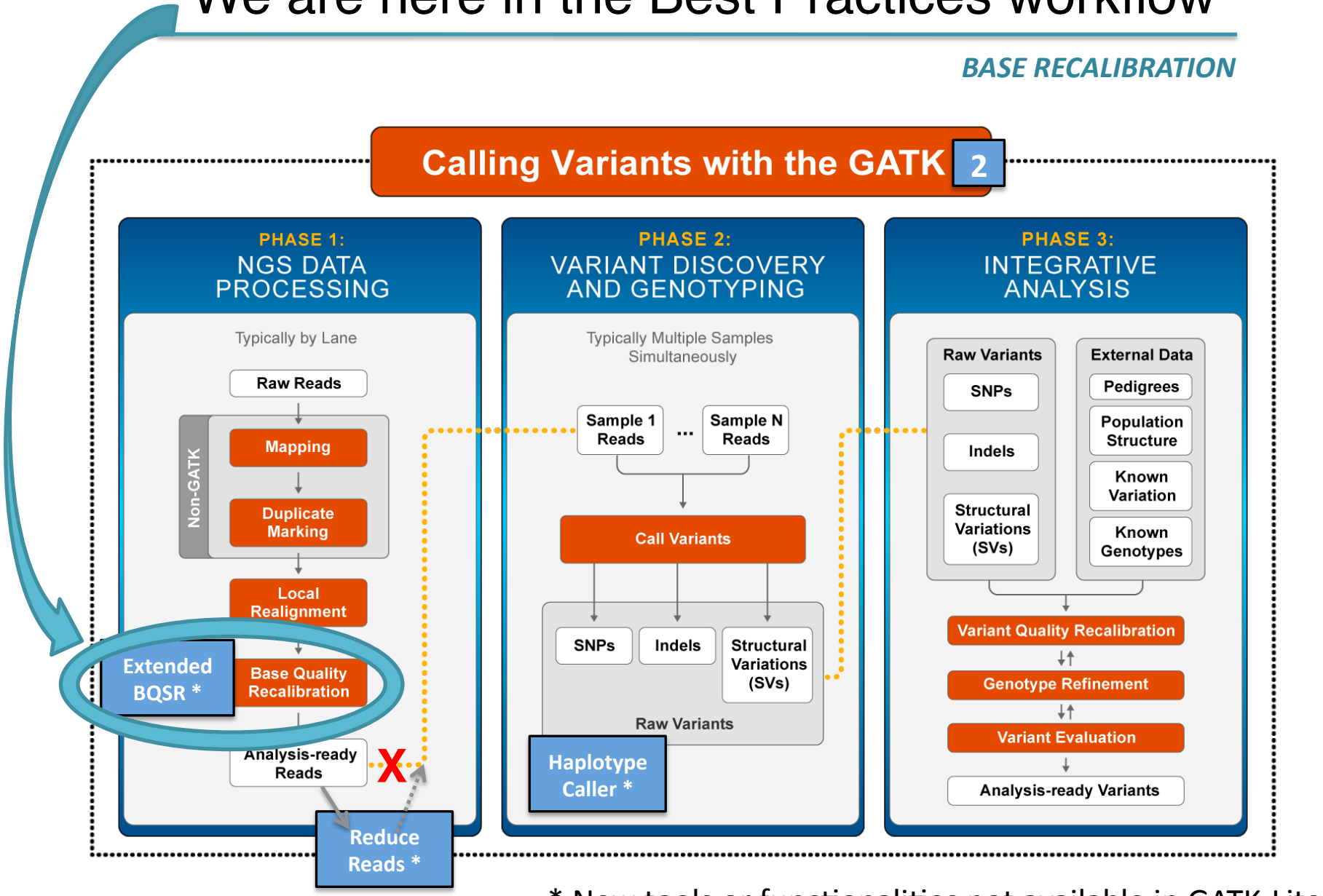
# Base Quality Score Recalibration

Assigning accurate confidence scores
to each sequenced base

# We are here in the Best Practices workflow

*BASE RECALIBRATION*




* New tools or functionalities not available in GATK-Lite

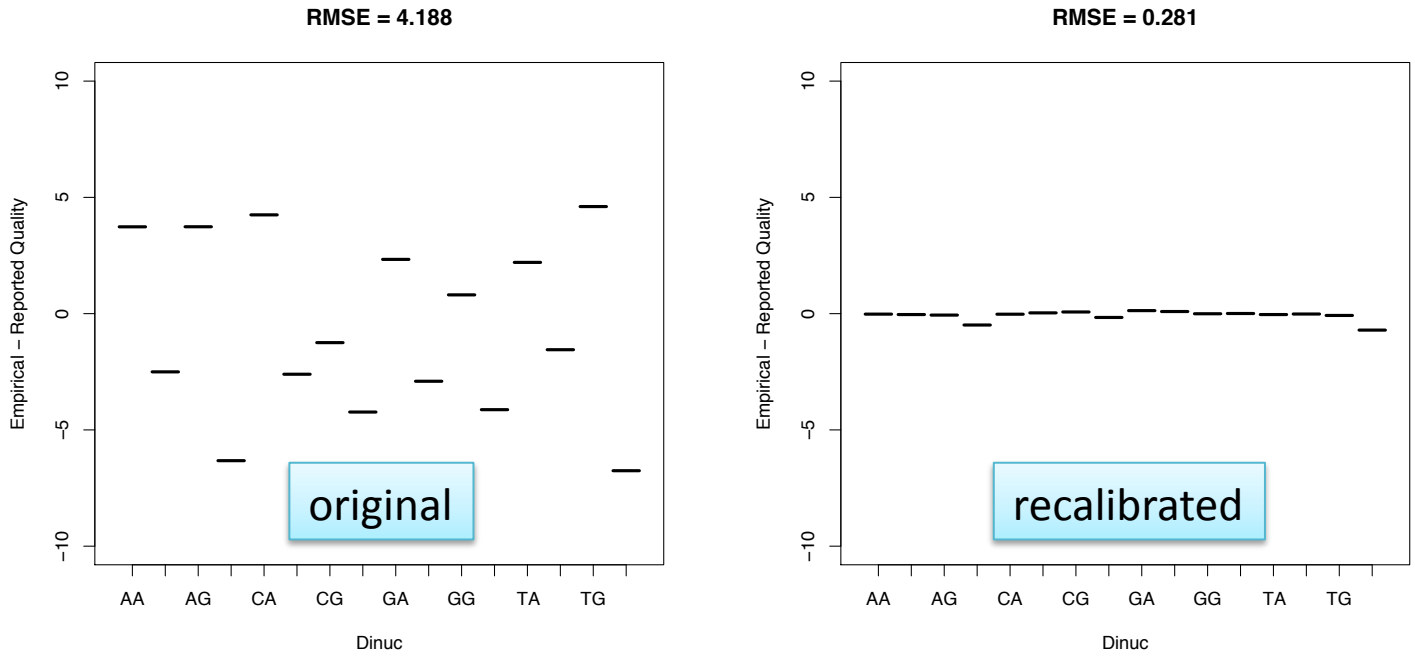
# PURPOSE

# Why recalibrate base qualities?

# The quality scores issued by sequencers are inaccurate and biased

- Quality scores are critical for all downstream analysis
- Systematic biases are a major contributor to bad calls

Example: Bias in the qualities reported depending on nucleotide context



# PRINCIPLES

Recalibration is empowered by looking at the accuracy of the entire lane's worth of data in aggregate

- Analyze covariation among several features of a base, e.g.:
  - Reported quality score
  - Position within the read (machine cycle)
  - Preceding and current nucleotide (sequencing chemistry effect)

- Apply covariates through a piecewise tabular correction to recalibrate the quality scores of all reads in a BAM file.

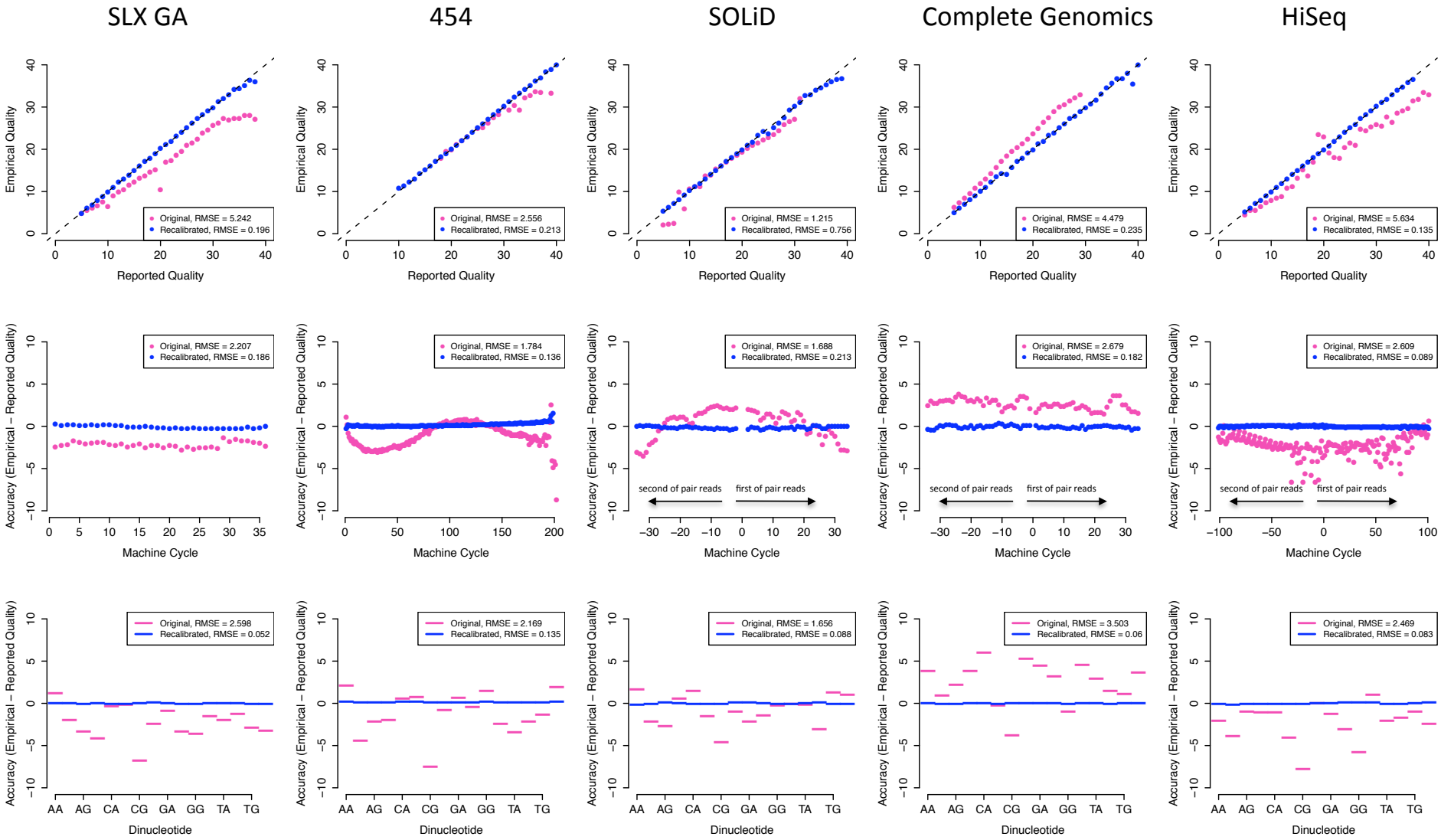
# How the covariates are analyzed

- Keep track of the number of observations and the number of times it was an error as a function of various error covariates.
  - Typically stratify the data by lane, original quality score, machine cycle, and sequencing context
  - Databases of known variants are used to discount most of the real genetic variation present in the sample
    - All other differences from the reference are assumed to be sequencing errors
    - Having done Indel Realignment first reduces noise from misalignments

$$\frac{\#\text{ of reference mismatches} + 1}{\#\text{ of observed bases} + 2} \rightarrow \text{PHRED-scaled quality score}$$

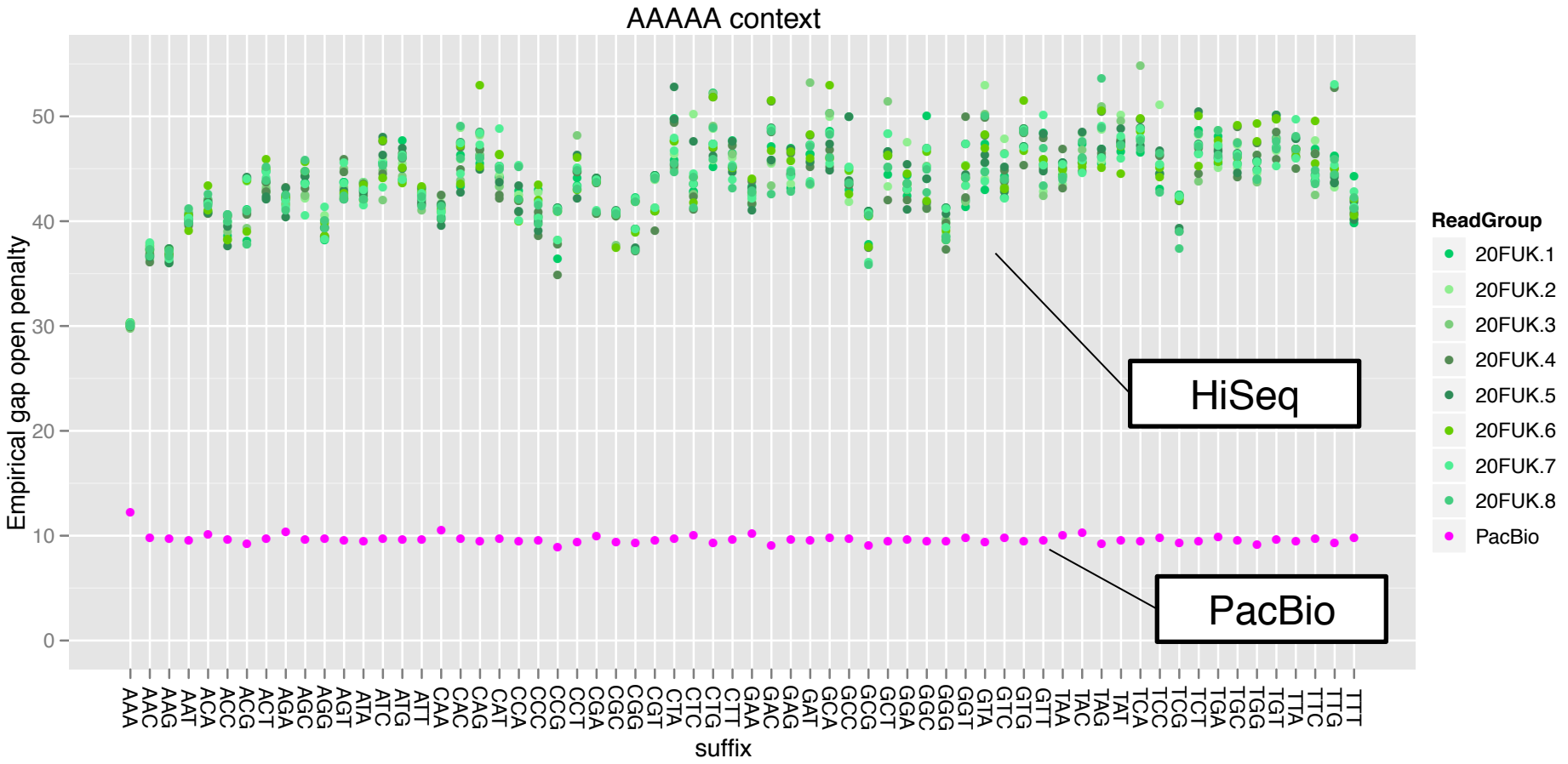
Highlighted as one of the major methodological advances of the 1000 Genomes Pilot Project!

# Base Quality Score Recalibration provides a calibrated error model from which to make mutation calls



New in 2

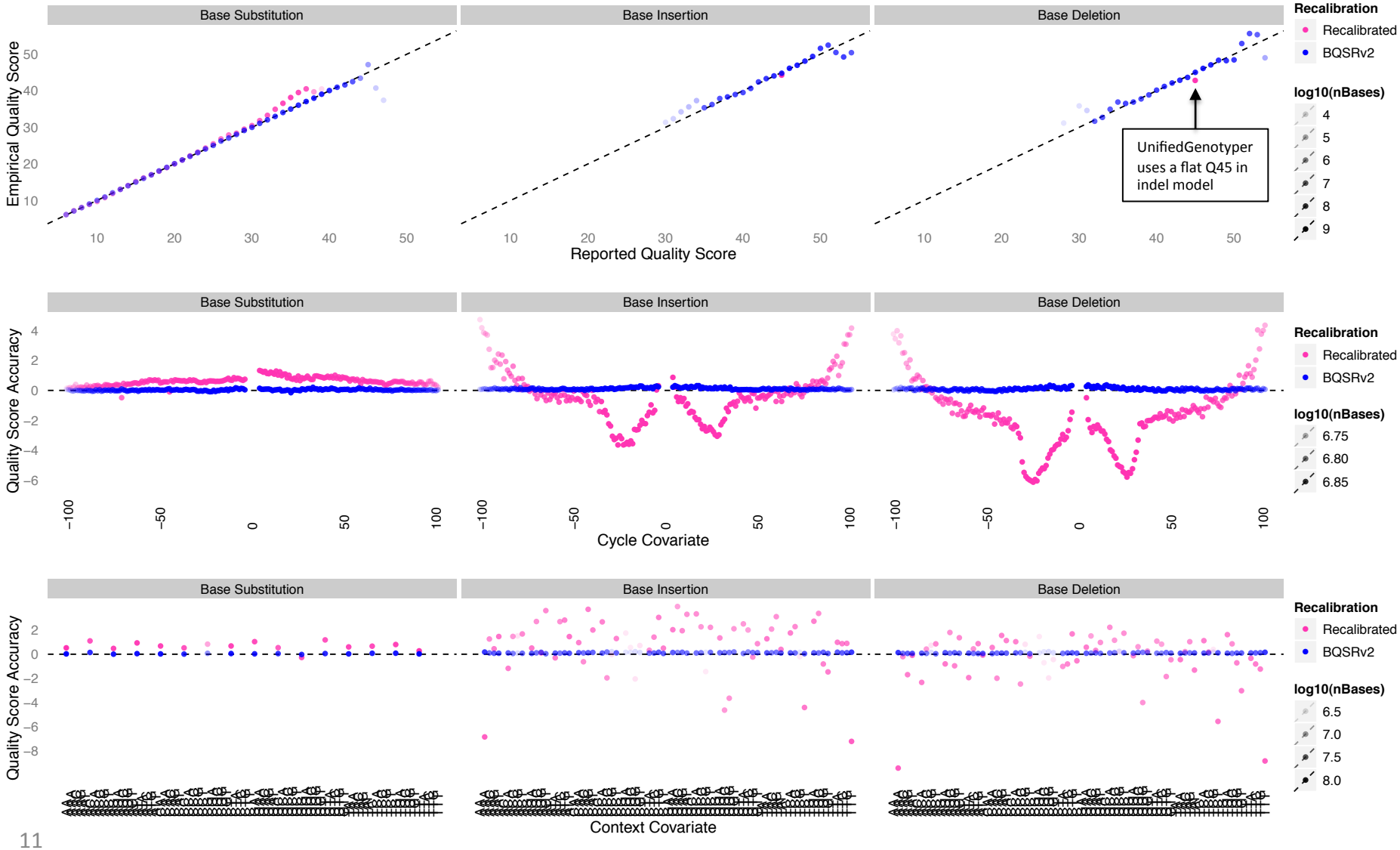
# Per-base indel error rate also varies by lane, sequence context and sequencing technology




Per-base indel error estimates are required for accurate indel calling, particularly on new technologies with indel-rich error model such as Pacific Biosciences.

New in 2

# Empirical estimates of base insertion and base deletion error rates unify SNP and indel error models

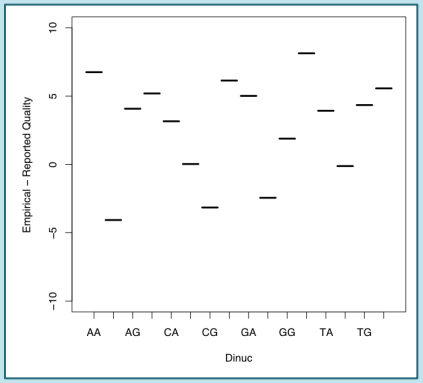


# PROTOCOL

# Base Recalibration steps/tools

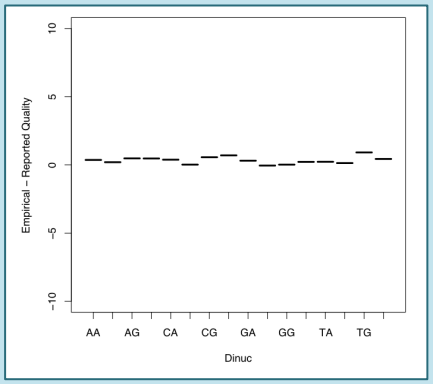
- Model the error modes and recalibrate qualities

  → **BaseRecalibrator**

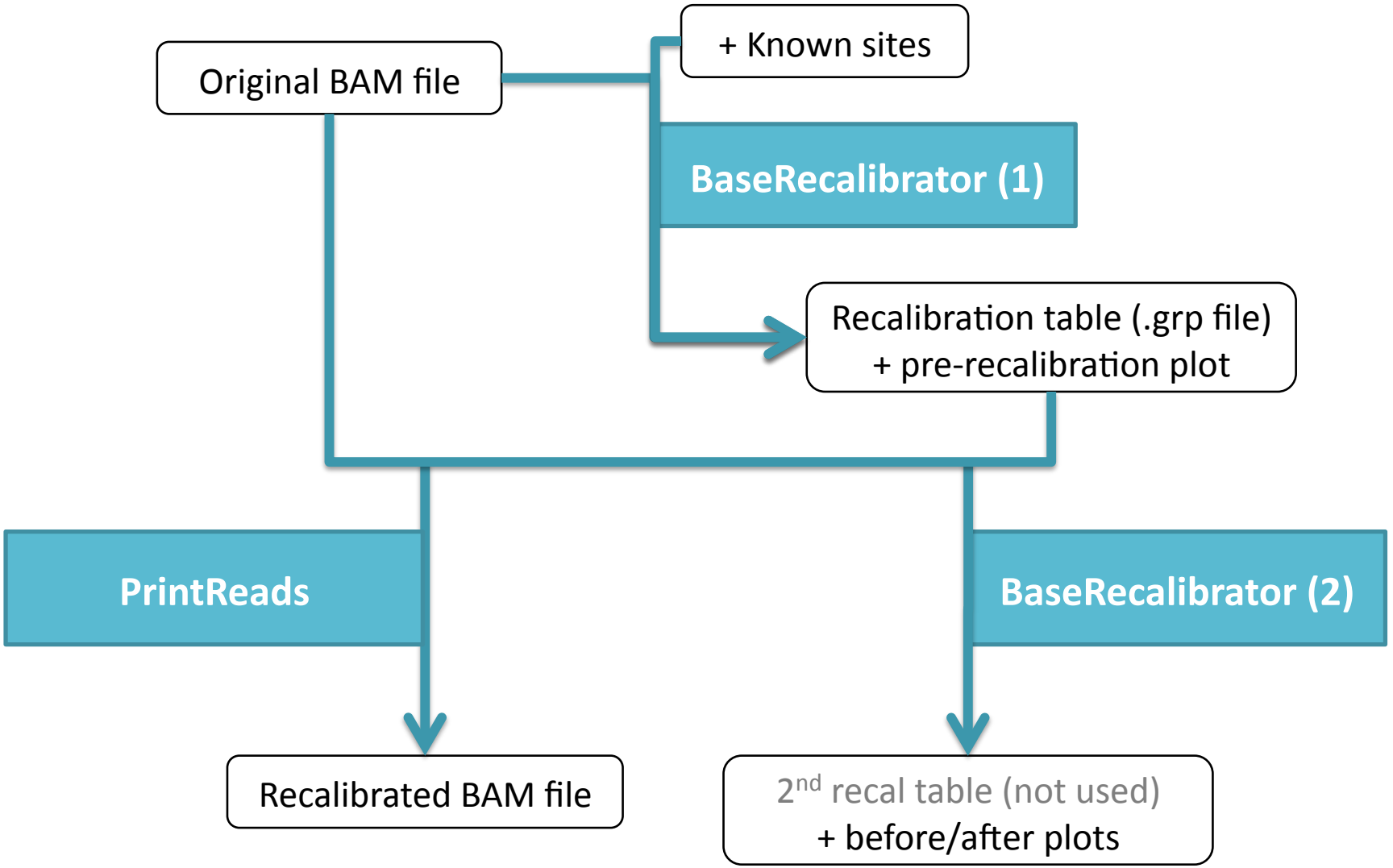



- Write the recalibrated data to file

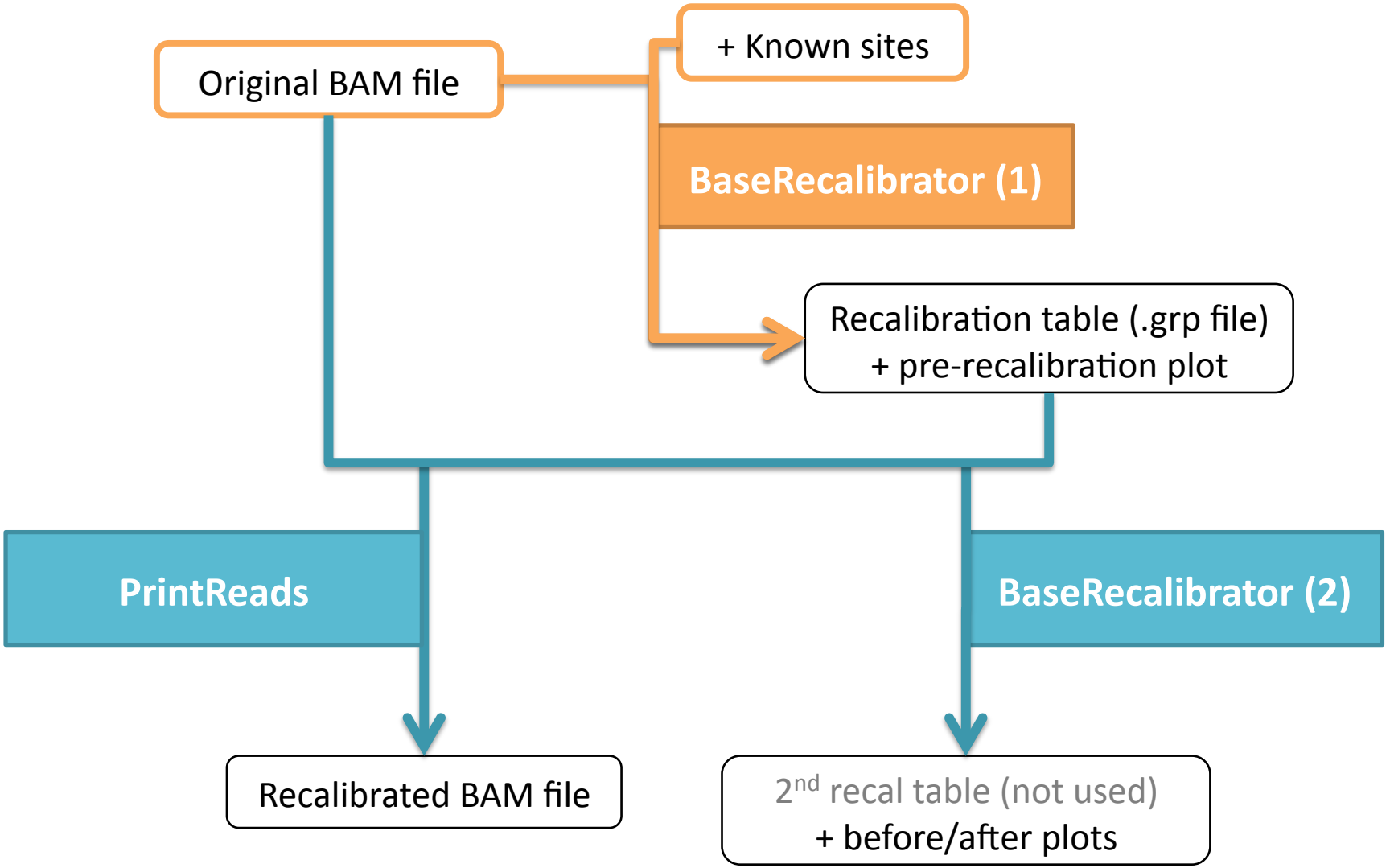
  → **PrintReads**



# Base Recalibration workflow



# Base Recalibration workflow



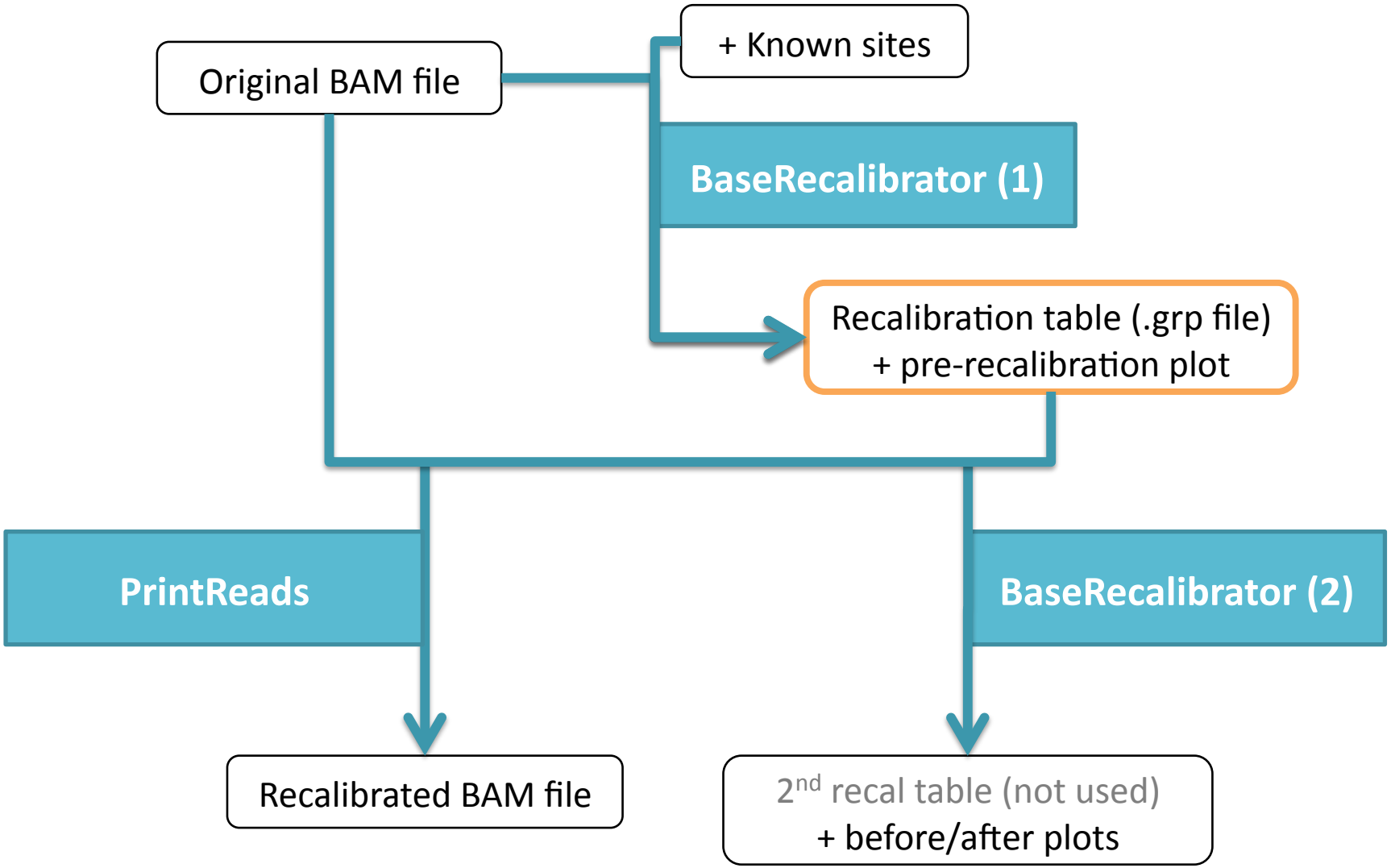
TOOL TIPS

# BaseRecalibrator

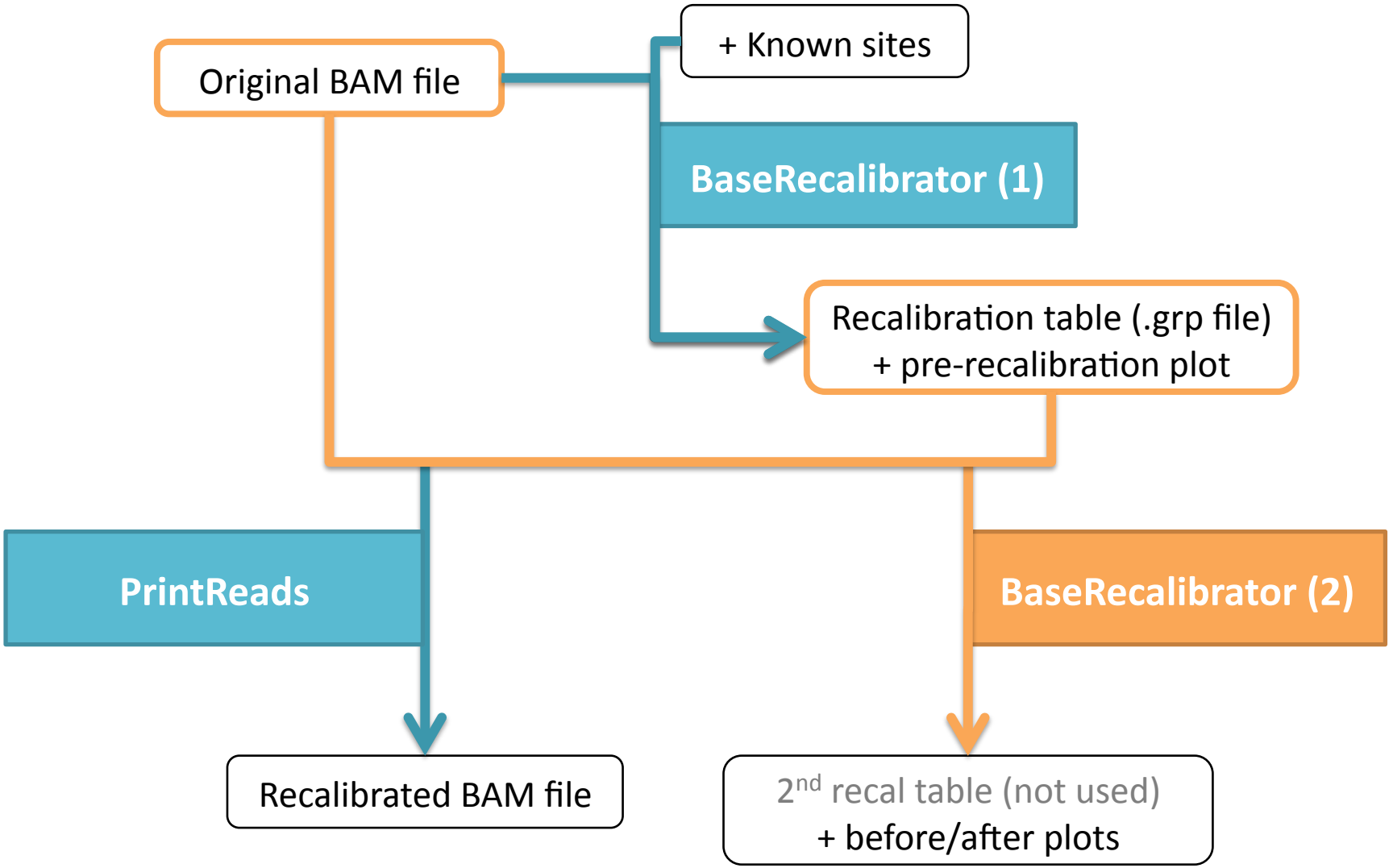
- Builds recalibration model and applies to data

```
java –jar GenomeAnalysisTK.jar –T BaseRecalibrator \
        –R human.fasta \
        –I realigned.bam \
        –knownSites dbsnp137.vcf \
        –knownSites gold.standard.indels.vcf \
        –o recal.grp \
        –plots recal.grp.pdf
```

# Base Recalibration workflow



# Base Recalibration workflow



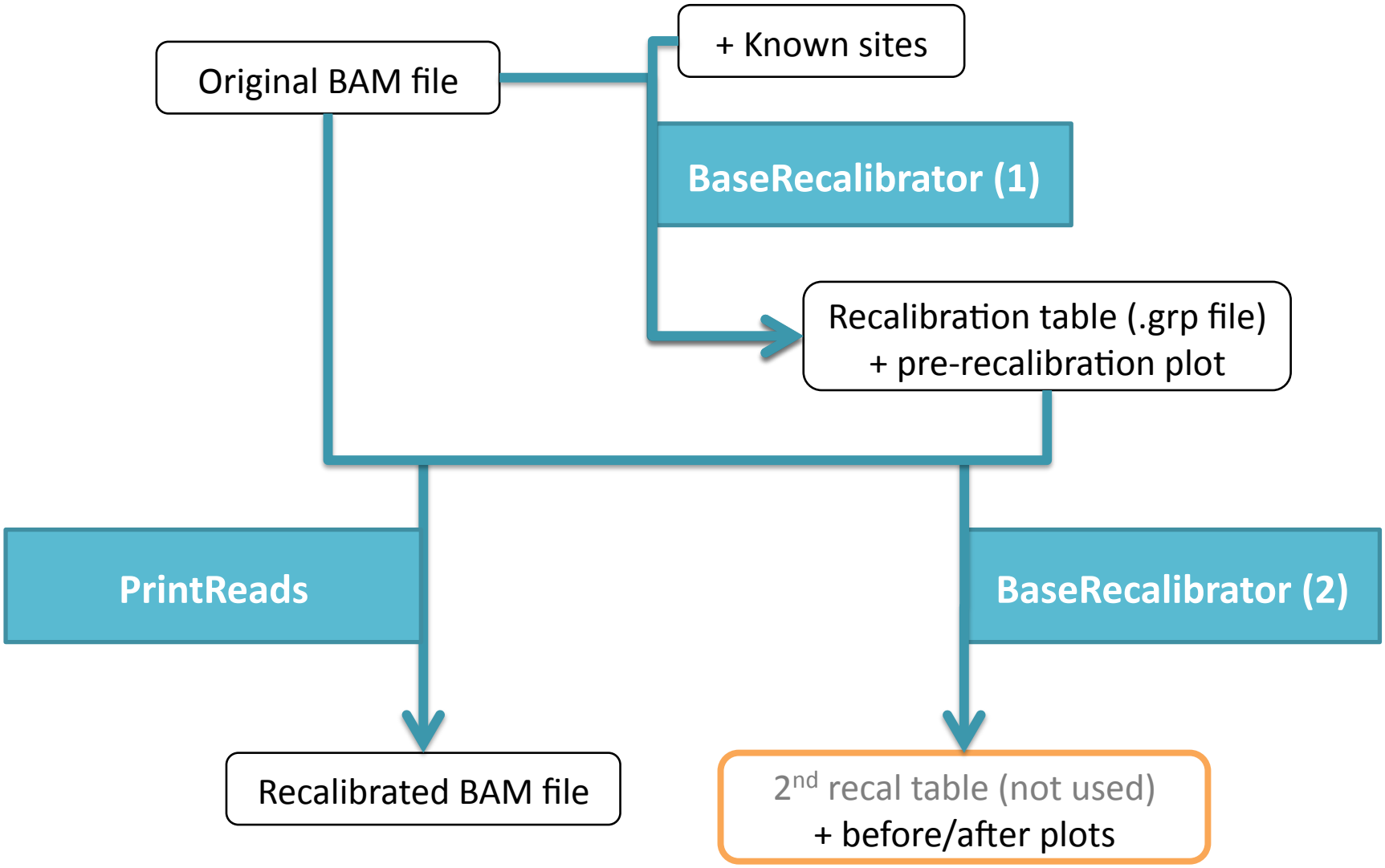
TOOL TIPS

# Base Recalibrator

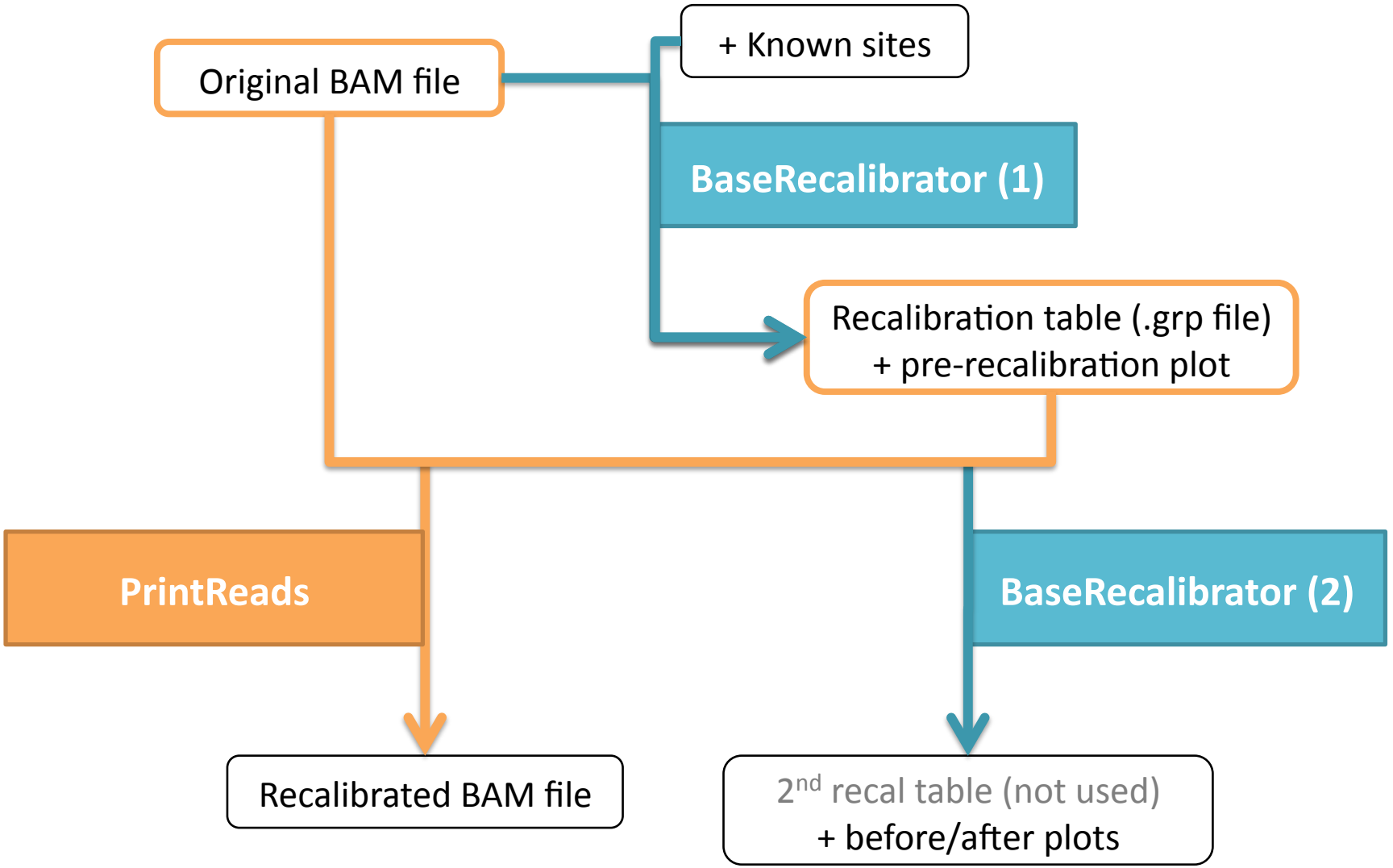
- Feed the first recalibration table to BaseRecalibrator to generate before/after plots

```
java –jar GenomeAnalysisTK.jar –T BaseRecalibrator \
      –R human.fasta \
      –I realigned.bam \
      –BQSR recal.grp \
      –o post_recal.grp \
      –plots post_recal.grp.pdf
```

# Base Recalibration workflow



# Base Recalibration workflow



TOOL TIPS

# Print Reads

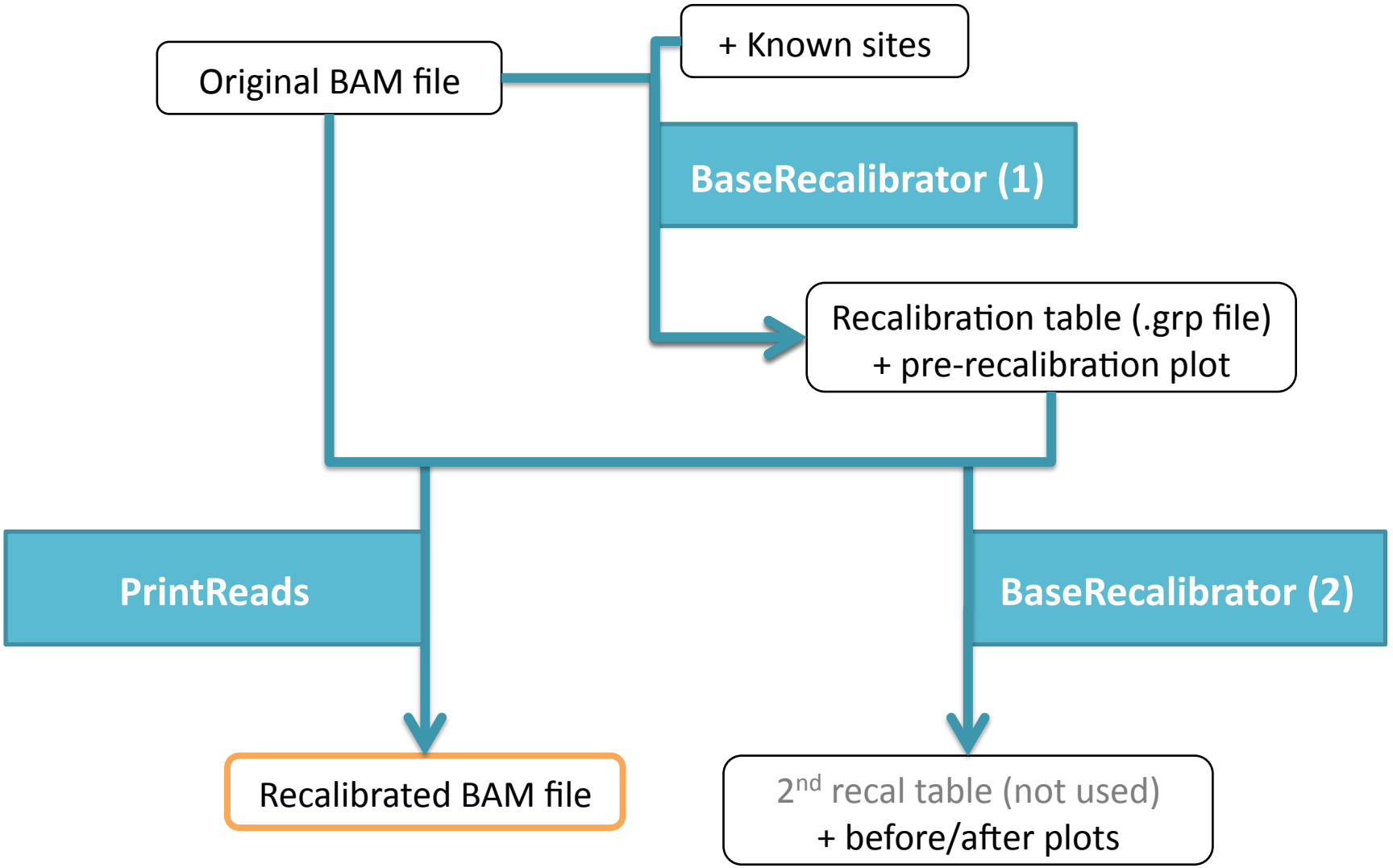
- General-use tool co-opted with –BQSR flag and fed a recalibration report

```
java –jar GenomeAnalysisTK.jar –T PrintReads \
      –R human.fasta \
      –I realigned.bam \
      –BQSR recal.grp \
      –o recal.bam
```

- Creates a new bam file using the input table generated previously which has exquisitely accurate base substitution, insertion, and deletion quality scores

- Original qualities retained with OQ tag

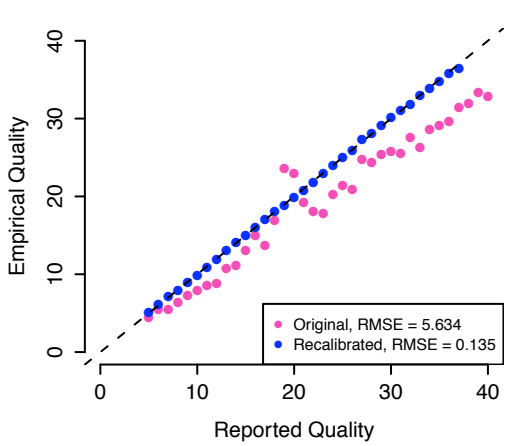
# Base Recalibration workflow

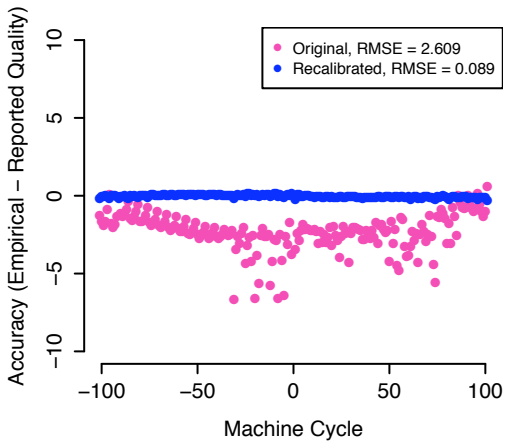


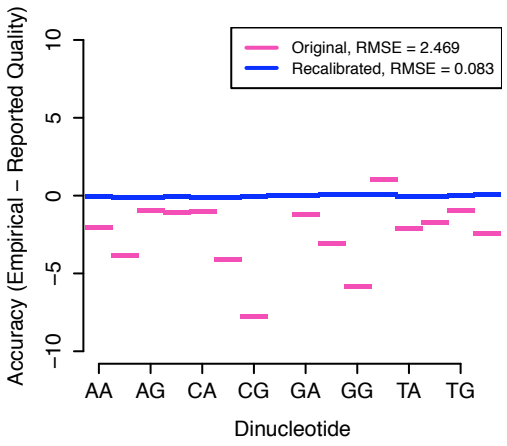
# RESULTS

# Did the recalibration work properly?

Post-recalibration quality scores should fit the empirically-derived quality scores very well; no obvious systematic biases should remain

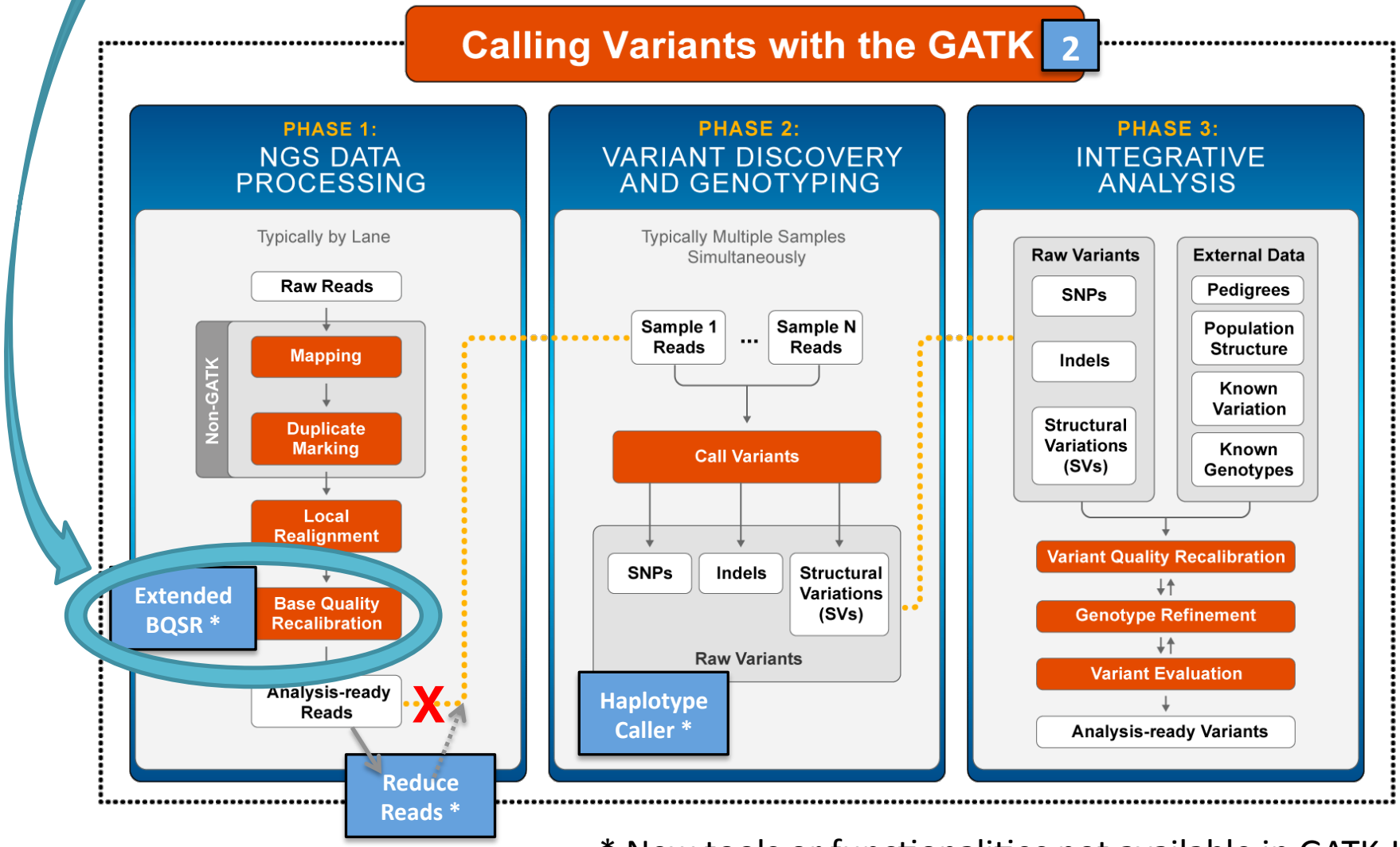


# We were here in the Best Practices workflow

***NEXT STEP: REDUCE READS***




* 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=44

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

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