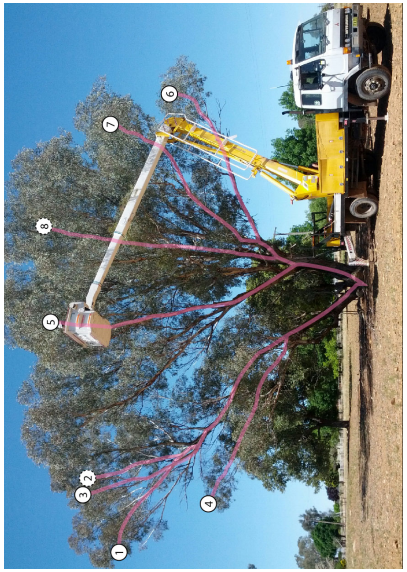
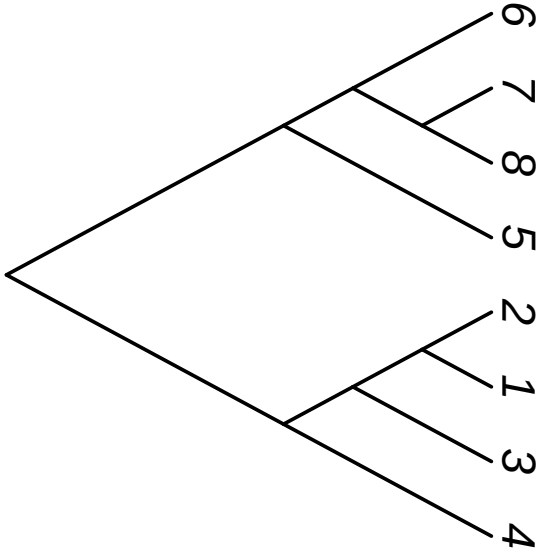
Methods for sensitive genotyping in nonmodel organisms

Adam Orr @AdamJOrr

12/1/19





Chapter 2 - Detecting Somatic Mutations in a Non-model Organism

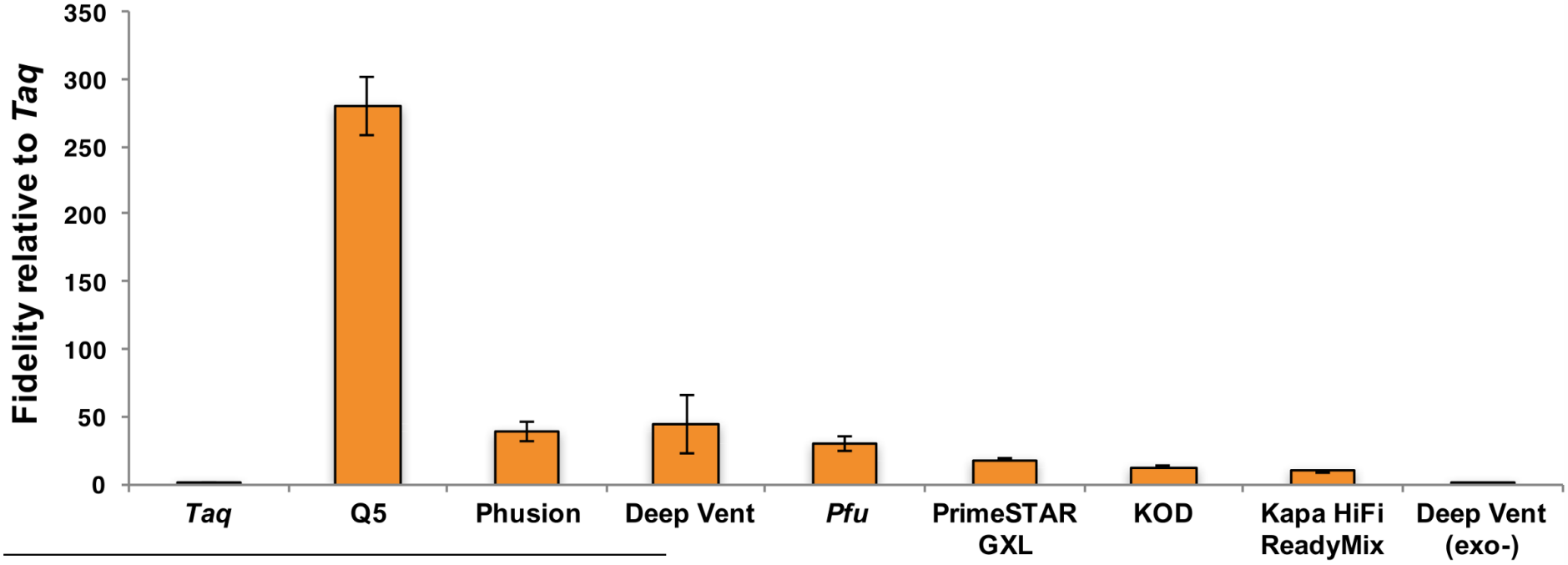
- Somatic mutations are difficult to detect
- Finding mutations in non-model organisms is also difficult
- What are the biggest challenges for sensitively detecting mutations in non-model organisms and how can we overcome them?

Why are somatic mutations difficult to detect?

Mutations are very rare, but sequencing errors are very common.

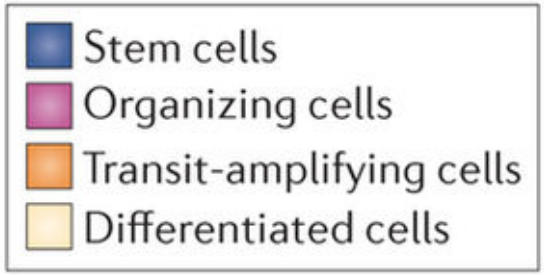
Sequencing error alone is $\sim 10^{-2}$ while mutation rate after error-checking is $\sim 10^{-9}$

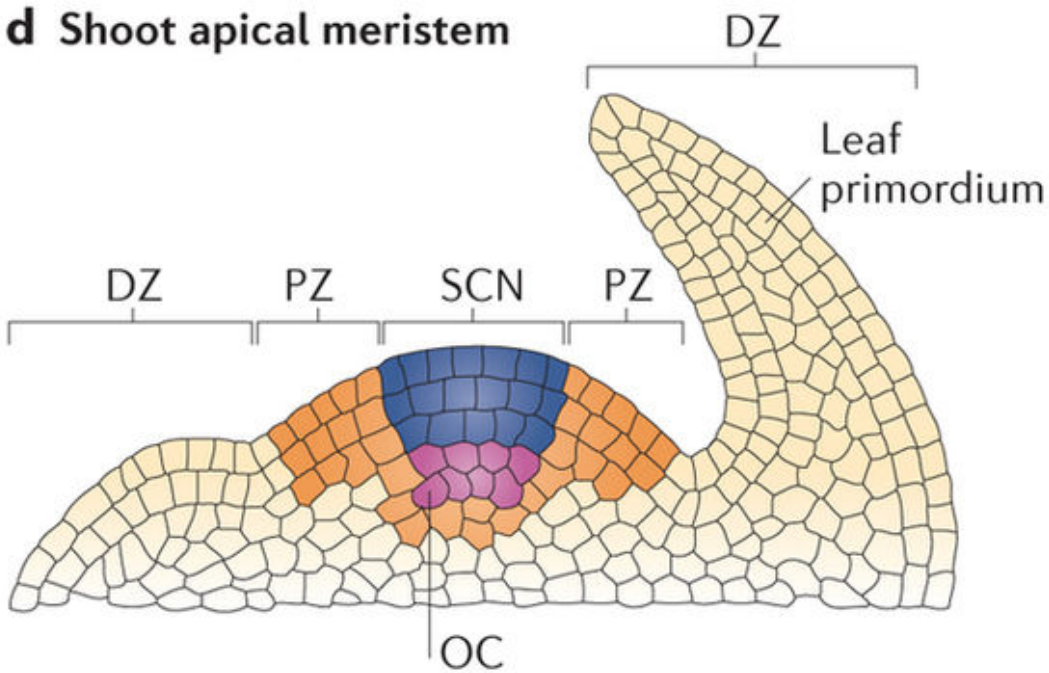
- Errors accumulate during PCR prior to sequencing - then propagate.
- *Taq* $\sim 10^{-4}$
- Technical error from sequencer



[0] Potapov V, Ong JL (2017) Examining Sources of Error in PCR by Single-Molecule Sequencing

Plants Grow Directionally





- The genetic structure of the plant *should* mirror its physical structure.

[0]Heidstra & Sabatini (2014) Plant and animal stem cells: similar yet different.

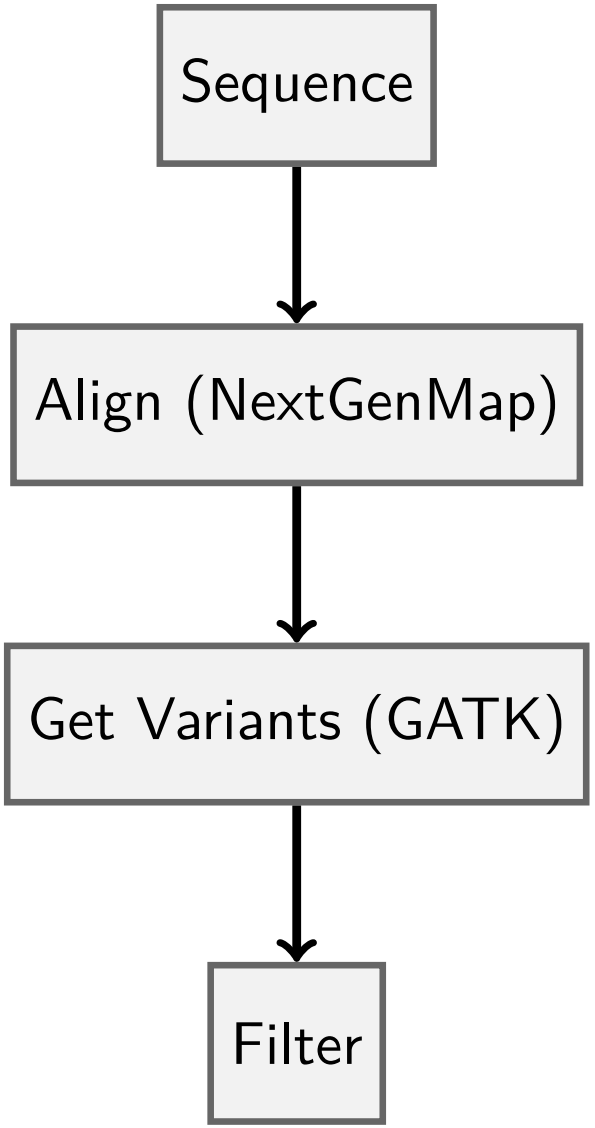
A Genetic Mosaic

- Edwards identified as mosaic in 1993[1]
- Sheep pen in Yeoval, New South Wales
- Differential oil production gives protection from Christmas beetles

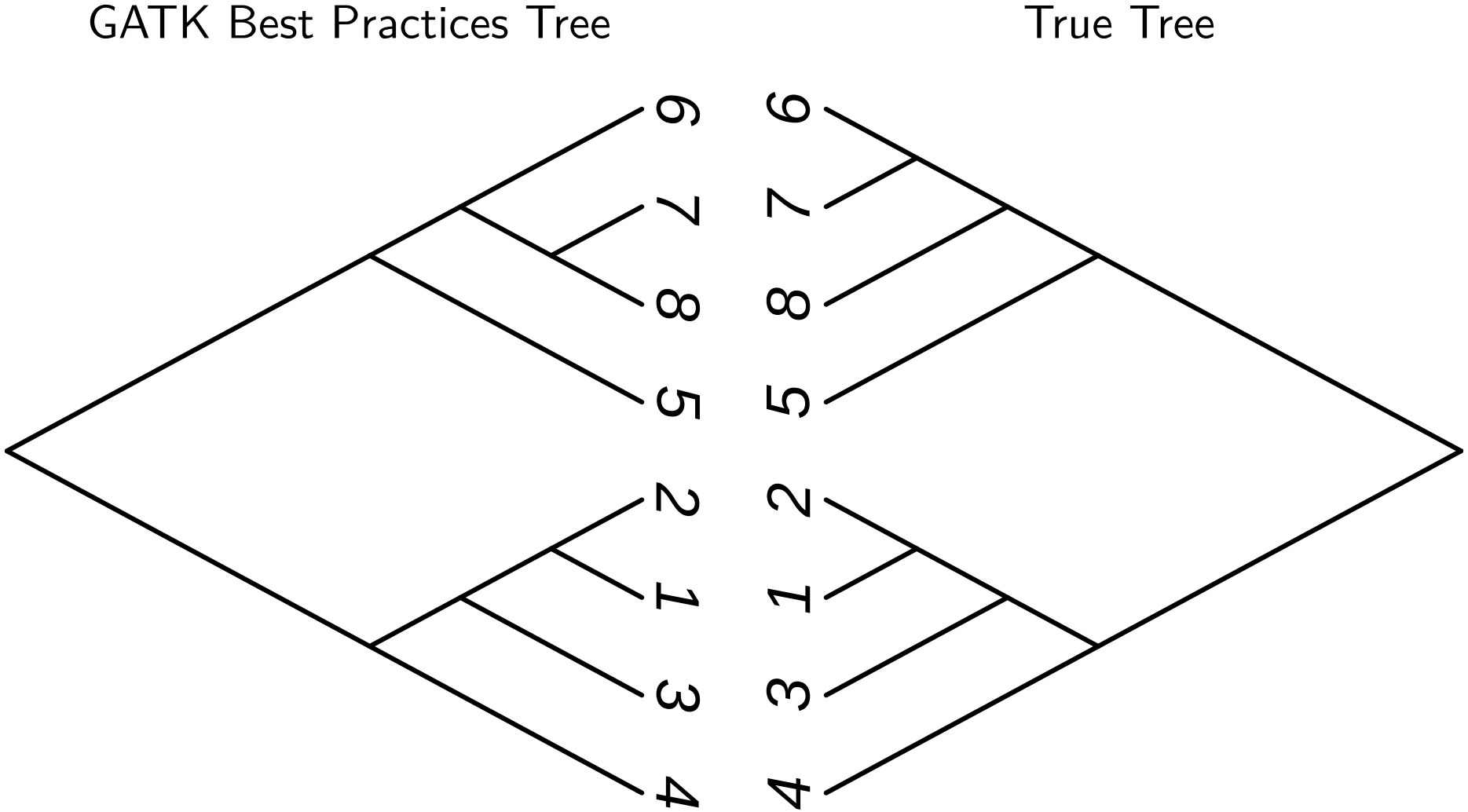
[1] *Edwards PB, Wanjura WJ, Brown WV. Oecologia 1993, 95:551–557.*

Study Methodology

- Sequence 8 samples in triplicate
- ∼10X coverage for each replicate
- Align sequence to genome of *Eucalyptus grandis*
- Use replicates to remove false positives



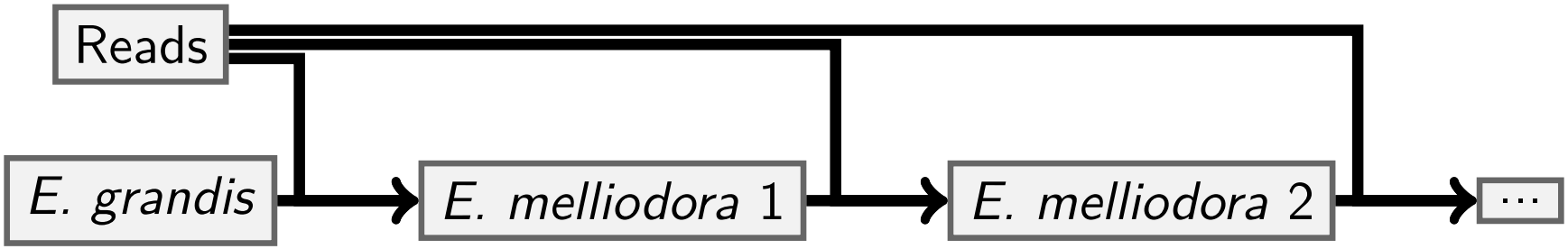
Mutation Pattern Approximately Matches Tree Structure



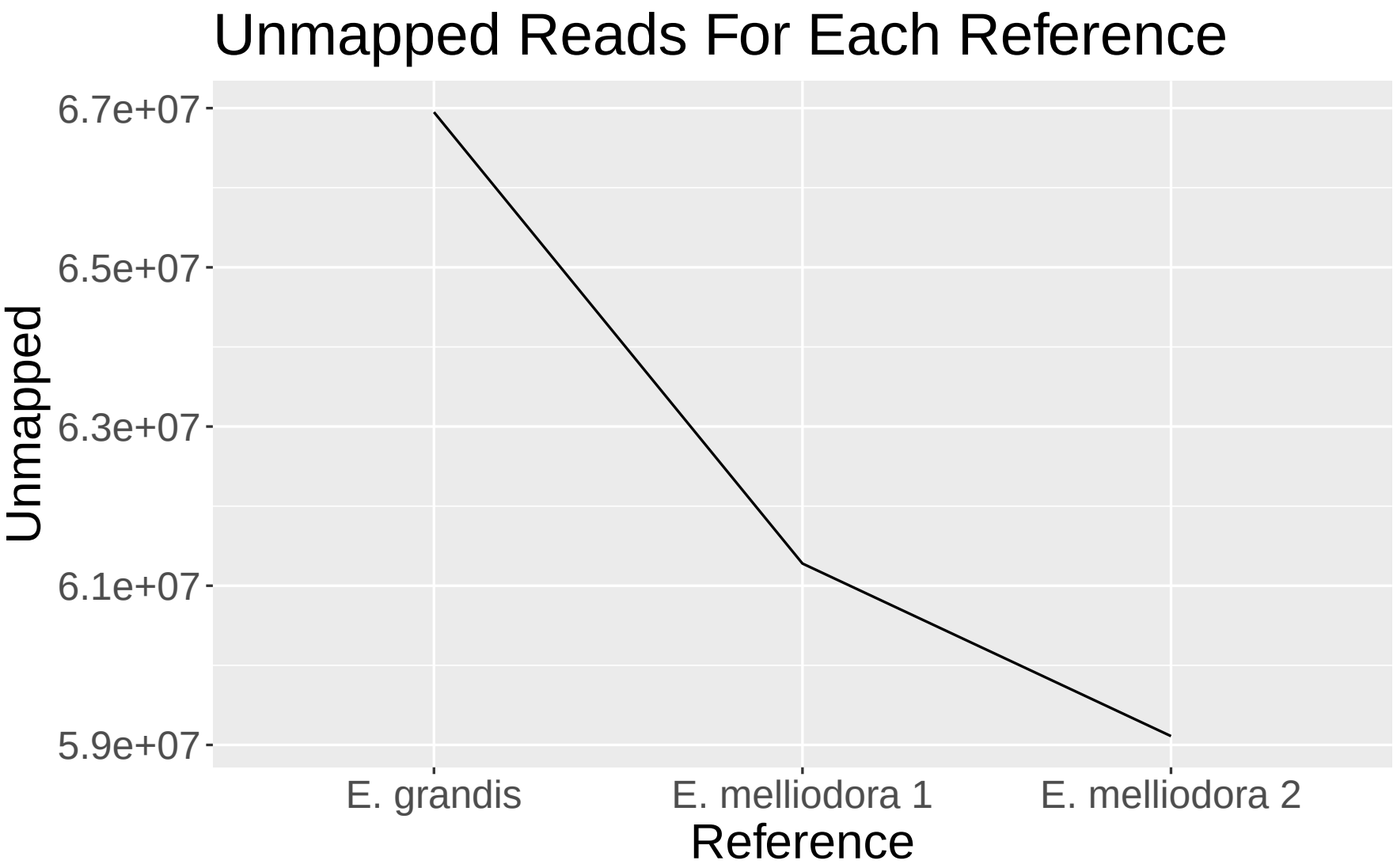
Most Reads Are Not Mapped to the *E. grandis* Reference

Approximating a Genome

Use *E. melliodora* genome as a starting place, then generate a new reference and map to that reference.



Our New Reference Has Fewer Unmapped Reads

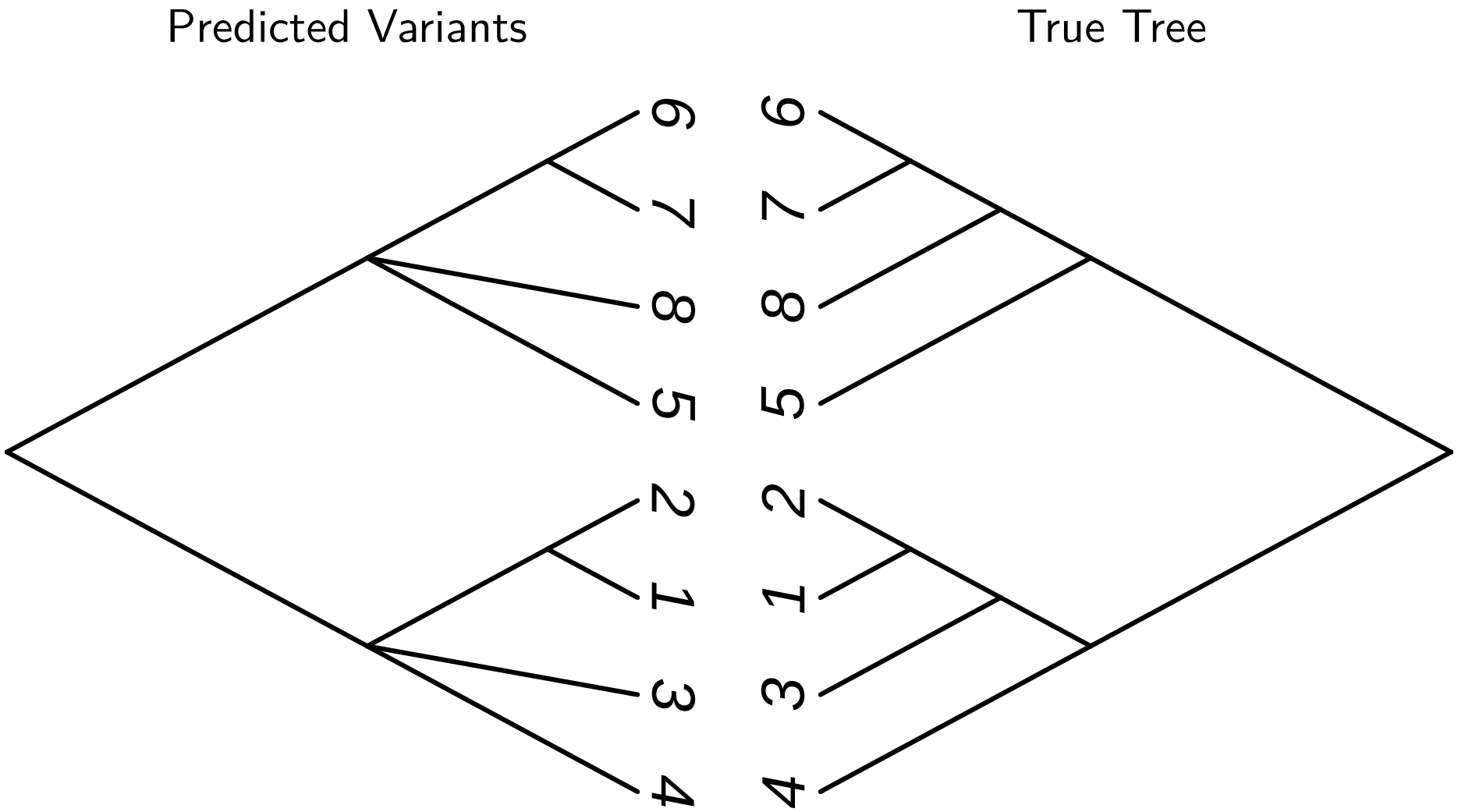


Filtering Variants

Remove variants likely from alignment errors:

- at sites with excessive depth (>500).
- with excessive levels of heterozygosity.
- within 50 bases of an indel.
- in repeat regions

Filtering and Reference Refinement Improve Tree Topology



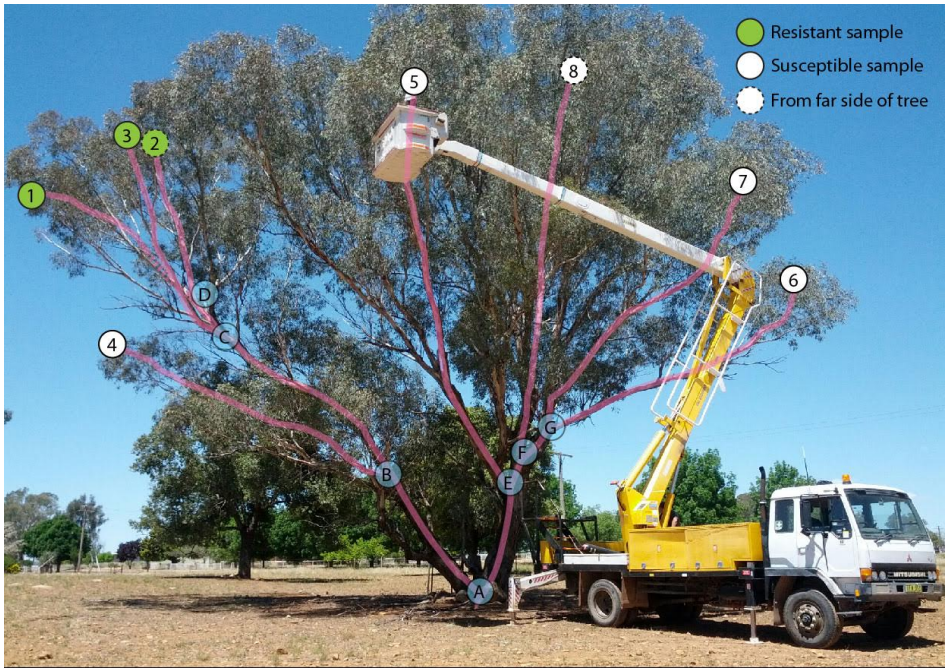
Using Tree Topology Gives Higher Recall Rate

- Thus, it's reasonable to assume the physical topology when inferring mutations
- *DeNovoGear* is a variant-calling method that uses information in the tree topology to call variants.
- By simulation, we introduced 14000 mutations on the tree

GATK	*DeNovoGear*
3859 mutations	4193 mutations
27%	30%

Mutation Rates

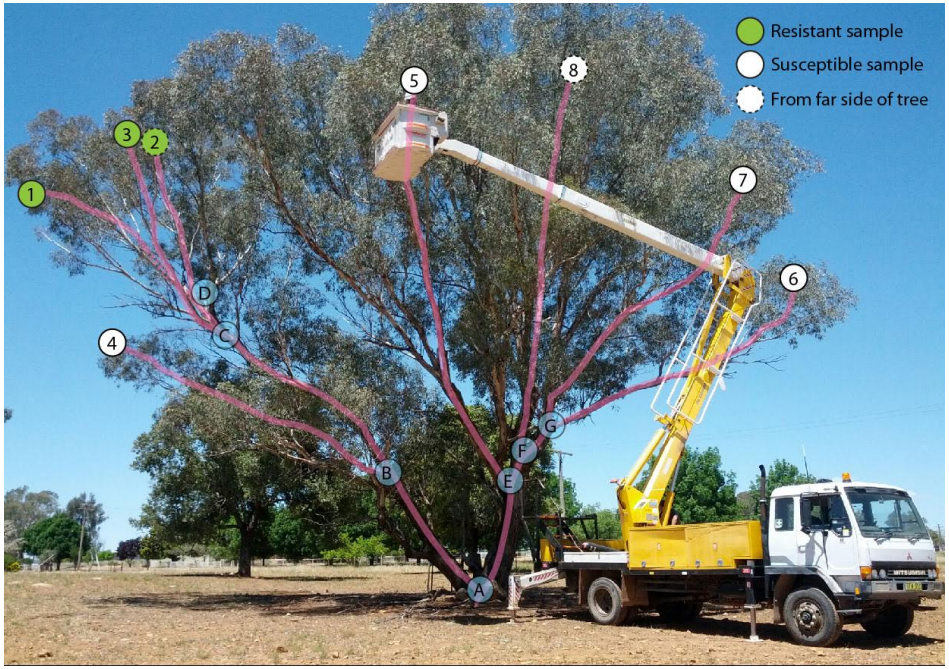
- Detected 90 mutations.
- 20 mutations in genes.
- Estimated recall of $\sim 30\%$.
- $90 \times \frac{1}{.3} = 300$ mutations.
- ~ 3.3 mutations per meter of length
- 2.7×10^{-9} mutations per base per meter
- Somatic mutations account for ~ 55 mutations per leaf tip.



Model Parameters

We studied *one* individual, but we can make conjectures about the population.

- The average height of a eucalypt is 22.5 M
- Mutation rate per base, per generation is 6.2×10^{-8}
- We estimated $\theta = 0.025$
- Since $\theta = 4N_e\mu$, $N_e = 102,000$



This per-generation rate is $\sim 10\times$ larger than *Arabidopsis*, but *Eucalyptus* is $100\times$ larger.

Chapter 3 - Base Quality Score Recalibration in Non-model Organisms

Errors make variant calling difficult - but we can predict them.

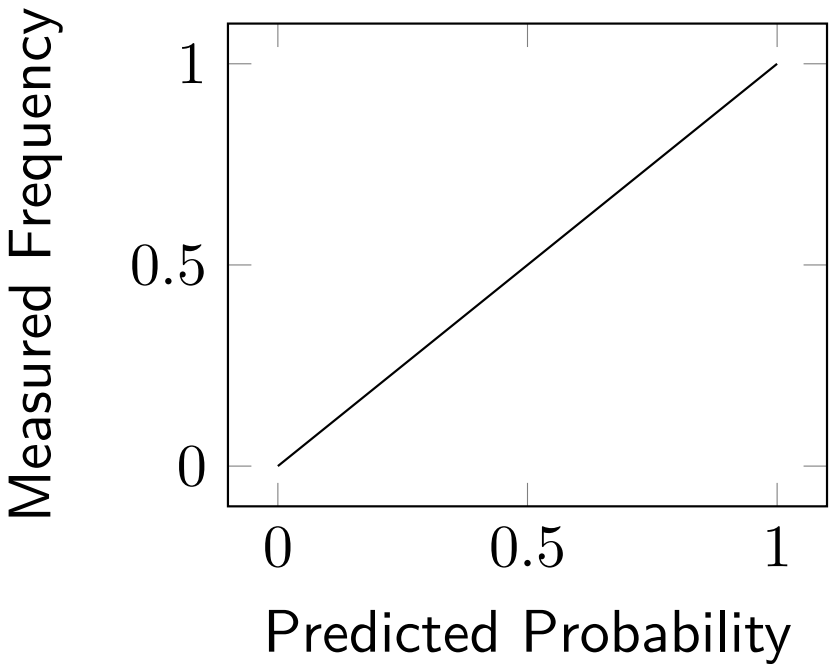
- FASTQ format data has a quality score
- Quality scores represent $P(error)$ on a phred scale.

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

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

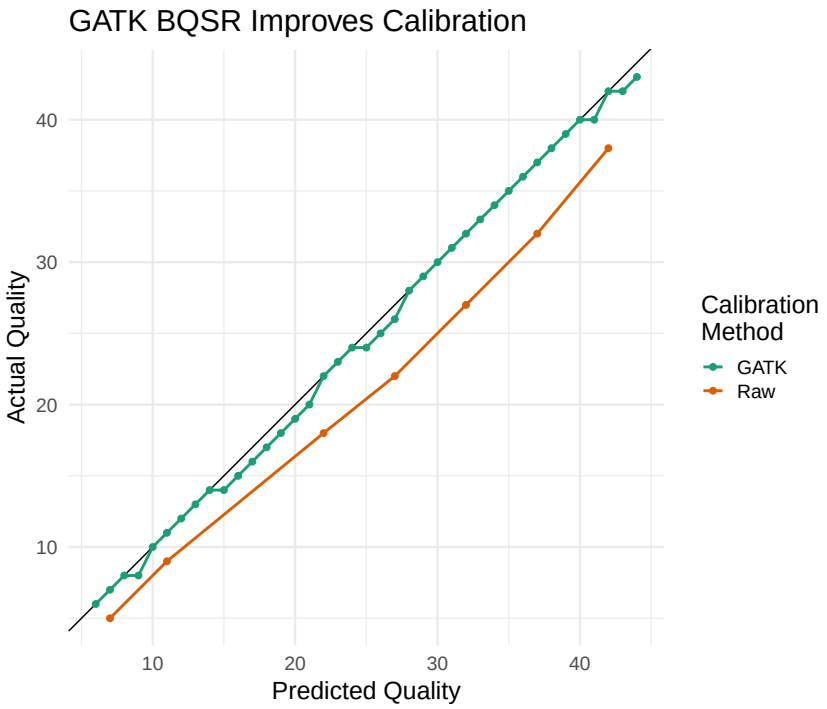
Quality scores are predictions

- A quality score is a **prediction** about whether a base call is correct.
- Predictions are said to be **calibrated** if the predicted event occurs as often as predicted.
- The weather forecast contains a **prediction** about whether it will rain.
- If it rains on a day with a 30% chance of rain, what does that mean?

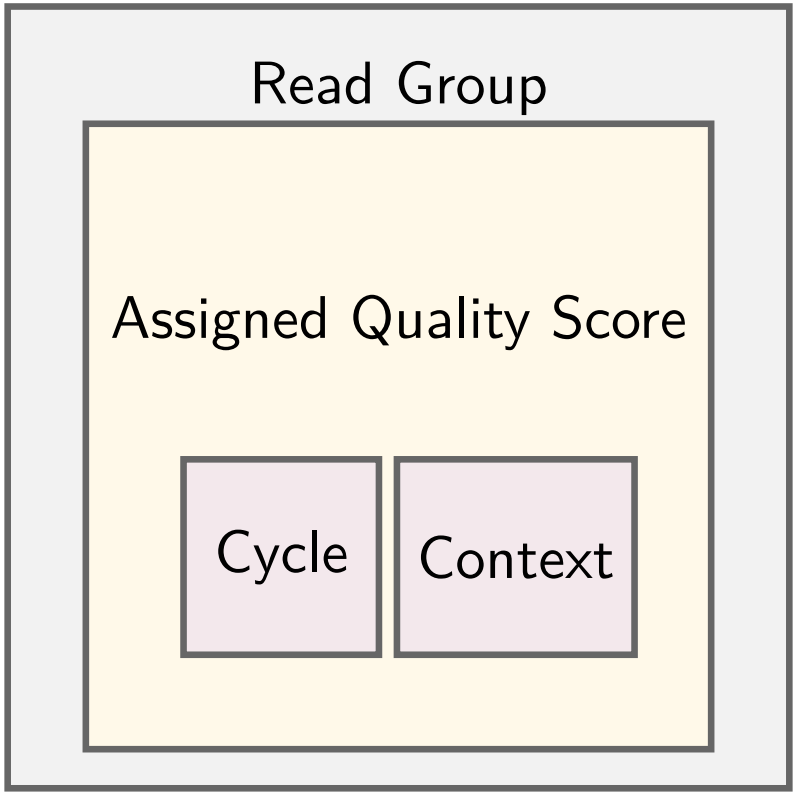


Quality scores aren't well-calibrated

- If quality scores *were* well-calibrated, it would be easier to identify errors
- Base Quality Score Recalibration can be done to fix calibration issues.
- Current GATK method for BQSR require a database of variable sites in your data then assumes mismatches at nonvariable sites are errors.



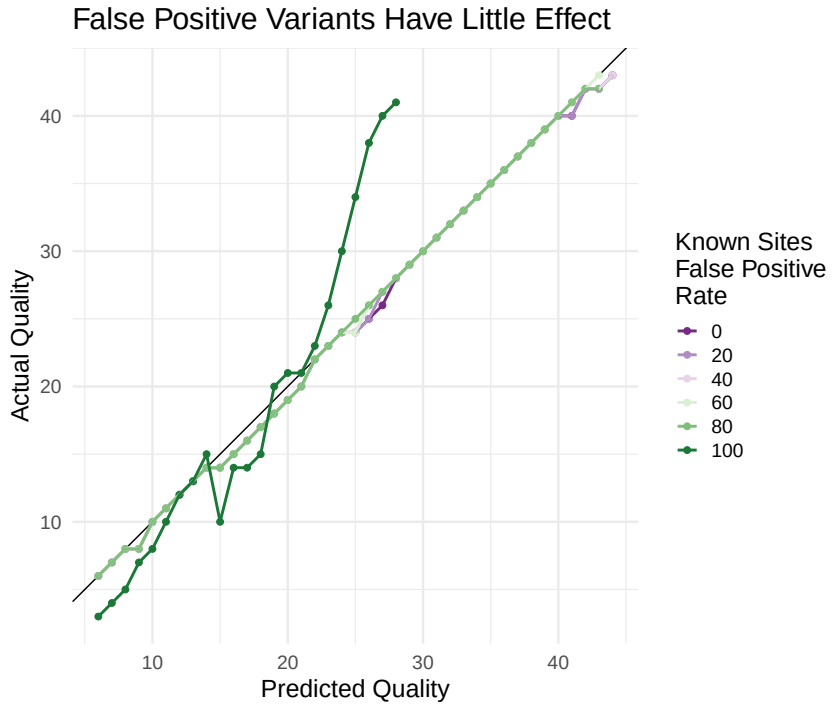
GATK BQSR uses a linear model to determine how much to adjust each quality score

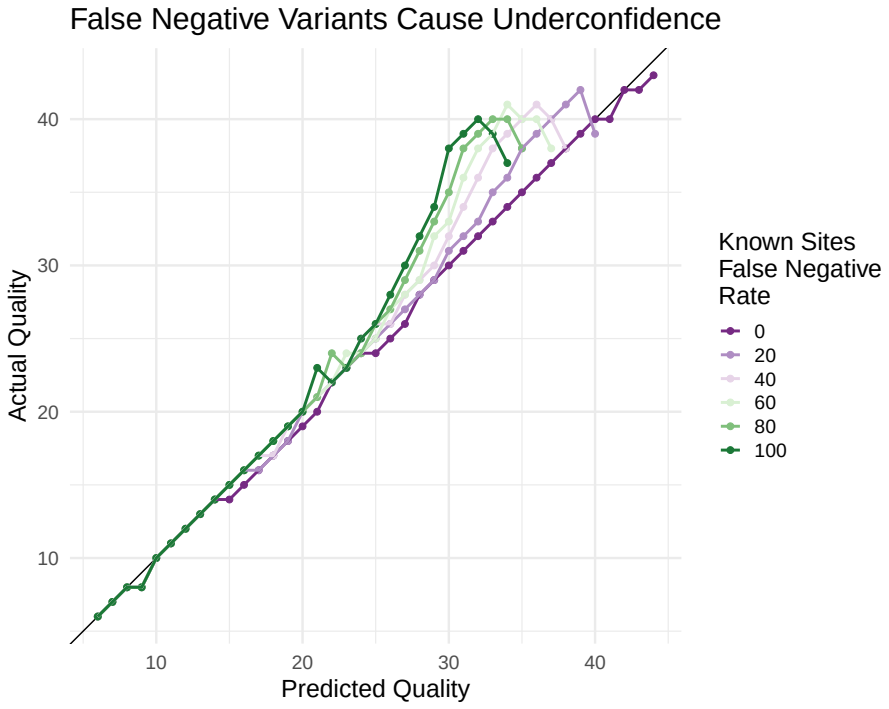


Questions

- How effective is BQSR, particularly if the reference and database of known variation are not good?
- What is the impact of BQSR on downstream variant calls?

BQSR is vulnerable to false negatives in the database of variable sites





Alternative approaches get around using a database of variable sites

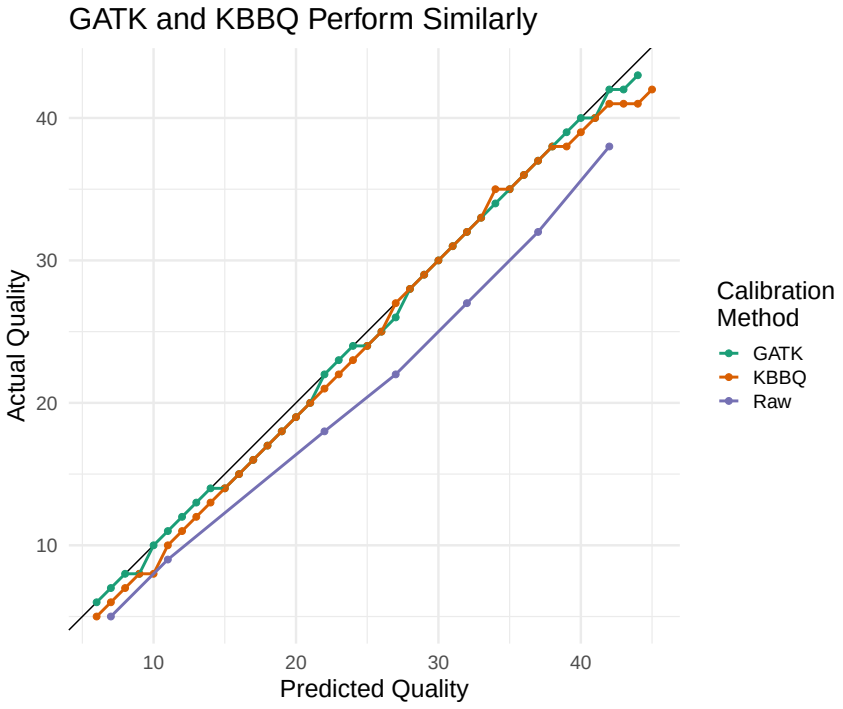
- Lacer uses singular value decomposition
- ReQON limits the number of errors there can be at a site
- Synthetic spike-ins

Error correctors can find some errors without a reference

- Error correction methods exist that use k-mers to identify errors rather than an alignment and reference.
- Most error correctors don't update quality scores; `Lighter` optionally updates quality scores of corrections to a set maximum value but this doesn't materially affect the calibration.

K-mer Based Base Quality score recalibration

- Combining error correction and BQSR is effective
- Method implemented in `kbbq` software



Future Plans

Calculate Brier scores for each calibration to have a quantitative comparison.

Evaluate downstream impact on quality of variant calls

- F-score of returned calls and AUC of variant quality
- How are reported annotations influenced by quality scores?
- Run `kbbq` on the Eucalyptus data and find how that changes the distribution of quality scores and the number of detected variants.

Chapter 4 - Base Quality Score Recalibration in Long Reads

- PacBio hi-fi reads are a consensus of many sub-reads; do these consensus reads have meaningful quality scores?
- Nanopore reads are said to be well-calibrated, but in Illumina different runs can produce different error profiles; is this the same in Nanopore?
- Genome In A Bottle has Illumina, PacBio, and Nanopore sequencing of the same sample.

How well-calibrated are long reads?

- Check data for calibration; if it's not well calibrated, try using GATK/`kbbq` and see if it works.
- For PacBio data can we use features of the subreads to make more accurate inferences?
- What is the best way to represent the read length covariate? Does it still matter?
- Nanopore uses 5 or 6bp basecalling models, so the context covariate in this case should be that long.
- Methylation?
- Does logistic regression make more sense for this data?

How does calibration improve variant calls in long reads?

- The biggest reason to use BQSR is to accurately classify bases as errors. If there is any reason to use it at all, it should be evident in noisy long read data!

Acknowledgements

- Advisor: Reed Cartwright @MinionLab
- Robert Lanfear, Australian National University @RobLanfear

Pipeline: https://github.com/adamjorr/somatic-variation
KBBQ: https://github.com/adamjorr/kbbq
Talk: https://github.com/adamjorr/talks

a conspiracy of μnions
cartwrig.ht/lab/

ASU BIODESIGN INSTITUTE
ARIZONA STATE UNIVERSITY

ASU SCHOOL OF Life Sciences
ARIZONA STATE UNIVERSITY

This work is supported by grants NIH R01-HG007178 and NSF DBI-1356548.