MPG NSG workshop I: Base quality score recalibration

Ryan Poplin

Genome Sequencing and Analysis
Broad Institute of Harvard and MIT

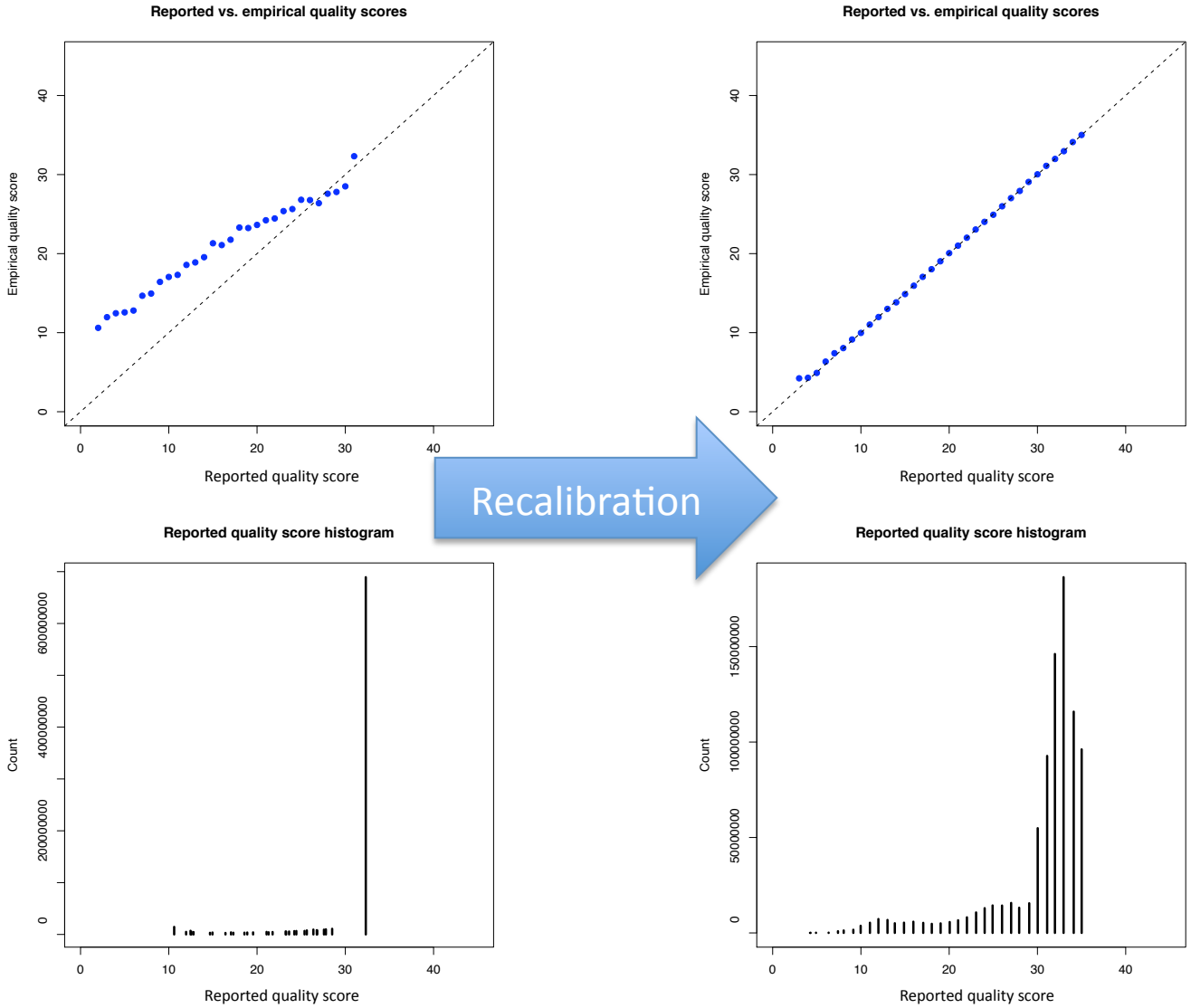
02/04/10

Introduction to Base Quality Score Recalibration

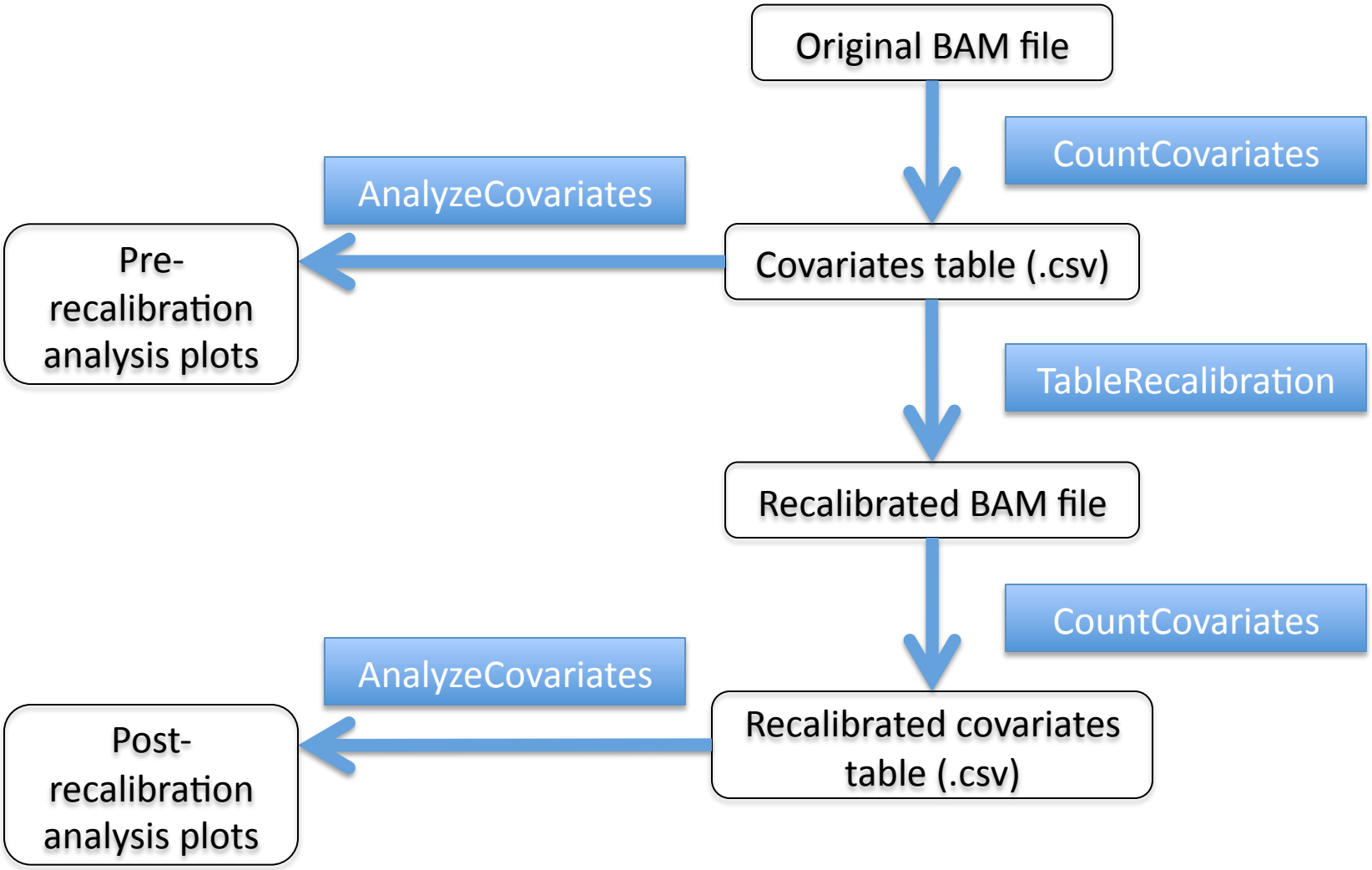
Sequencers provide estimates of error rate per nucleotide

… but they aren't very accurate

… and they aren't very informative



Recalibration workflow



Running CountCovariates

```
java -Xmx4g -jar GenomeAnalysisTK.jar
  -R Homo_sapiens_assembly18.fasta
  -D dbsnp_129_hg18.rod
  -I original.bam
  -T CountCovariates
  -cov ReadGroupCovariate
  -cov QualityScoreCovariate
  -cov DinucCovariate
  -cov CycleCovariate
  -recalFile table.recal_data.csv
```

List of known polymorphic sites is strongly recommended so they won't count against base's mismatch rate

List of covariates to be used in the recalibration calculation

CSV file containing covariate counts

Table recalibration file (table.recal_data.csv)

```
# Counted Bases    143745620
ReadGroup,QualityScore,Dinuc,Cycle,nObservations,nMismatches,Qempirical
SRR001802,2,AA,-8,165,17,10
SRR001802,2,AA,-2,91,10,10
SRR001802,2,AA,3,5,4,1
SRR001802,2,AA,4,9,4,4
SRR001802,2,AA,7,12,4,5
```

See http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration for more information

Running TableRecalibration

```
java -Xmx4g -jar GenomeAnalysisTK.jar
  -R Homo_sapiens_assembly18.fasta
  -I original.bam
  -T TableRecalibration
  -recalFile table.recal_data.csv
  -outputBam recal.bam
```

Table recalibration file from CountCovariates step

The full recalibrated bam file

A recalibrated copy of the original BAM file

See http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration for more information

Running AnalyzeCovariates

A separate .jar file distributed with the GATK

```
java -Xmx4g -jar AnalyzeCovariates.jar
   -outputDir /path/to/output_dir/
   -resources resources/
   -recalFile table.recal_data.csv
```

The directory in which to place the output analysis plots

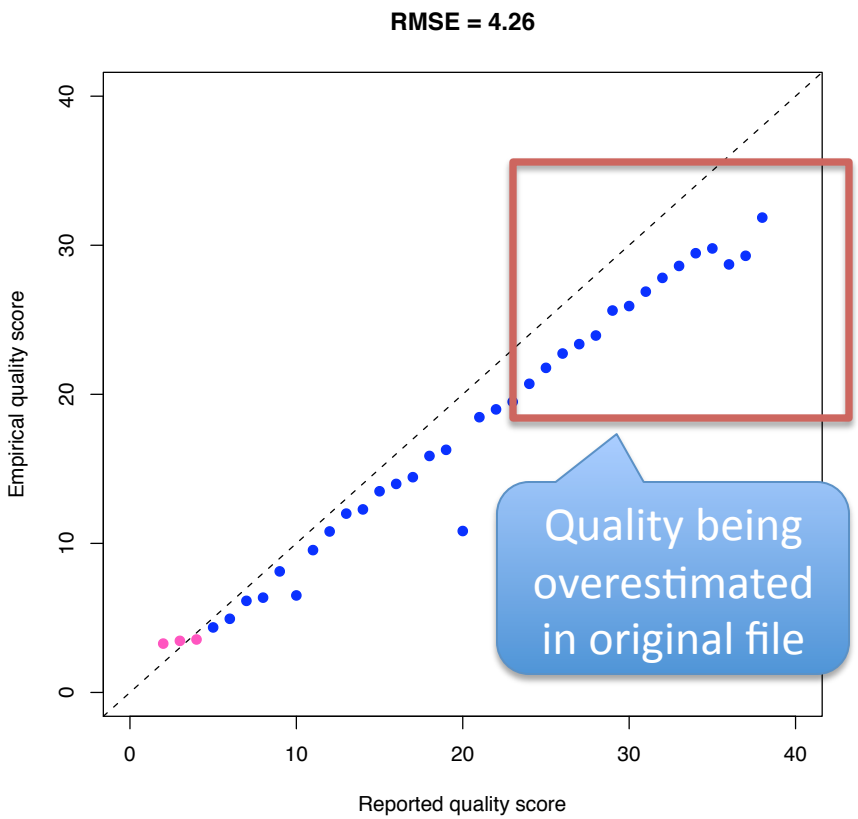
Points to the GATK installation's directory of R scripts which are used for plotting the data

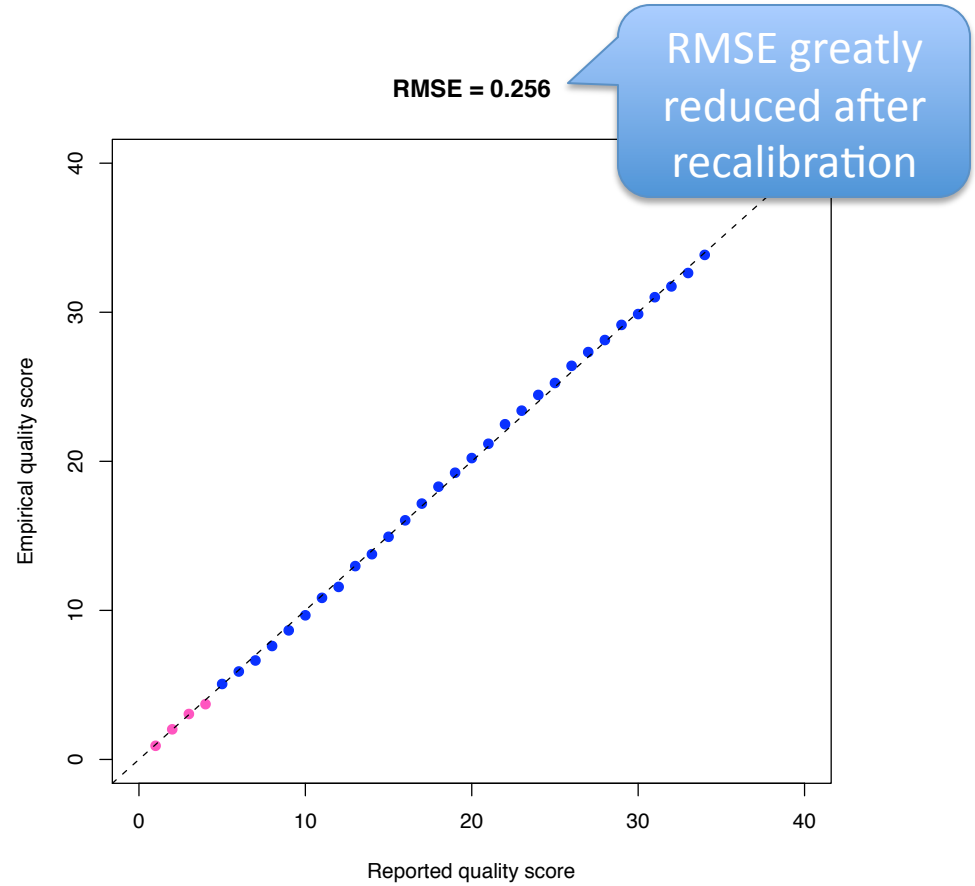
Table recalibration file from either the before or after CountCovariates step

Many plots of base quality versus each covariate

See http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration for more information

Reported vs Empirical Quality



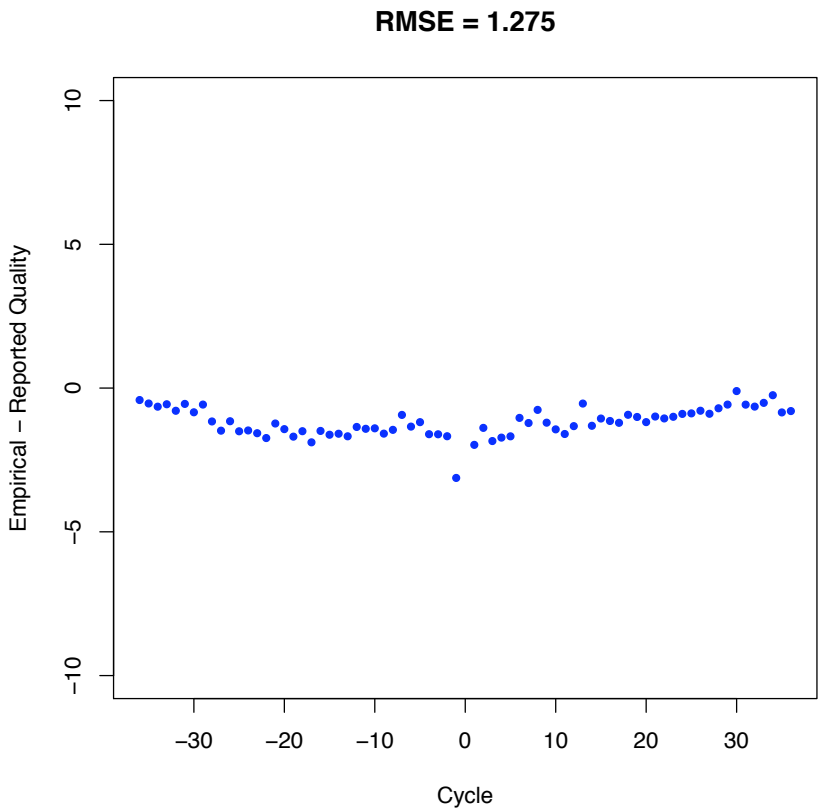


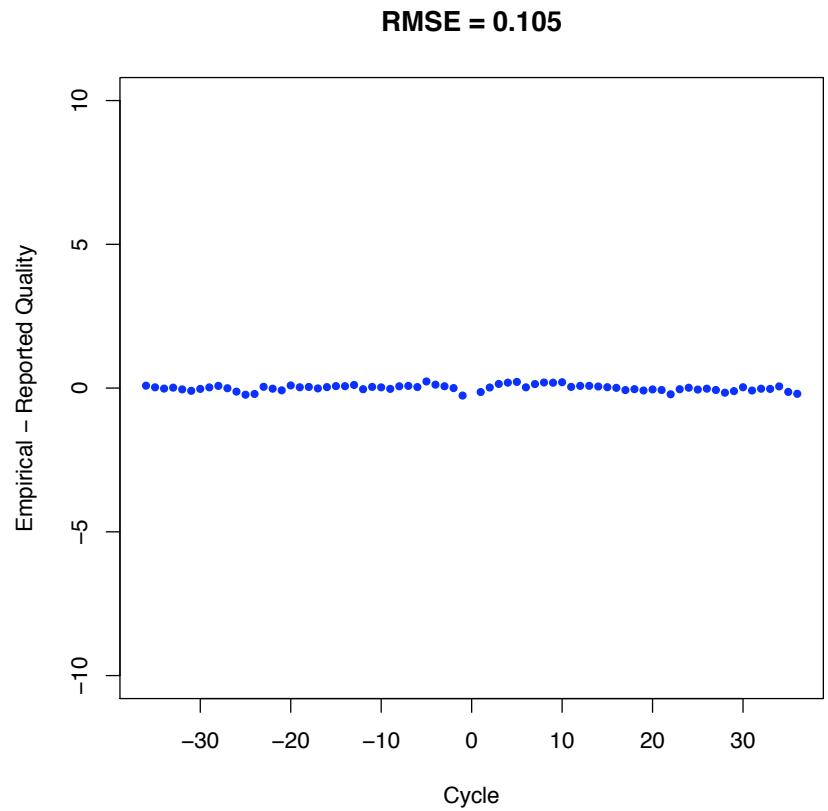
Before Recalibration

After Recalibration

* Data from 1KG, whole genome, single sample, deep coverage, Illumina data

Residual Error by Machine Cycle



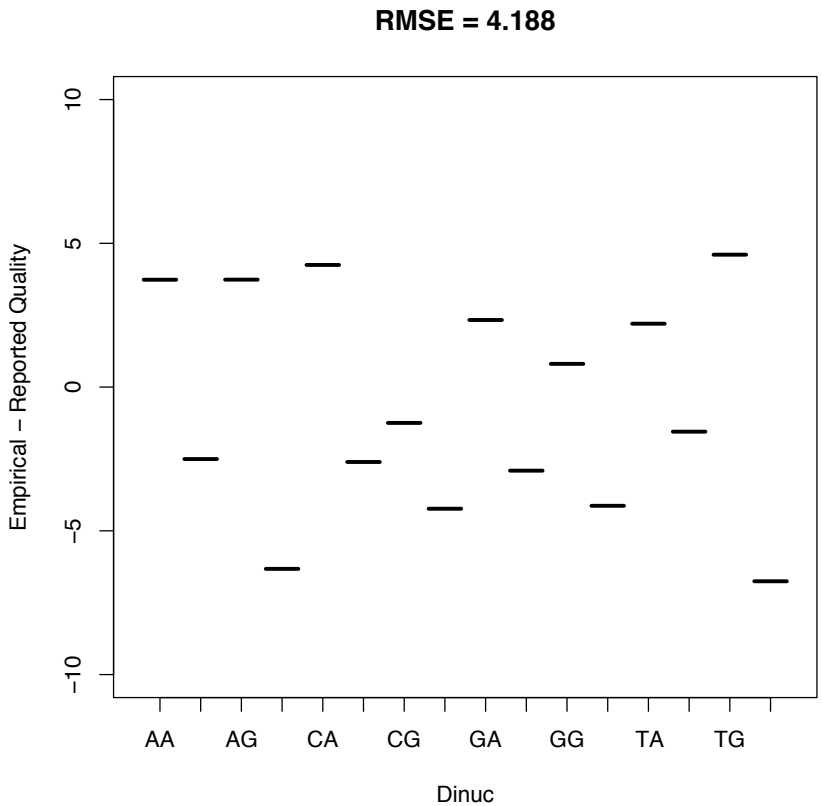


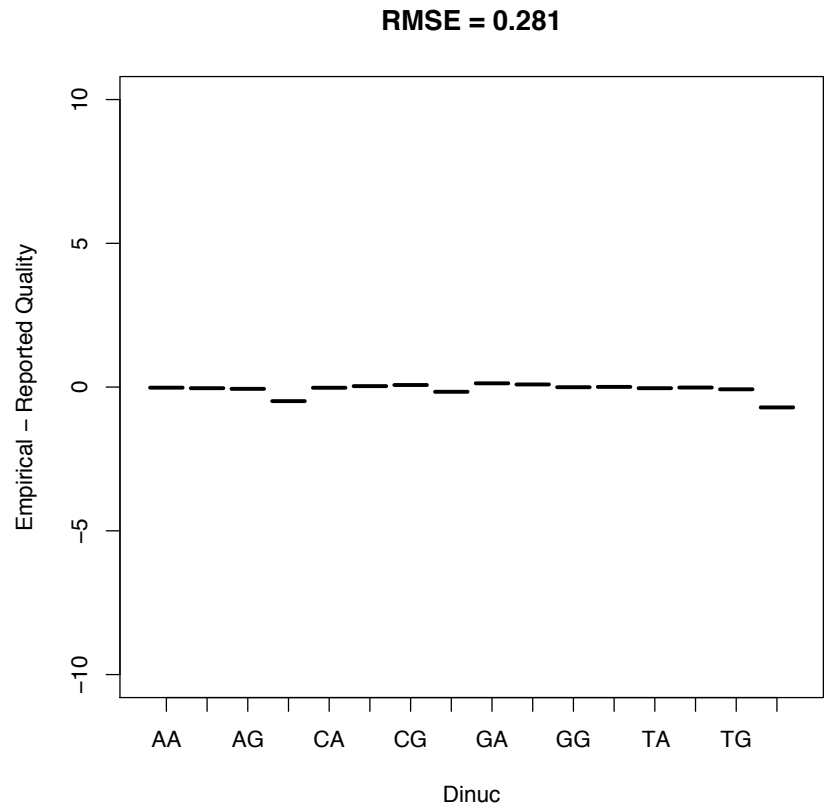
Before Recalibration

After Recalibration

* Data from 1KG, whole genome, single sample, deep coverage, Illumina data

Residual Error by Dinucleotide





Before Recalibration

After Recalibration

* Data from 1KG, whole genome, single sample, deep coverage, Illumina data

Recalibration Tidbits

- Many additional analysis plots available
- Many advanced options described on wiki
- GATK Recalibrator now in the Picard pipeline
  - All recent Broad-produced data is already recalibrated