KBBQ: A reference-free method for base quality score recalibration

Adam J Orr [1,2] @AdamJOrr, Reed A Cartwright [1,2] @MinionLab

[1]School of Life Sciences, Arizona State University
[2]Biodesign Institute, Arizona State University

Introduction

- Illumina sequencing reads contain errors
- Errors make mutation detection difficult
- Quality scores represent $P(error)$ on a phred scale.

$$P(error) = 10^{\frac{-Q}{10}}$$

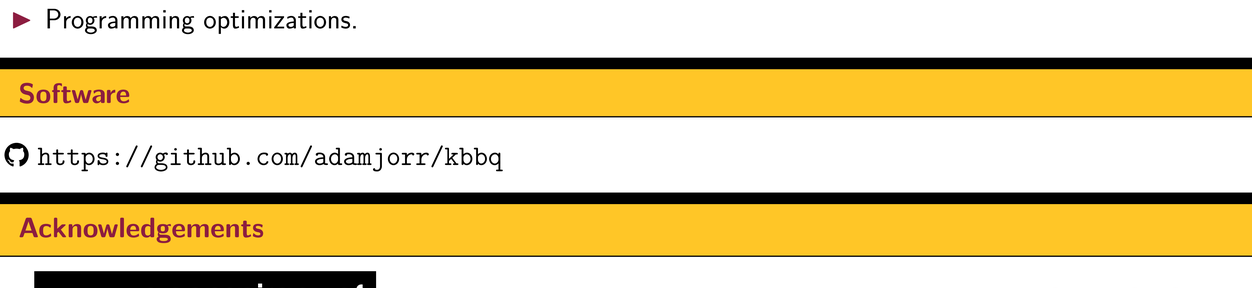
$$Q = -10 \log_{10} P(error)$$

- Quality scores are sometimes binned to reduce file size

Score	Bin	$P(error)$
0	N	1.0000
10	15	0.1000
20	27	0.0100
30	33	0.0010
40	40	0.0001

Methods: Base Quality Score Recalibration - BQSR

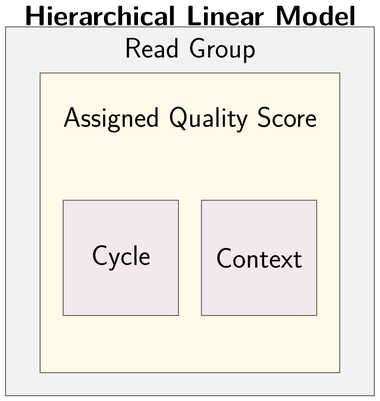
Base Quality Score Recalibration is a technique to improve calibration of the original quality scores. However, **BQSR normally requires a reference genome and a database of variable sites**. Our reference-free method uses a method similar to the `lighter` (Song, Florea, and Langmead 2014) error corrector to find erroneous bases rather than comparing the sequence to a reference. You can use your favorite error corrector and use those corrections as input to `kbbq`.



Methods: Hierarchical Linear Model

The recalibration uses a hierarchical linear model to predict the true probability of error given a set of covariates. Each covariate causes the predicted quality score to shift up or down from the predicted score one level above it in the hierarchy. The relevant covariates are:

- Read Group
- Original Assigned Quality Score
- Position in read and whether read is forward or reverse (Cycle)
- Base called and the prior base call (Context)



Methods: Evaluation

- DNA from CHM1 and CHM13 human hydatidiform mole cell lines was mixed to generate a synthetic diploid dataset for benchmarking (Li et al. 2018).
- We analyzed Illumina reads aligned on Chromosome 1 and compared recalibrated quality scores generated using our method and the standard GATK method.

Next Steps

- Test performance of GATK BQSR with varying levels of false positives.
- Linear model improvements
- Programming optimizations.

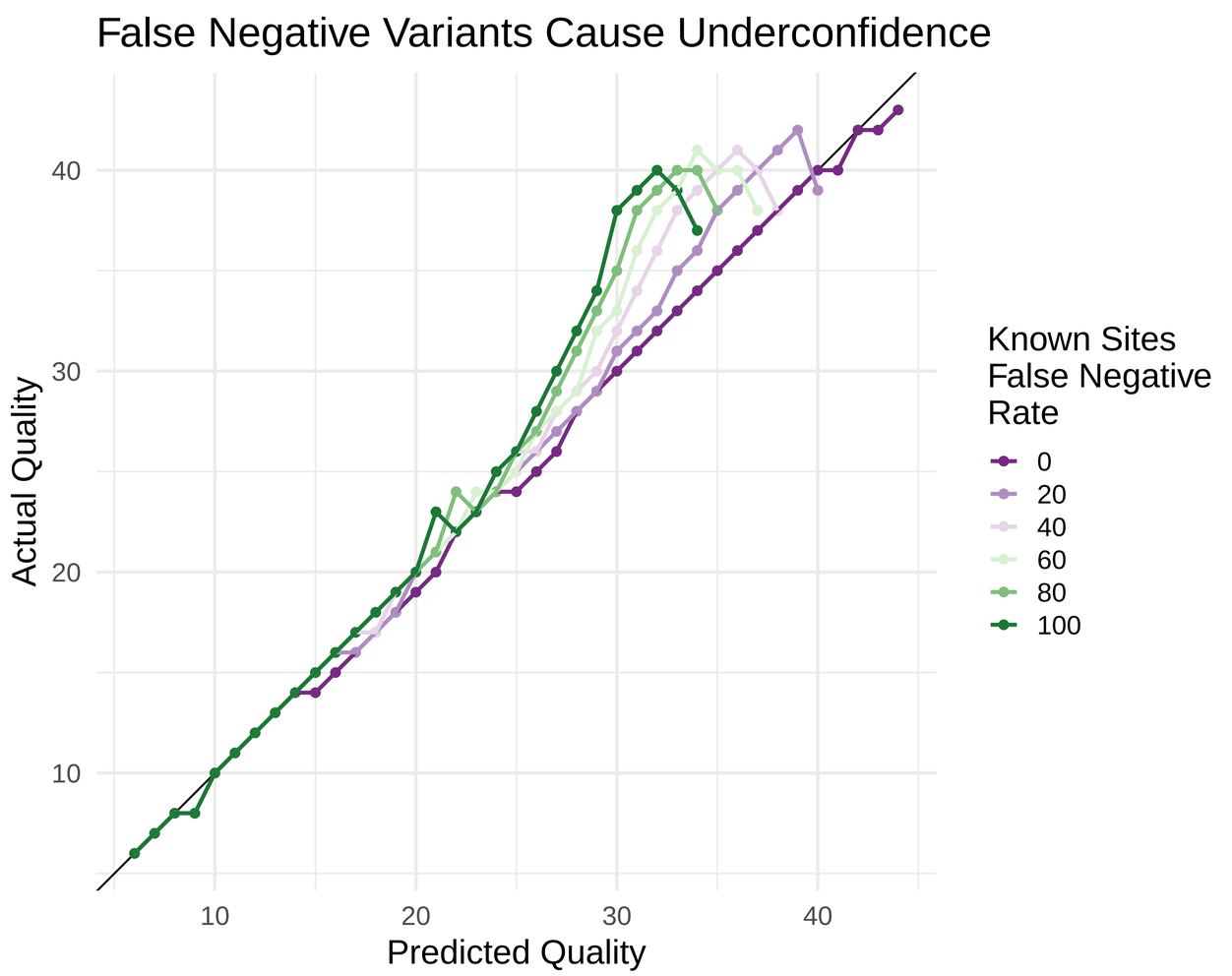
Software

`https://github.com/adamjorr/kbbq`

Acknowledgements

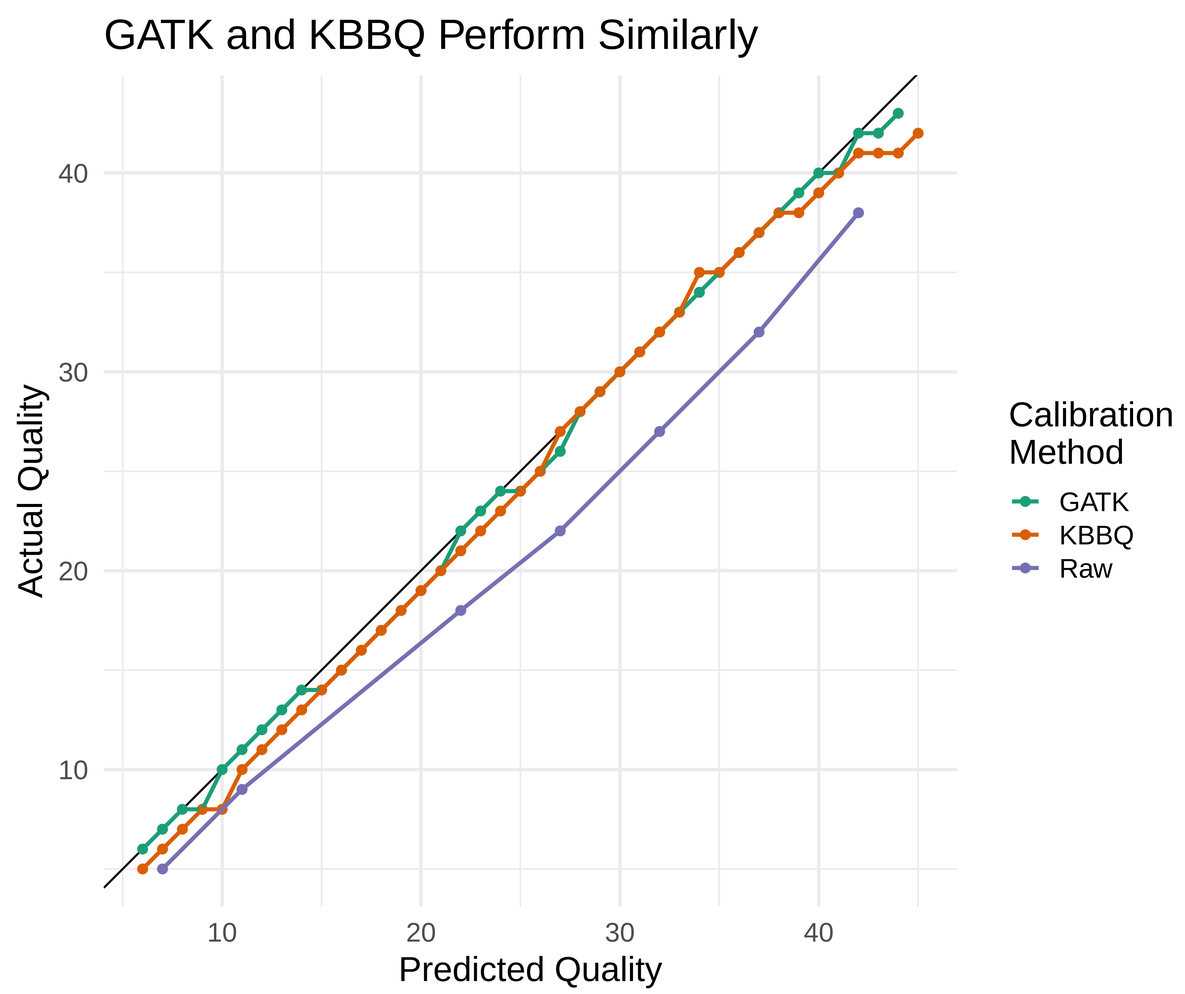
a conspiracy of unions cartwrig.ht/lab/
ASU School of Life Sciences Arizona State University
ASU Biodesign Institute Arizona State University

Results: GATK BQSR is vulnerable to false negatives in the known sites input.



KBBQ doesn't require *a priori* knowledge of variable sites, so is unaffected by misspecification of known sites.

Results: Similar performance, no reference or variants required



Both methods improve calibration and resolution, while KBBQ has relaxed input requirements.