Variant Calling

Cesky Krumlov

May 25, 2022

Marking/Removing Duplicates

- Reads can be artificial duplicates
  - PCR duplicates during library prep
  - Optical duplication (reads one cluster as two)
- These are not independent observations
  - Skew results for depth counts and allele frequency
  - Reads with PCR errors double counted
- We want to either mark or remove these

Detecting Duplicates

- Single reads
  - Reads are same strand and start at same position
- Paired reads
  - Both reads of pair start at same positions
  - Much more predictive than for single reads
- If physically close on sequencer, call optical
- Mark instead of remove
  - Allows data to be retained, software can ignore

Detecting without Reference

- Mostly used for metagenomic analysis
- Detect reads/pairs with identical starts
  - First 6-12 bases are exactly the same
- Align those reads to each other
- Remove reads/pairs which meet alignment thresholds

Should You Ignore Duplicates?

- Yes, if high complexity and low coverage
  - Almost all duplicates likely artifact
- Low complexity or high coverage less clear
  - You expect some number of random duplicates
  - These may be real independent data
  - Discarding them may skew results
- Much harder to accurately call for single reads

Base Quality Scoring

- Quality score measures probability of a base being incorrect: P(error) = $10^{-Q/10}$
  - Q = 10, p(error) = 0.1
  - Q = 20, p(error) = 0.01
- To get the Q value: Q = $-10 \bullet \log_{10}(P(\text{error}))$

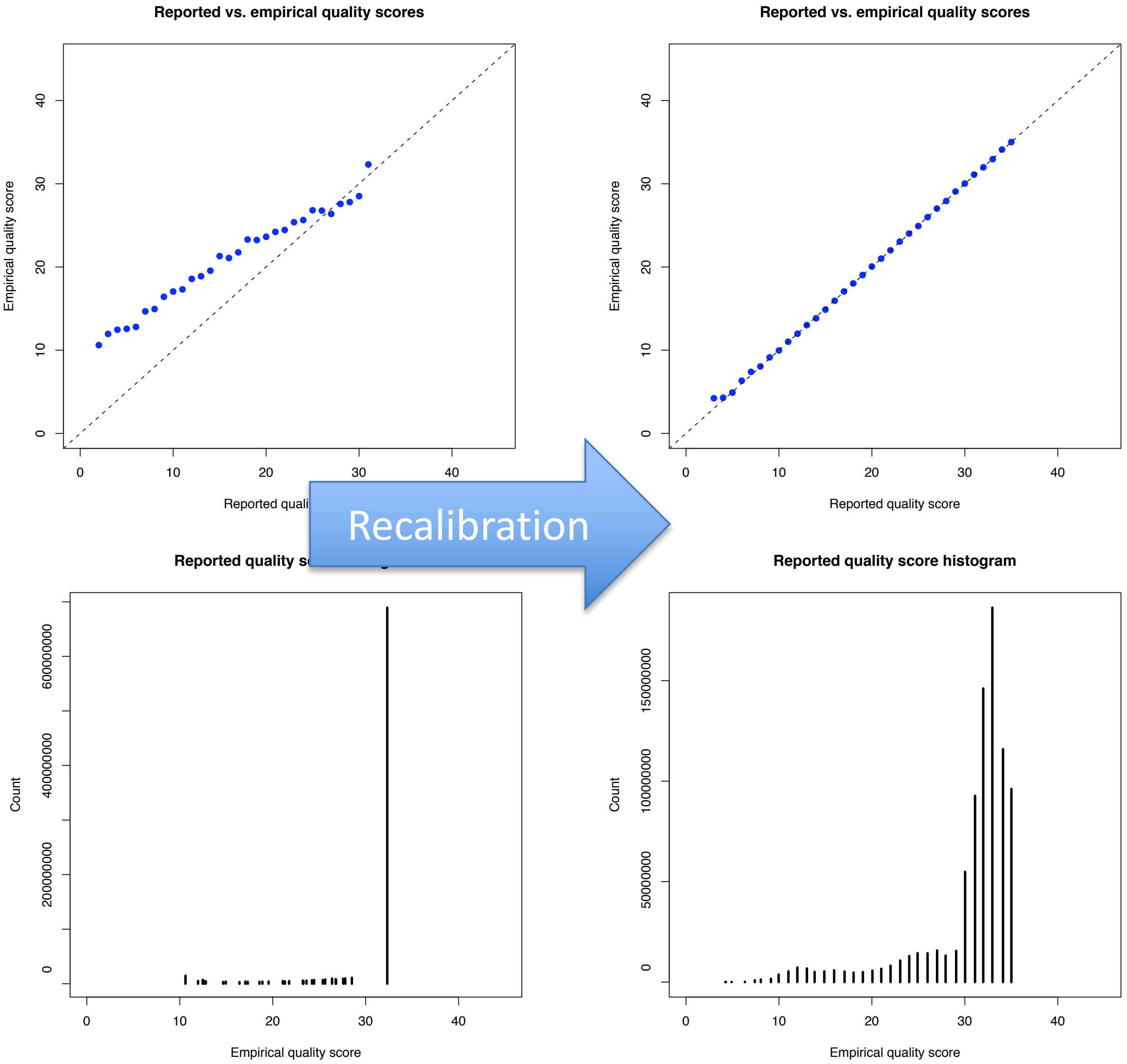
Base Quality Scoring

- Can help use of poor quality data
- Scalar quality says nothing about other bases
- Can use to weight value of bases in a single read for consensus building or SNP calling
- At very deep coverage, can be less important
- Still valuable for variant filtering

Alignment Quality Scoring

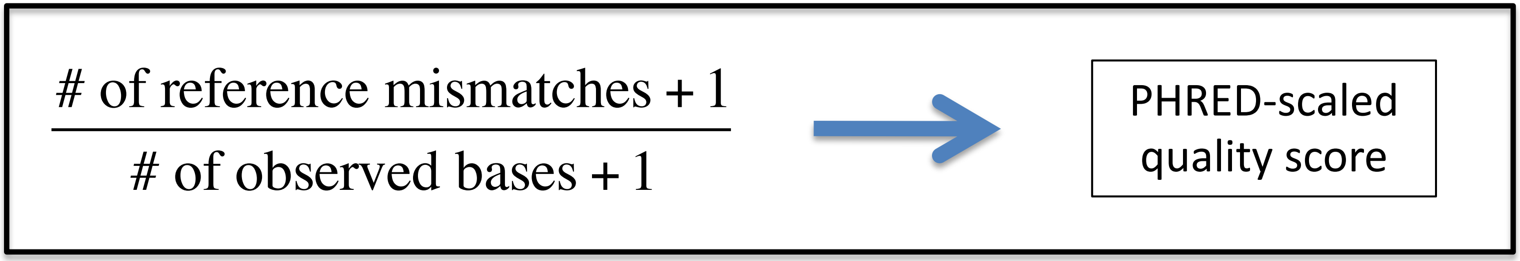
- Measures the probability the placement of a read on the reference is in error
- Roughly a measure of how likely it is that the read has enough errors that another placement is correct
- Generally, reads with multiple identical matches have mapping quality 0
- Useful for SNP calling and QC of alignments

Base Quality Recalibration



Recalibration Method

- Bin each base according to
  - Read group
  - Called quality
  - Position in read
  - Local dinucleotide context
- Score observed quality for each bin

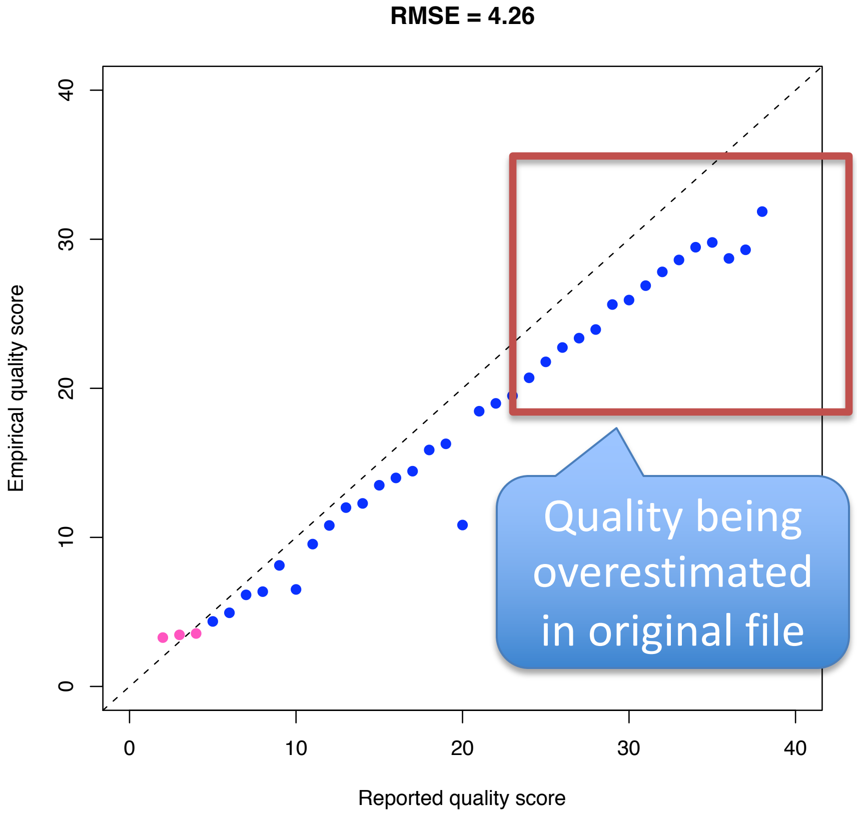


$$\frac{\text{\# of reference mismatches} + 1}{\text{\# of observed bases} + 1}$$

PHRED-scaled quality score

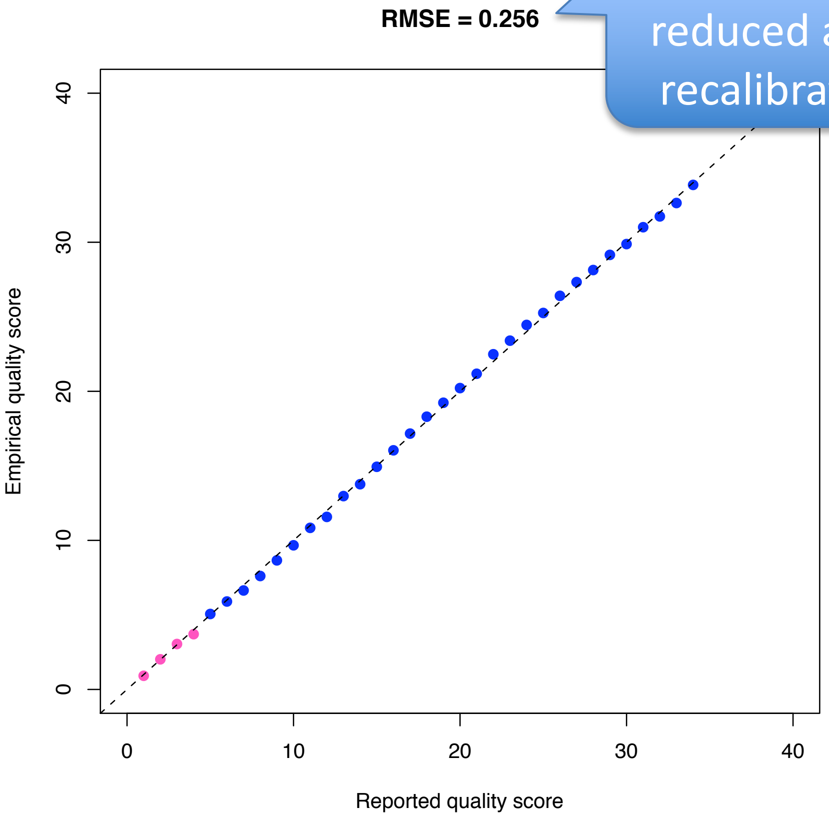
- Ignore known SNP positions

Reported vs Empirical Quality



Quality being overestimated in original file

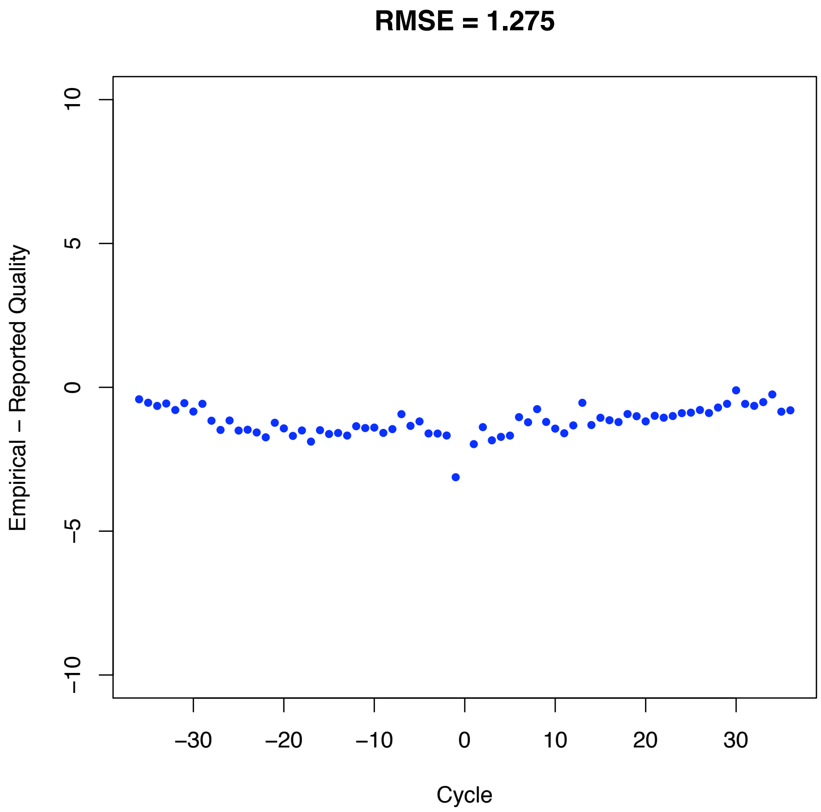
Before Recalibration

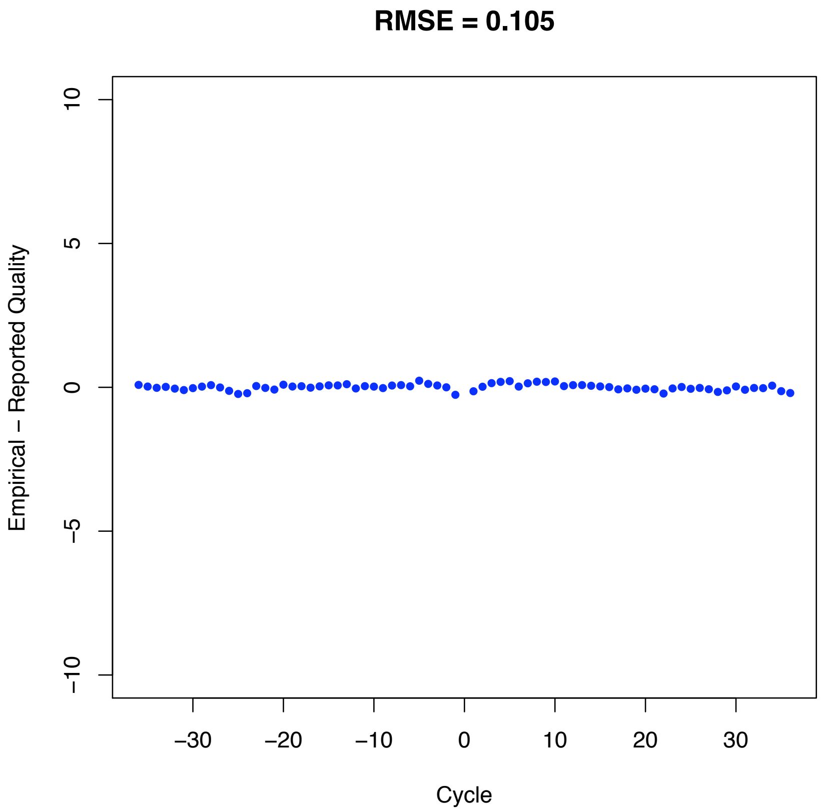


RMSE greatly reduced after recalibration

After Recalibration

Residual Error by Machine Cycle

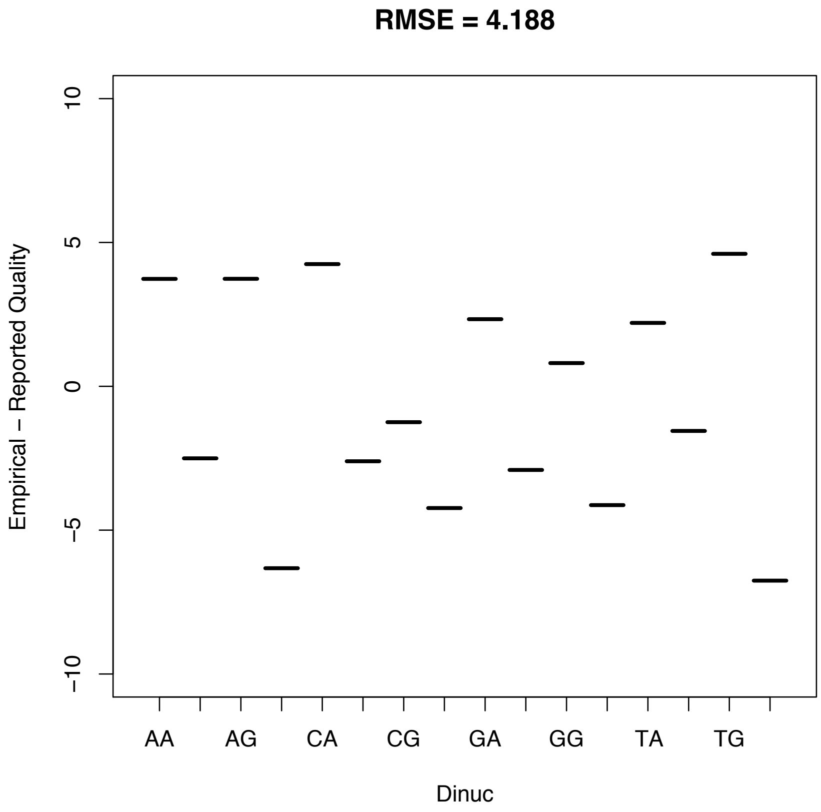




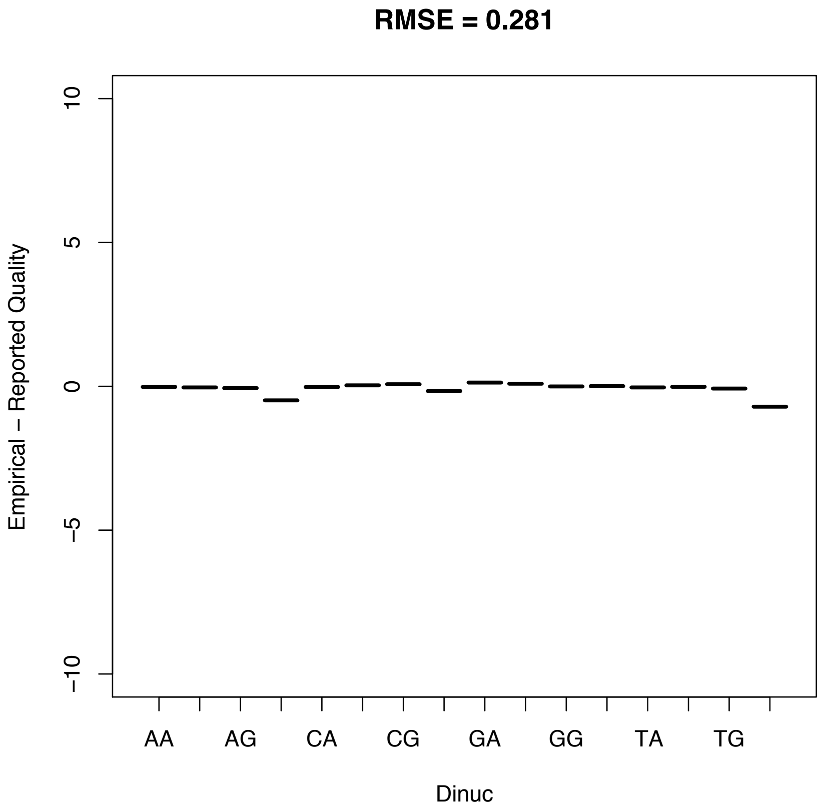
Before Recalibration

After Recalibration

Residual Error by Dinucleotide

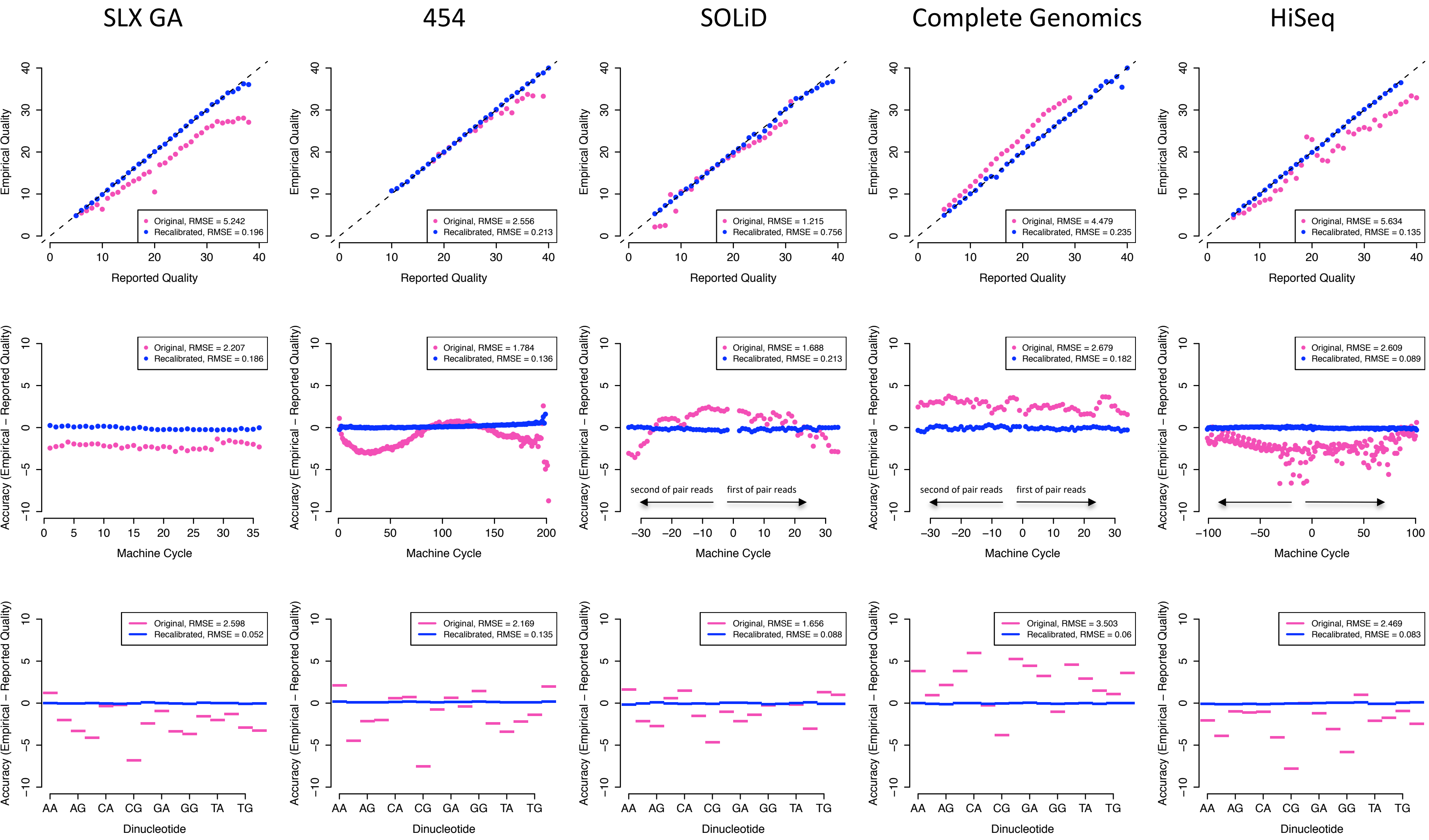


Before Recalibration



After Recalibration

Results by Platform



Calling Variants

- Distinguish real variants from errors

- Tradeoff between sensitivity and specificity

Simple Pileup Methods

- Count calls at each site
  - Compare each call to reference or majority
- Are there more of a base than expected?
  - Based on random error model
  - Note that all platforms have non-random error
- Most appropriate for pooled data
- With explicit genotypes, more information is available

Bayesian Methods

- Assign calls to specific genotypes
  - Requires a ploidy model
- Compute the probability of each genotype given the data
  - Accounts for error probabilities
  - Also considers allele balance, priors on variation
- Make better use of all available data

Population Aware Calling

- Real variation has expected distribution between individuals
- Variants observed at high frequency in a population more likely real in a given sample
- Variants seen with skewed allele distributions are more likely artifact
    - Always heterozygous or homozygous
    - Out of Hardy-Weinberg equilibrium

Haplotype Aware Callers

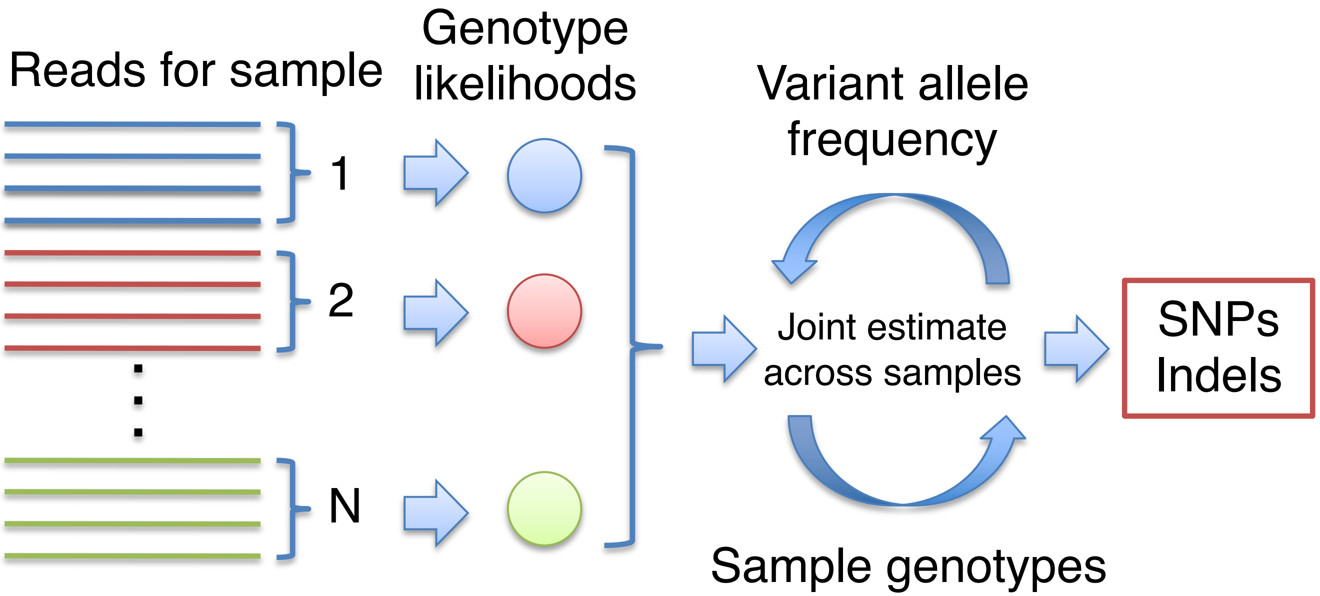
- Consider population data at multiple loci
- Essentially imputing variants during calling
  - Reduce likelihood of calls not in linkage
  - Fill in missed variants predicted by linkage
- Require extensive population data for training

Genotype Likelihood per Sample

$$\overbrace{L(G|D)}^{\text{Likelihood for the genotype}} = \overbrace{P(G)}^{\text{Prior for the genotype}}\overbrace{P(D|G)}^{\text{Likelihood of the data given the genotype}} = \overbrace{\prod_{f_i \in \{fragments\}} P(f_i|G)}^{\text{Inference from reads and bases to sequenced DNA fragment to chromosomes}}$$

- Genotype likelihoods describe the probability of the reads for each genotype (AA, AC, …, GT, TT) at each locus
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only “good reads and bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS

Multi-sample Calling



Simultaneous estimation of:

- Allele frequency (AF) spectrum: Pr{AF = i | D}
- The prob. that a variant exists: Pr{AF > 0 | D}
- Assignment of genotypes to each sample

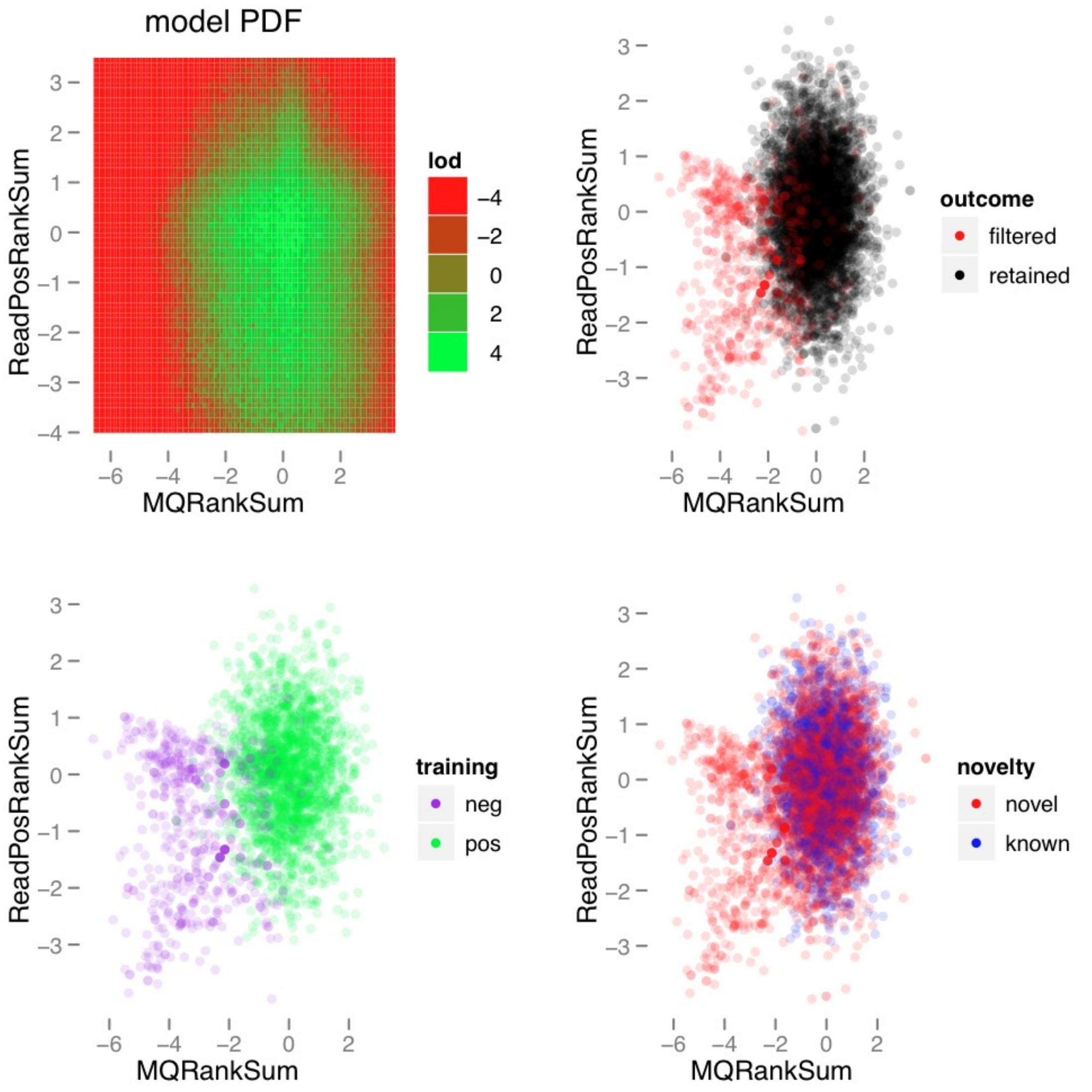
Filtering Variants

- Even the most advanced variant calling models can be fooled by systematic error
- Sensitivity comes from the caller
- Precision comes from filtering
- Certain artifact patterns can be recognized
  - Variant calls are biased to one strand, ends of reads, low quality bases, low mapping quality, etc.
  - Variant positions have unusual read depth

Variant Quality Score Recalibration

- Similar to base quality recalibration
- Consider factors used in filtering variants
- Compare known variant sites to all sites
- Build models of the probability of a variant matching the profile of known variants
- Dynamic determination of filtering cutoffs
- Requires large data set, known variant set

Example of Quality Recalibration


model PDF
ReadPosRankSum
MQRankSum
lod
outcome
filtered
retained
training
neg
pos
novelty
novel
known