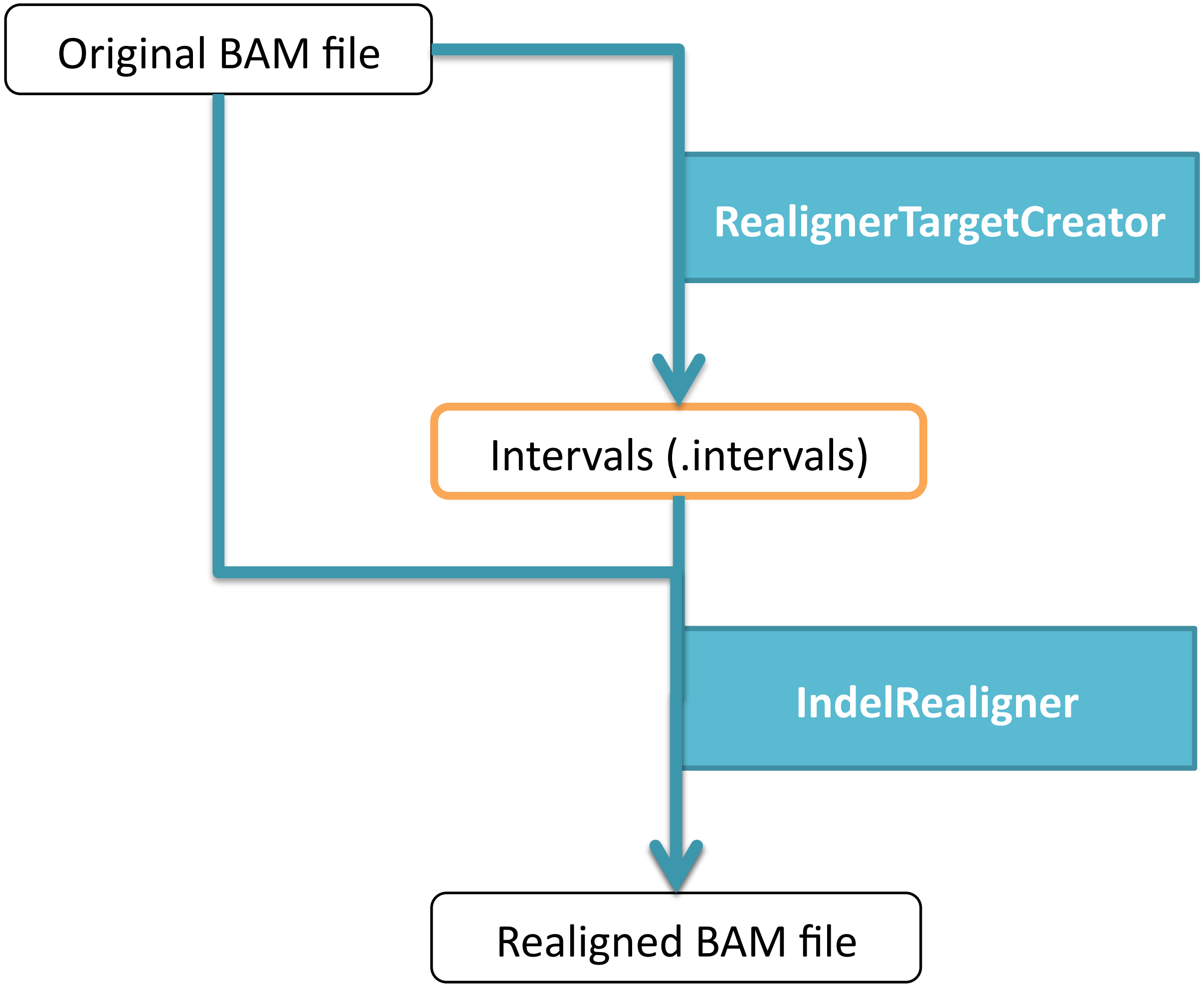
RealignerTargetCreator

- Pre-processing step to find intervals that may need realignment

```
java –jar GenomeAnalysisTK.jar –T RealignerTargetCreator \
        –R human.fasta \
        –I original.bam \
        –known indels.vcf \
        –o realigner.intervals
```

- Input BAM file not necessary if processing only at known indels
- Using a list of known indels will both speed up processing and improve accuracy, but is not required

Indel Realignment workflow



Original BAM file

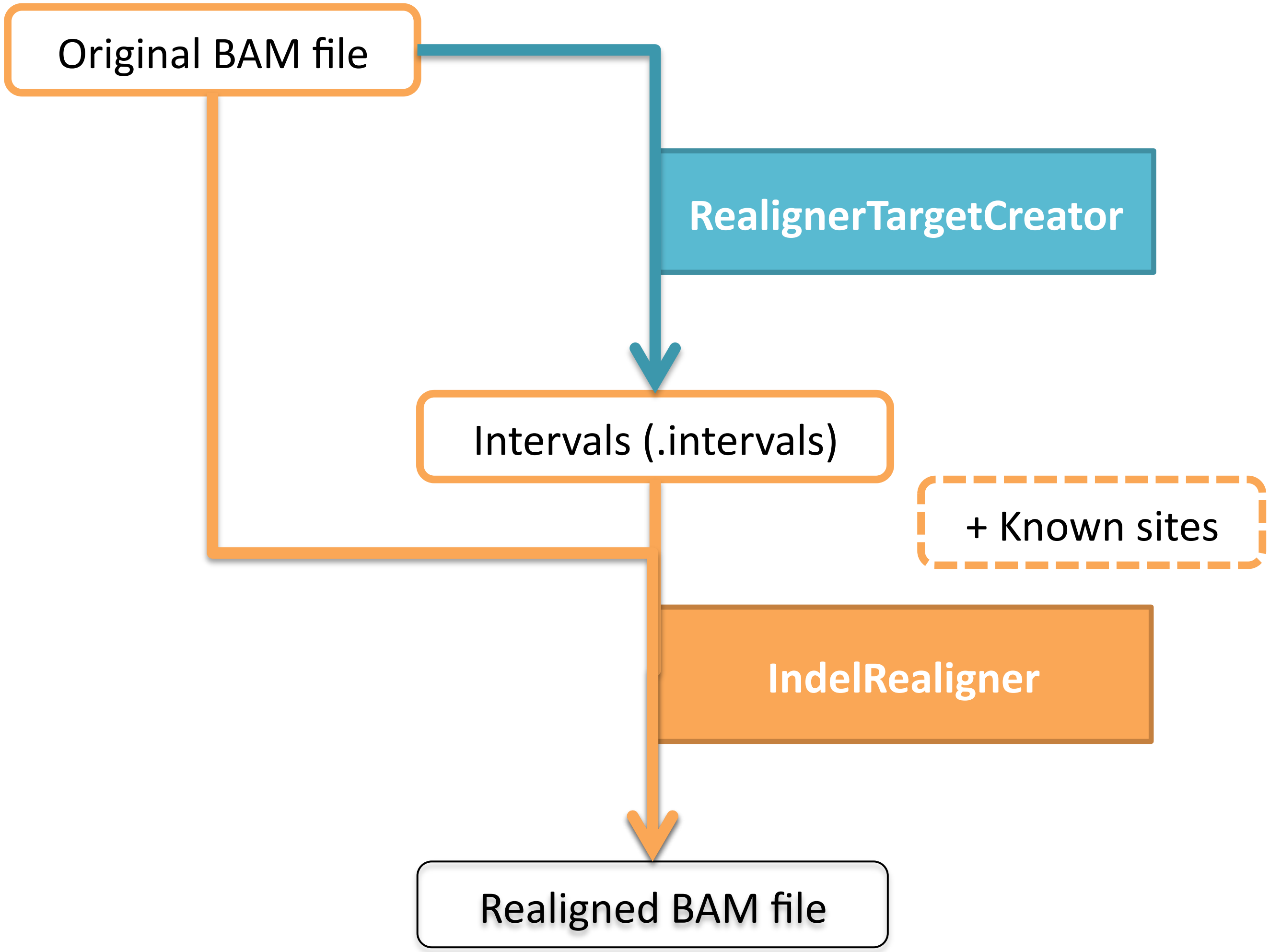
RealignerTargetCreator

Intervals (.intervals)

IndelRealigner

Realigned BAM file

Indel Realignment workflow



TOOL TIPS

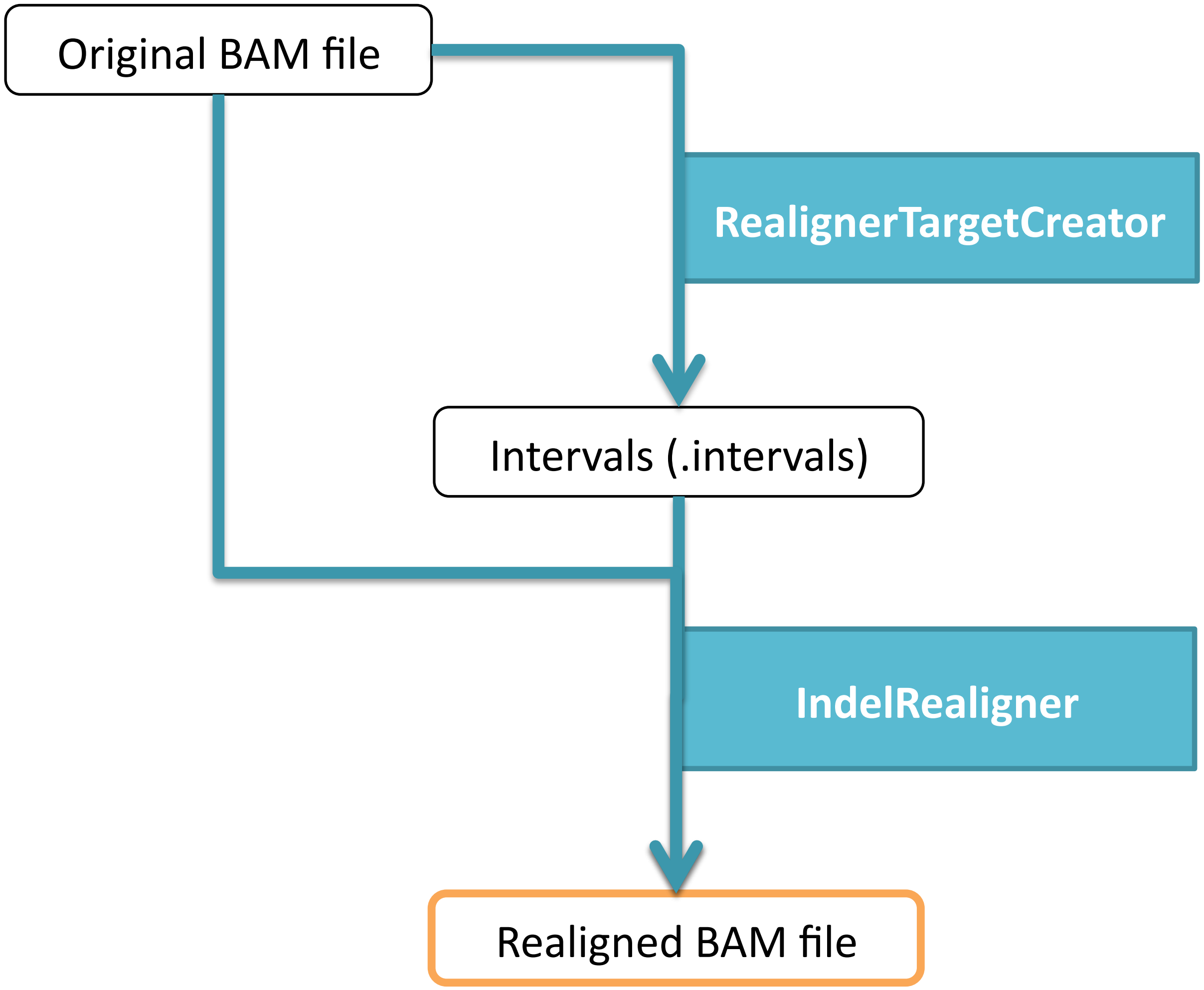
IndelRealigner

- Attempts realignment at RTC target intervals

```
java –jar GenomeAnalysisTK.jar –T IndelRealigner \
        –R human.fasta \
        –I original.bam \
        –known indels.vcf \
        –targetIntervals realigner.intervals \
        –o realigned.bam
```

- Must use same input file(s) used in RealignerTargetCreator step
- Processing options
  - Only at known indels: much faster, accurate for ~90-95% of indels
  - At indels seen in the original BAM alignments: the recommended mode
  - Using full Smith-Waterman realignment: most accurate, but heavy computational cost and not really necessary with the new techs

Indel Realignment workflow



Original BAM file

RealignerTargetCreator

Intervals (.intervals)

IndelRealigner

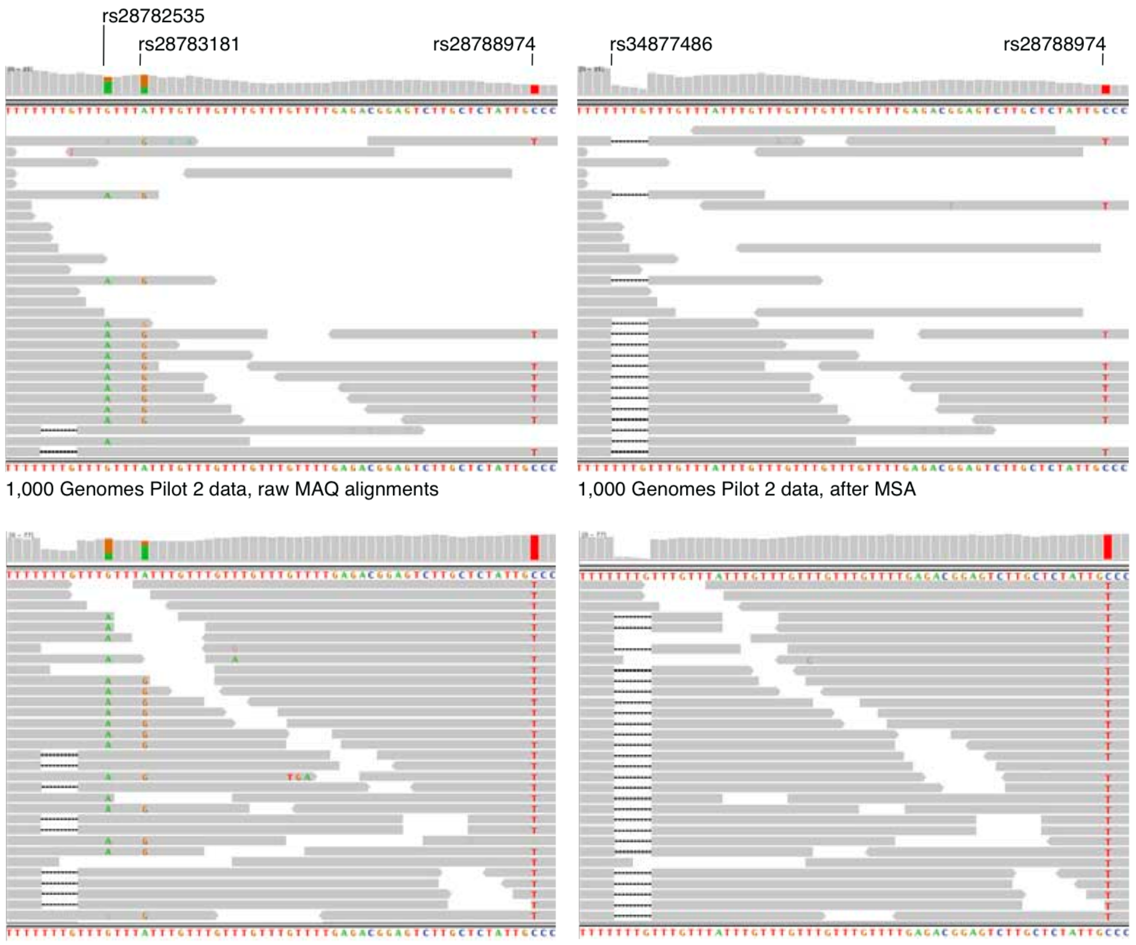
Realigned BAM file

RESULTS

This is what a realigned BAM looks like

NA12878, chr1:1,510,530-1,510,589

Before



1,000 Genomes Pilot 2 data, raw MAQ alignments

1,000 Genomes Pilot 2 data, after MSA

HiSeq data, raw BWA alignments

HiSeq data, after MSA

After

DePristo, M., Banks, E., Poplin, R. et. al, A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Gen.

Did the realignment work properly?

- Indel Realigner changes the CIGAR string of realigned reads but maintains the original CIGAR (with OC tag)
  - So it’s very easy to check that realignment was performed and/or how many reads were adjusted
- BUT no formal measure to assess the accuracy or completeness of the realignment process

Is realignment still necessary with latest software?

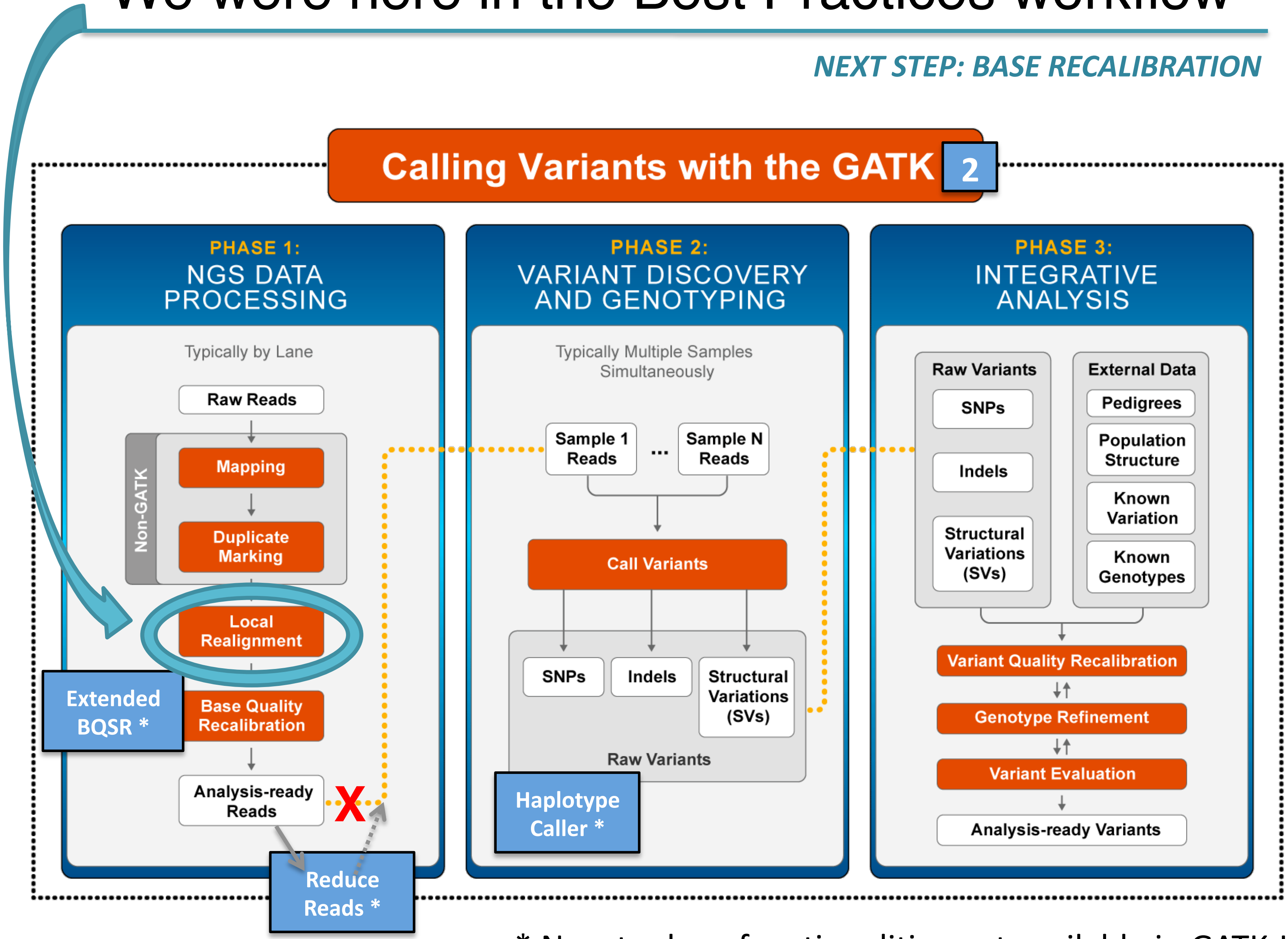
- Latest tools being implemented for discovering mutations all include some sort of assembly step (for which upstream realignment is not really helpful).

- BUT big improvement for Base Quality Score Recalibration when run on realigned BAM files (artifactual SNPs are replaced with real indels).

- Also still useful for legacy tools, e.g. full realignment should be performed if using the GATK's Unified Genotyper.

We were here in the Best Practices workflow

NEXT STEP: BASE RECALIBRATION



* New tools or functionalities not available in GATK-Lite

Further reading

http://www.broadinstitute.org/gatk/guide/topic?name=intro

http://www.broadinstitute.org/gatk/guide/topic?name=best-practices

http://www.broadinstitute.org/gatk/guide/article?id=38

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_indels_IndelRealigner.html

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_indels_RealignerTargetCreator.html