GATK Best Practices for Variant Discovery

Base Quality Score Recalibration

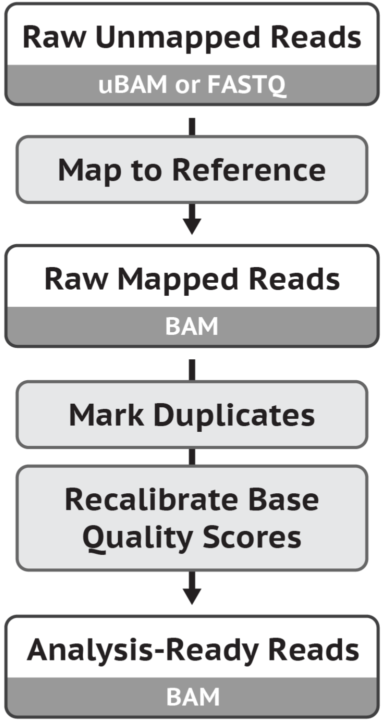
Assign an accurate confidence score
to each sequenced base

BROAD INSTITUTE

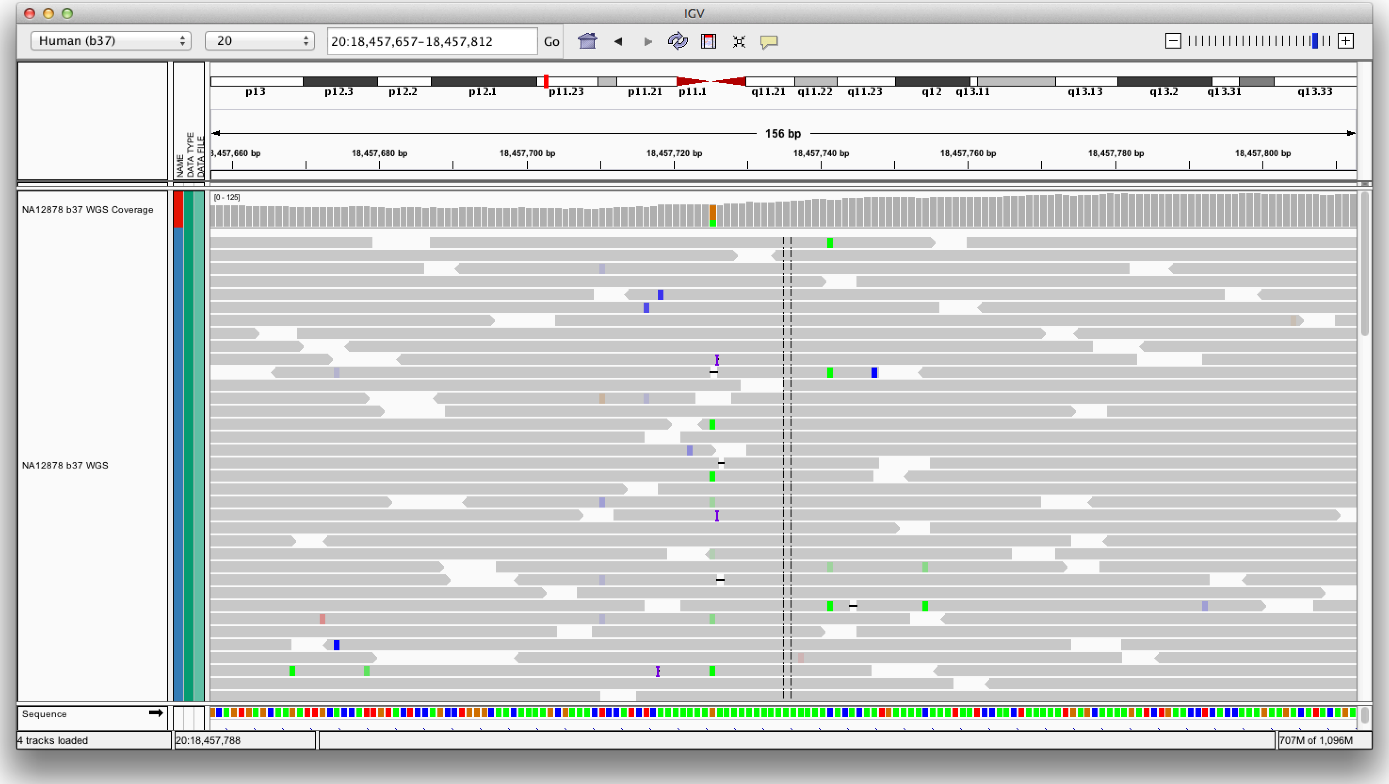
https://software.broadinstitute.org/gatk/

gatk

Data Pre-processing for Variant Discovery



Real data is messy -> properly estimating the evidence is critical



Human (b37)

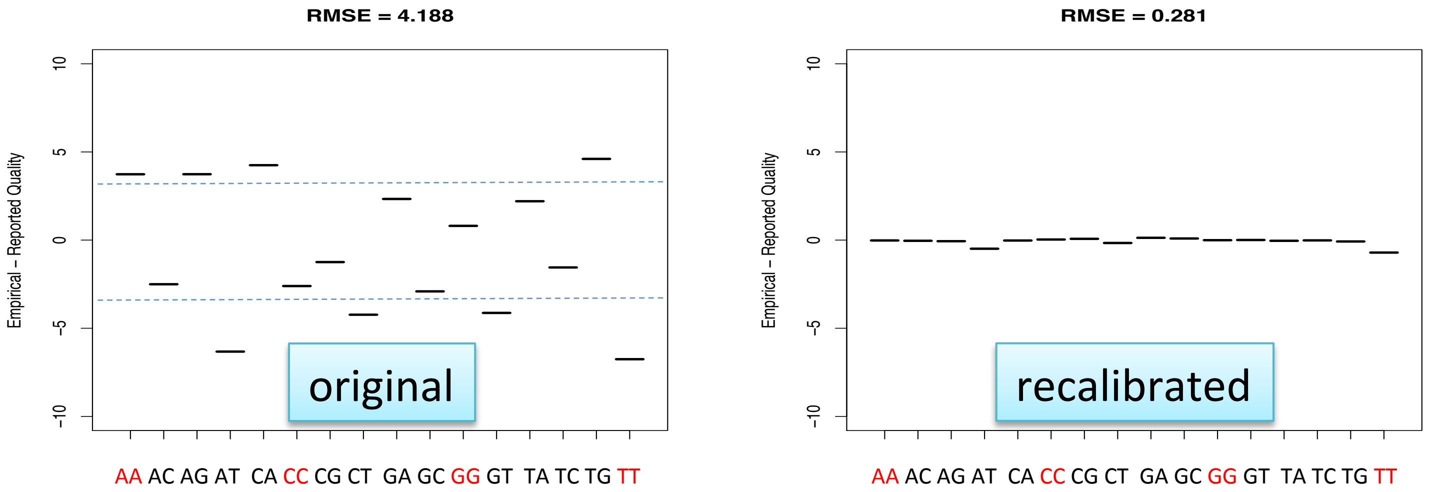
20

20:18,457,657-18,457,812

Quality scores issued by sequencers can be **inaccurate** and **biased**

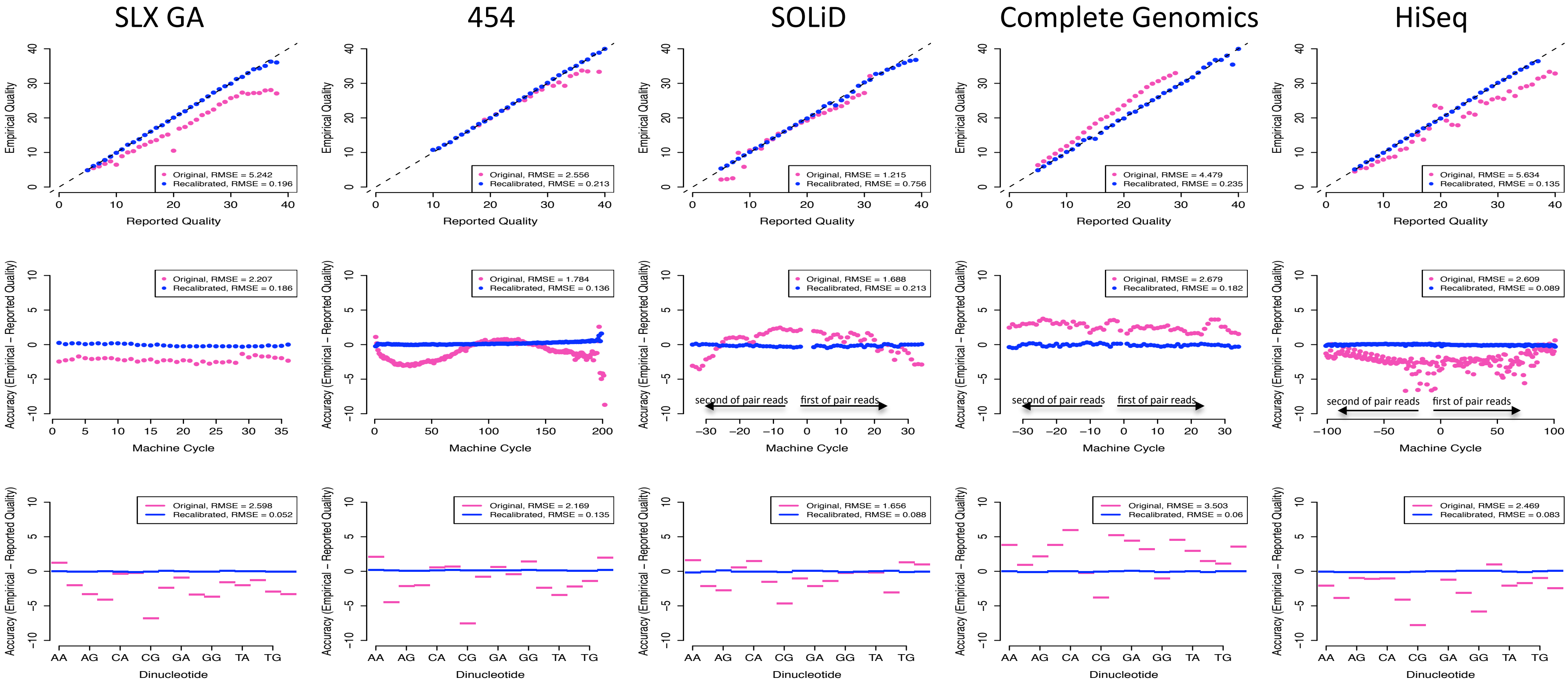
- Quality scores are critical for all downstream analysis
- Systematic biases are a major contributor to bad calls

Example of bias: qualities reported depending on nucleotide context



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

Different sequencing technologies / machines have different error modes


SLX GA

454

SOLiD

Complete Genomics

HiSeq

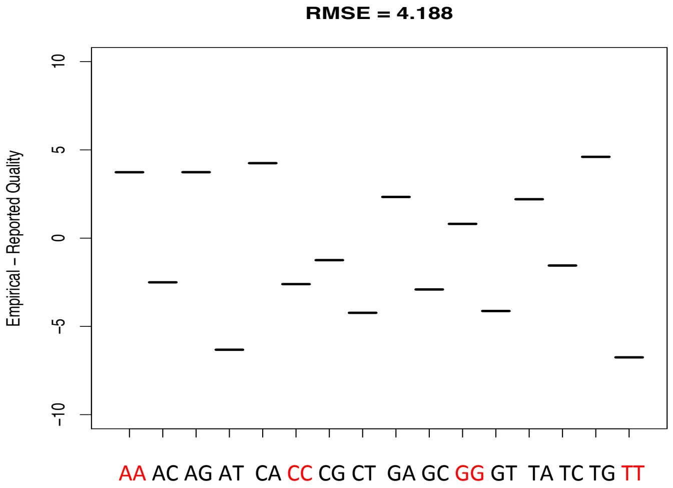
How do we identify the error modes in the data?

- Systematic biases can be found by looking at covariates:
  - Read group sample

    per-lane, per-sample
  - Reported base quality score
  - Position within the read

    machine cycle, first or second of pair
  - Sequence context

    e.g. di- and tri-nucleotide; for chemistry effects



- Calculate error empirically and find patterns in how error varies with basecall features

How do we calculate the empirical qualities?

- Any sequence mismatch = error ***except known variants*!***

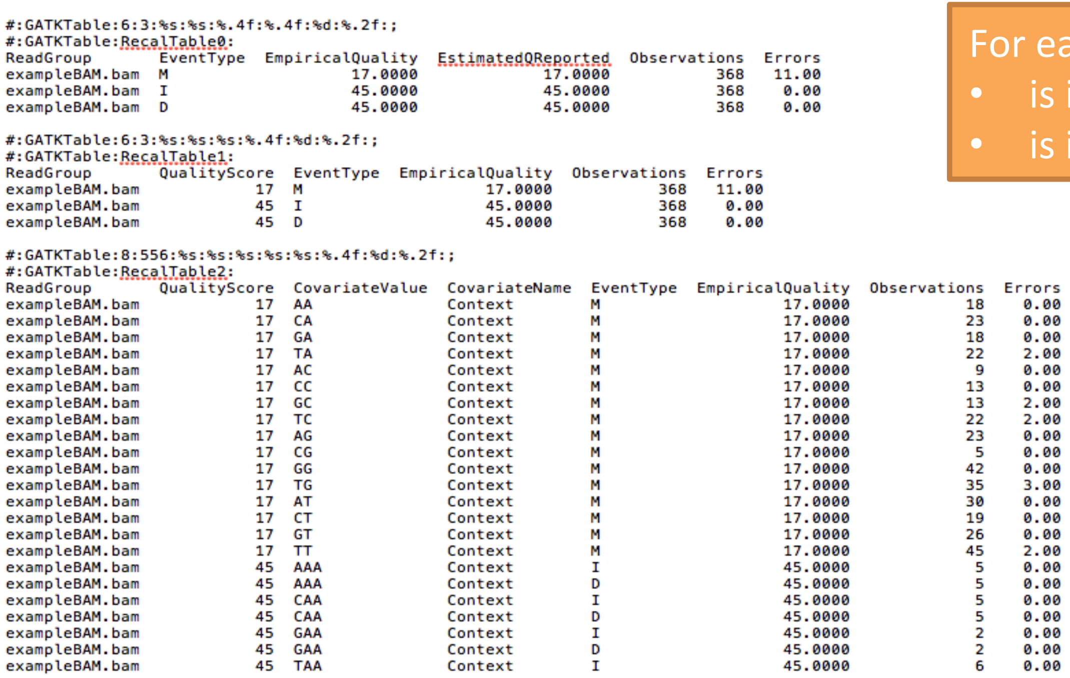
- Keep track of number of observations and number of errors as a function of various error covariates

  (lane, original quality score, machine cycle, and sequencing context)

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

** If you don’t have known variation, bootstrap it!*

Applying recalibration is simple



```
#:GATKTable:6:3:%s:%s:%.4f:%.4f:%d:%.2f:;
#:GATKTable:RecalTable0:
ReadGroup        EventType   EmpiricalQuality  EstimatedQReported  Observations  Errors
exampleBAM.bam   M                    17.0000             17.0000           368   11.00
exampleBAM.bam   I                    45.0000             45.0000           368    0.00
exampleBAM.bam   D                    45.0000             45.0000           368    0.00

#:GATKTable:6:3:%s:%s:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable1:
ReadGroup        QualityScore  EventType  EmpiricalQuality  Observations  Errors
exampleBAM.bam             17  M                   17.0000           368   11.00
exampleBAM.bam             45  I                   45.0000           368    0.00
exampleBAM.bam             45  D                   45.0000           368    0.00

#:GATKTable:8:556:%s:%s:%s:%s:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable2:
ReadGroup        QualityScore  CovariateValue  CovariateName  EventType  EmpiricalQuality  Observations  Errors
exampleBAM.bam             17  AA              Context        M                   17.0000            18    0.00
exampleBAM.bam             17  CA              Context        M                   17.0000            23    0.00
exampleBAM.bam             17  GA              Context        M                   17.0000            18    0.00
exampleBAM.bam             17  TA              Context        M                   17.0000            22    2.00
exampleBAM.bam             17  AC              Context        M                   17.0000             9    0.00
exampleBAM.bam             17  CC              Context        M                   17.0000            13    0.00
exampleBAM.bam             17  GC              Context        M                   17.0000            13    2.00
exampleBAM.bam             17  TC              Context        M                   17.0000            22    2.00
exampleBAM.bam             17  AG              Context        M                   17.0000            23    0.00
exampleBAM.bam             17  CG              Context        M                   17.0000             5    0.00
exampleBAM.bam             17  GG              Context        M                   17.0000            42    0.00
exampleBAM.bam             17  TG              Context        M                   17.0000            35    3.00
exampleBAM.bam             17  AT              Context        M                   17.0000            30    0.00
exampleBAM.bam             17  CT              Context        M                   17.0000            19    0.00
exampleBAM.bam             17  GT              Context        M                   17.0000            26    0.00
exampleBAM.bam             17  TT              Context        M                   17.0000            45    2.00
exampleBAM.bam             45  AAA             Context        I                   45.0000             5    0.00
exampleBAM.bam             45  AAA             Context        D                   45.0000             5    0.00
exampleBAM.bam             45  CAA             Context        I                   45.0000             5    0.00
exampleBAM.bam             45  CAA             Context        D                   45.0000             5    0.00
exampleBAM.bam             45  GAA             Context        I                   45.0000             2    0.00
exampleBAM.bam             45  GAA             Context        D                   45.0000             2    0.00
exampleBAM.bam             45  TAA             Context        I                   45.0000             6    0.00
```

For each base in each read:

- is it in AA context? -> adjust by X points
- is it at 3rd position? -> adjust by Y points

Generates exquisitely accurate base substitution, insertion and deletion quality scores

Base recalibration (BQSR) overview

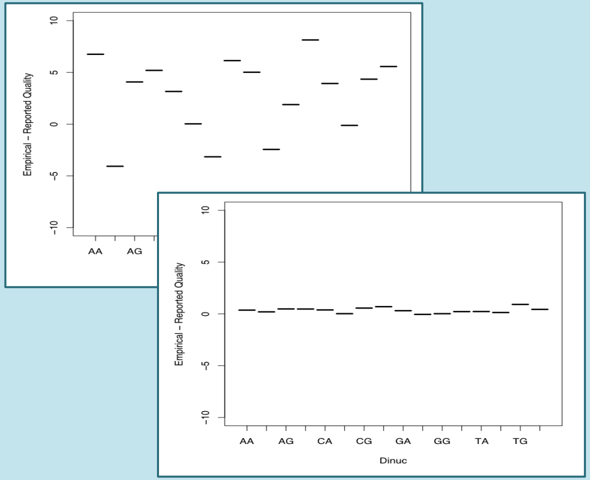
- Model the error modes and compute adjustments
  → **BaseRecalibrator**

- If parallelizing over a sample, combine scattered tables
  → GATK4: **GatherBQSRReports**

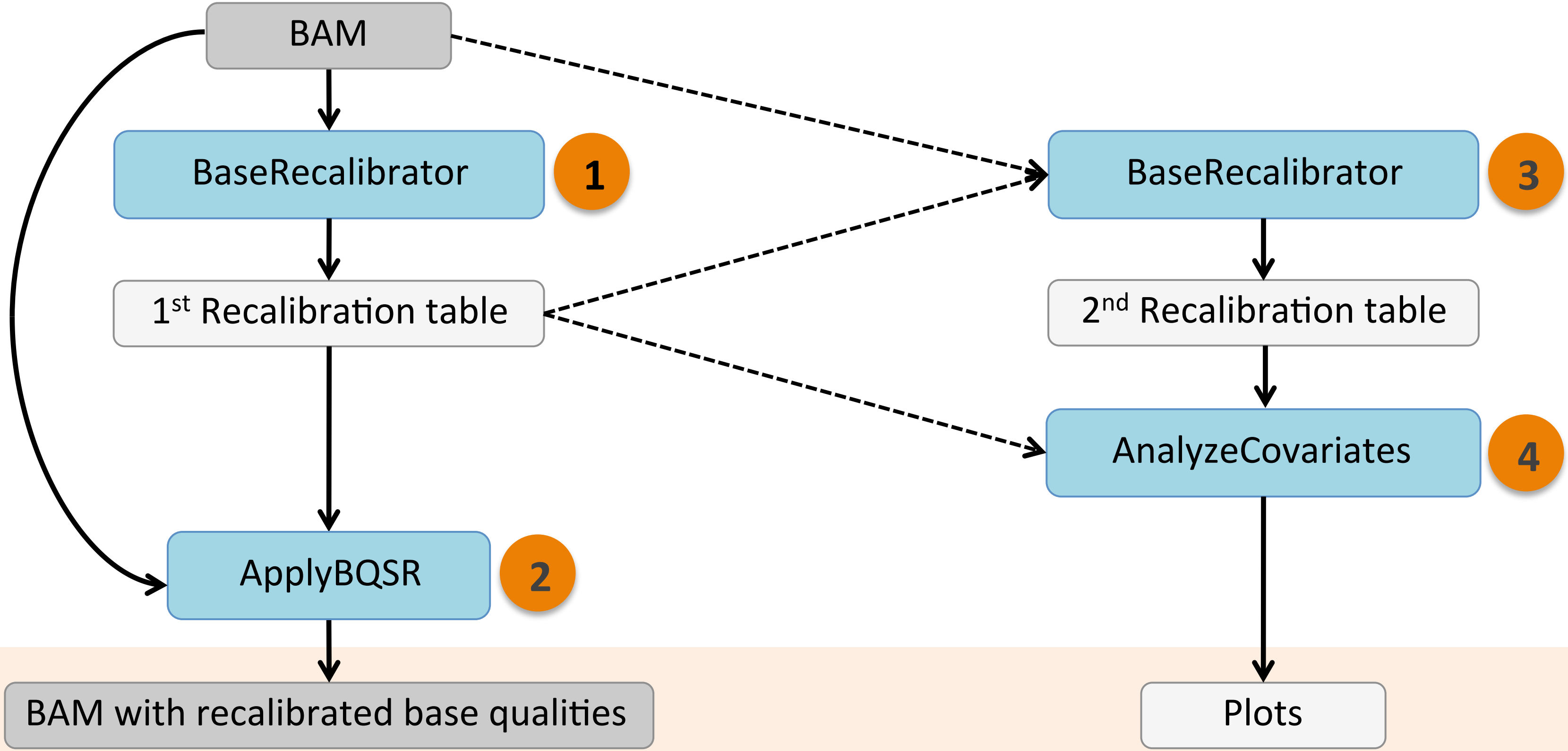
- Apply recalibration adjustments to BAM
  → GATK3: **PrintReads**
  → GATK4: **ApplyBQSR**

- Make before and after plots

→ **AnalyzeCovariates**



Two complementary paths: data processing and quality control



Steps 1 and 3: Calculate covariate bias with BaseRecalibrator

Build base recalibration model

```
gatk BaseRecalibrator \
    -R ref.fasta \
    -I sample.bam \
    -knownSites snps.vcf.gz \
    -knownSites indels.vcf.gz \
    -O recal.table
```

To generate the 2nd recal table, include the 1st with:

```
    -bqsr 1st_recal.table
```

Step 2: Apply recalibration with ApplyBQSR

Recalibrate base qualities in GATK3:

```
gatk ApplyBQSR \
    -R ref.fasta \
    -I sample.bam \
    -bqsr recal.table \
    -O sample_bqsr.bam
```

To bin quals (an example implementation):

```
    -SQQ 10 -SQQ 20 -SQQ 30 -SQQ 40
```

To emit original quals to OQ tag:

```
    --emit_original_quals
```

Some BQSR options that impact BAM file compression

- Bin BQs using `--static_quantized_quals` (`-SQQ`)
  - Our germline production pipelines use four bins at 10, 20, 30 and 40 (https://software.broadinstitute.org/gatk/documentation/article?id=7899)
  - Rounds in probability space, e.g. 7 to 12 rounds to 10.
- Original qualities (OQ) are tossed by default
  - Retain with `--emit_original_quals`
- BQs less than 6 are untouched. Change threshold with `--preserve_qscores_less_than`
- Our tools currently do not use base indel quality scores (BI and BD tags).
  - GATK4 ApplyBQSR omits these by default
  - GATK3 recalibration emits these. Remove with `--disable_indel_quals`

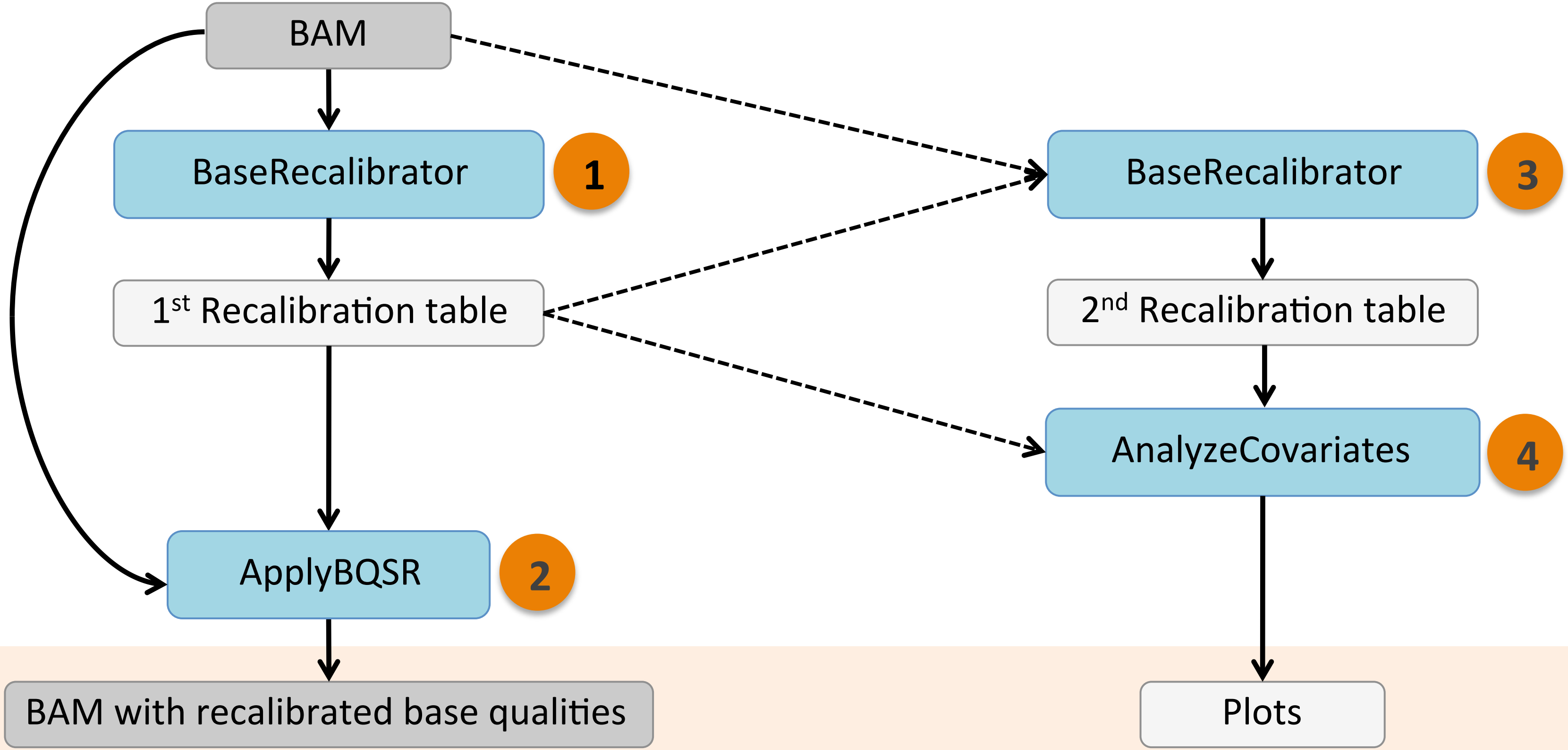
Example recalibrated SAM record with OQ tag

Recalibrated Base Qualities

```
ACCTTCCCCCAGCCCCTACCCCCAGACAGGCCCCGGTGTGTTGTGTTCCCCTCCCTCT
GTCCATGTGTTCTCATTGTTCAACTCTCATTTATGAGTGAGAACATCGGGGGTTTGGT
TTTCTGTTCTTGGATTAGTTTGGTGAGAATGATGG  <;<>==>=>>6>=>>>??
+<>>>?3::*<>8=>>8?/=.3/7;<<;>=???>???@=1=>=?+=>?
=.<=A@;??,>?=;4:?>1>+>=?:@=>?/;4??<@+??9<;+8/
<-,?:<@>:@=/-.@>=@9/?)=6???+:@=B=#######   MC:Z:151M MD:Z:
108T29C12 PG:Z:MarkDuplicates.4  RG:Z:H01PE.2 NM:i:2 MQ:i:
0  OQ:Z:AAFFAFJFJJ<FFJJJJJ-AJJJJ7AA-AJ<FJJJJ-F-7-
<AAAAJFJJJFJJJJF-FFFJ-FFJF-FFJJAJJ-FJAA7AAF-F-FFJAJAFF-
A7FFAJ-FFFAA-<-A--F<AJF<FA---AFAF<-F-A7FFF-<FAJA#######
   UQ:i:24   AS:i:141
```

Original Base Qualities

Two complementary paths: data processing and quality control

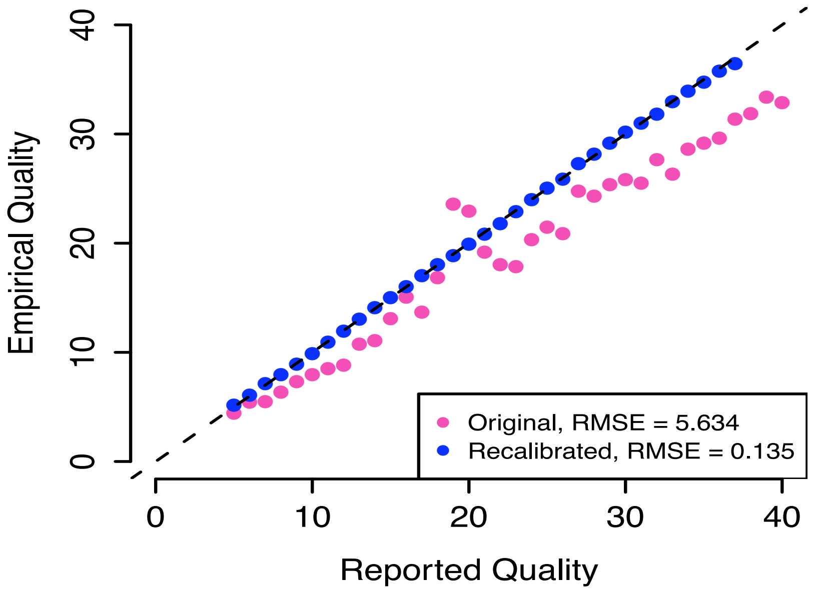


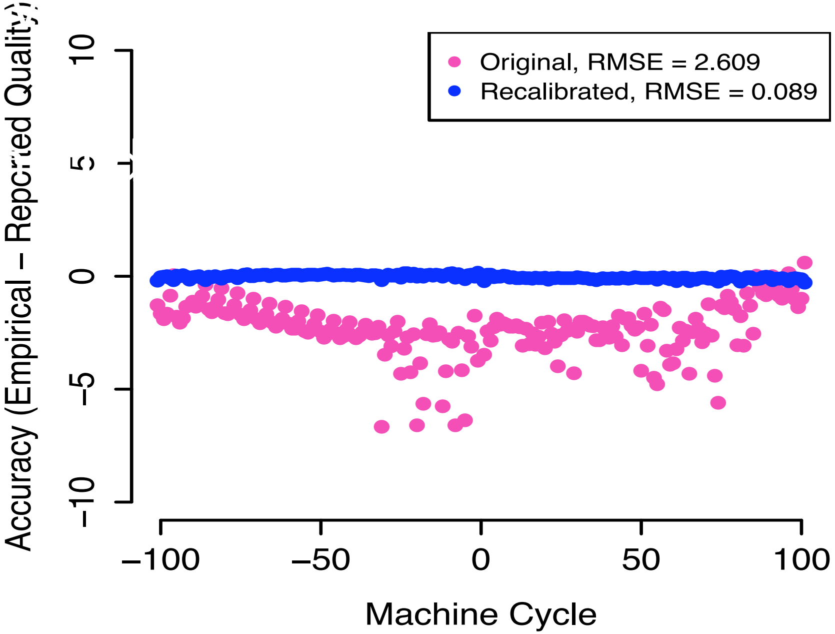
Step 4: Make QC plots with AnalyzeCovariates

Generate before and after plots

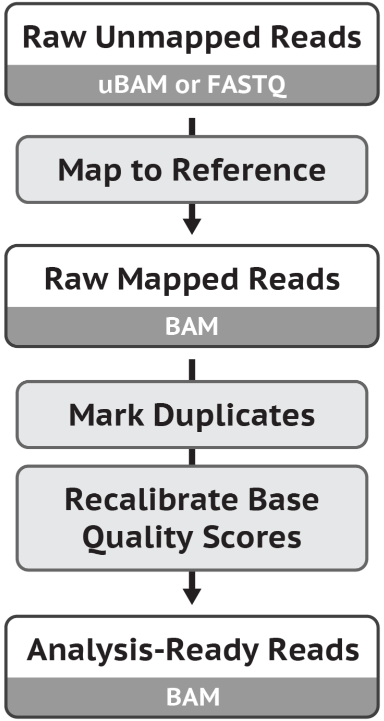
```
gatk AnalyzeCovariates \
   -before 1st_recal.table \
   -after 2nd_recal.table \
   -plots plots.pdf
```

Plots show effectiveness of recalibration





Data Pre-processing for Variant Discovery


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM