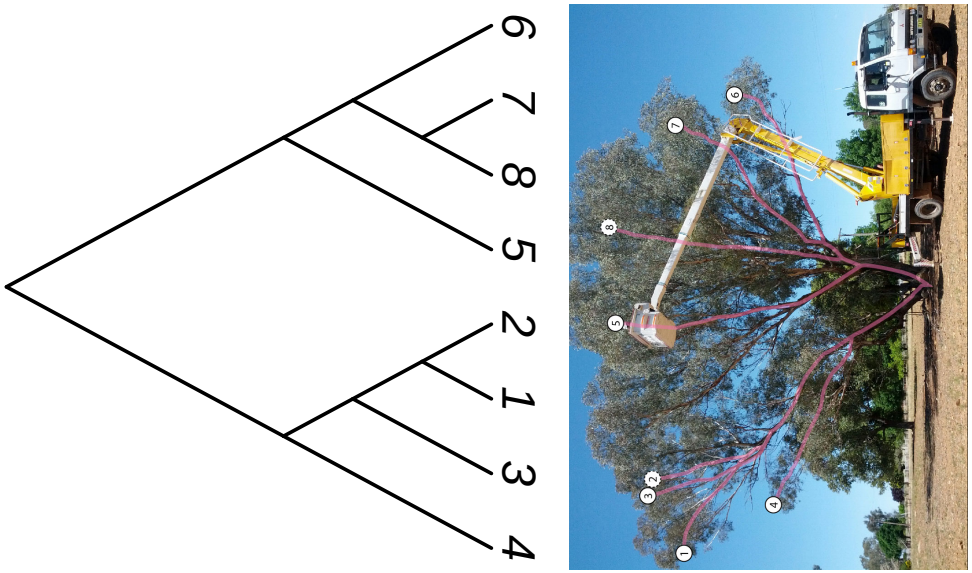
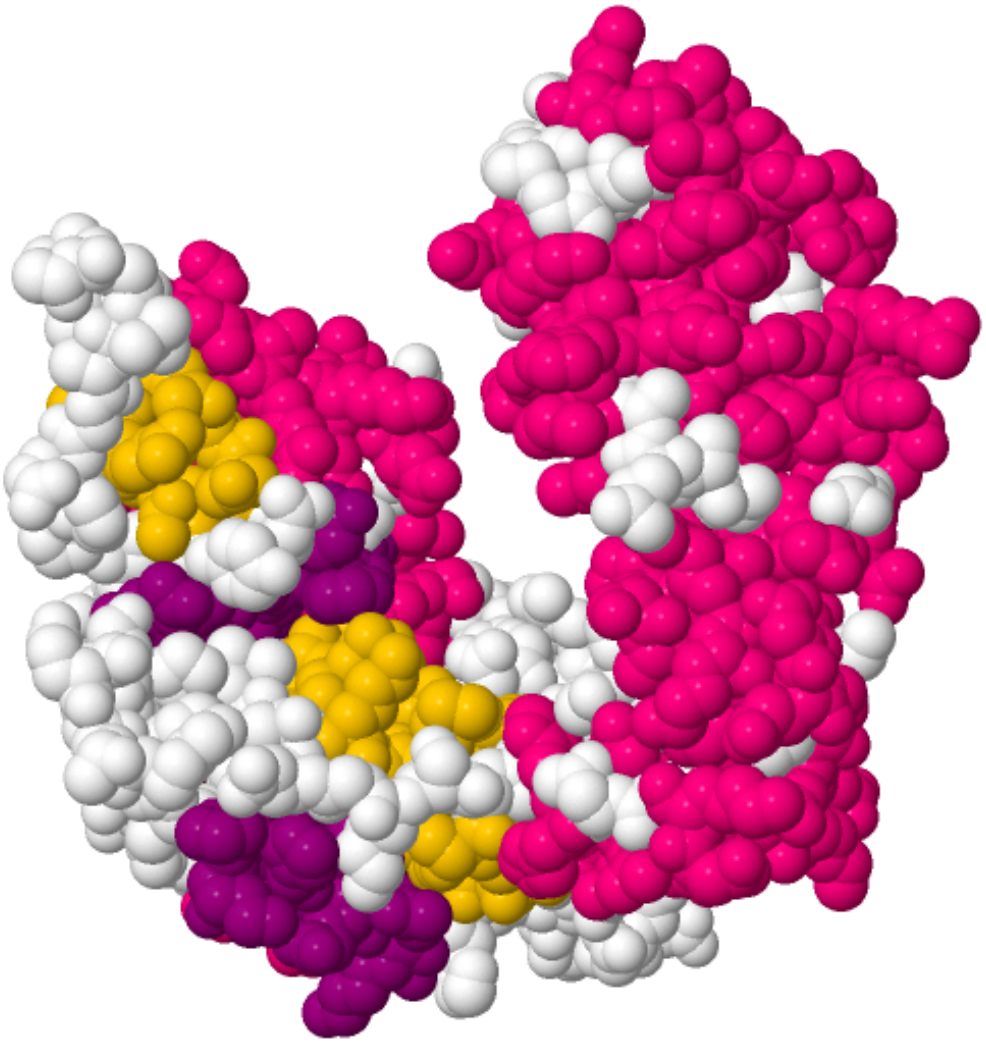
Methods for sensitive genotyping in nonmodel organisms

Adam Orr @AdamJOrr

5/17/19



Somatic Mutations Occur During Replication Even Without Exposure to Mutagens



- DNA Polymerase β
- Mutation rate $\sim 10^{-9}$

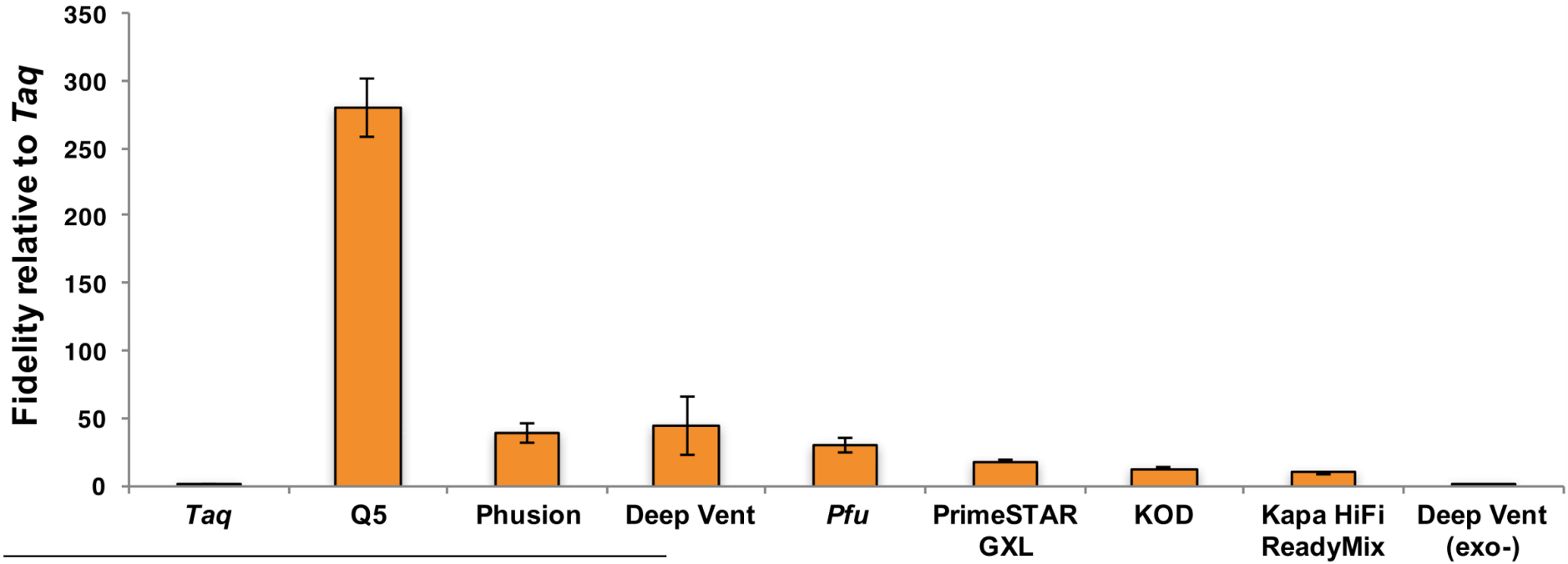
PDB: 7ICG

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

Why Care About Somatic Mutations?

Disease

- Cancer

Development

- Understanding the relationship between tissues

Agriculture

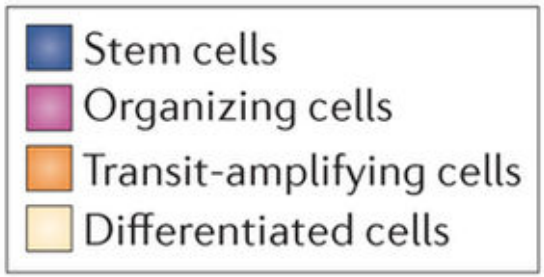
- Looking for interesting phenotypes in clonally reproducing species

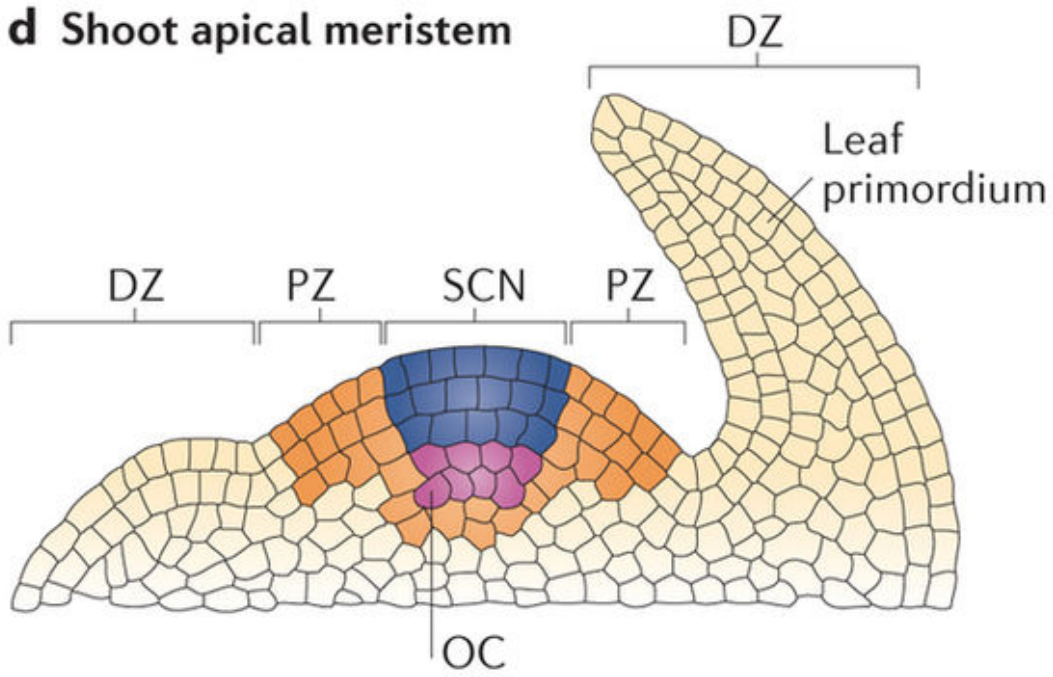
Evolution

- Determining the relationship between somatic and germline mutation rate

[0] https://commons.wikimedia.org/wiki/File:White_nectarine_and_cross_section02_edit.jpg

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.*

Project Goals

Is it possible?

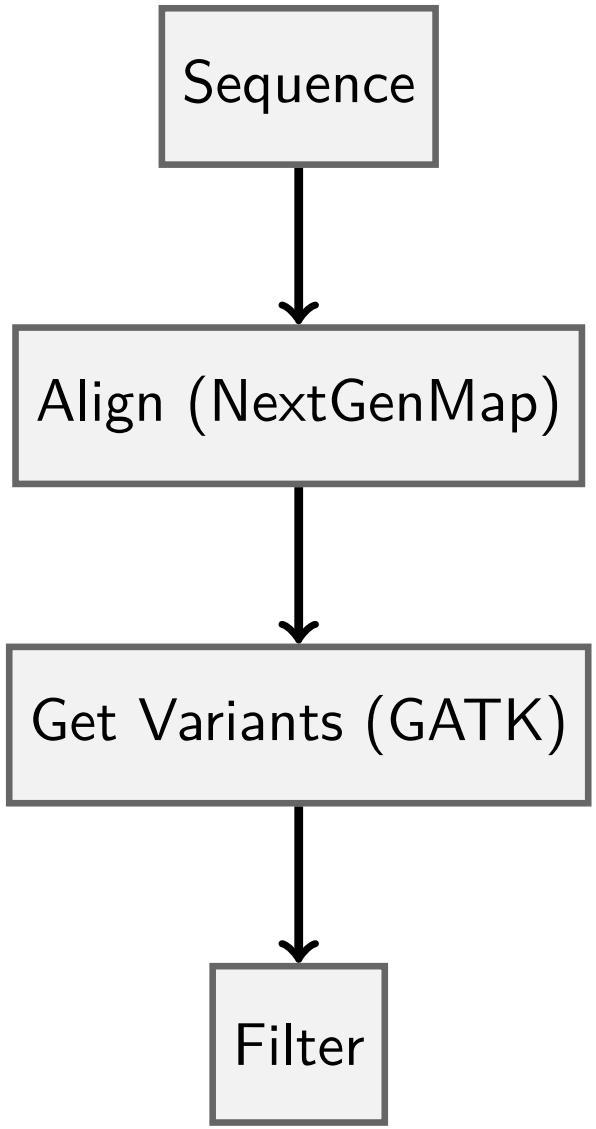
Can we detect mutations with sufficient accuracy? A good test would be to reconstruct its physical structure.

What does the lack of a segregated germline mean for evolution?

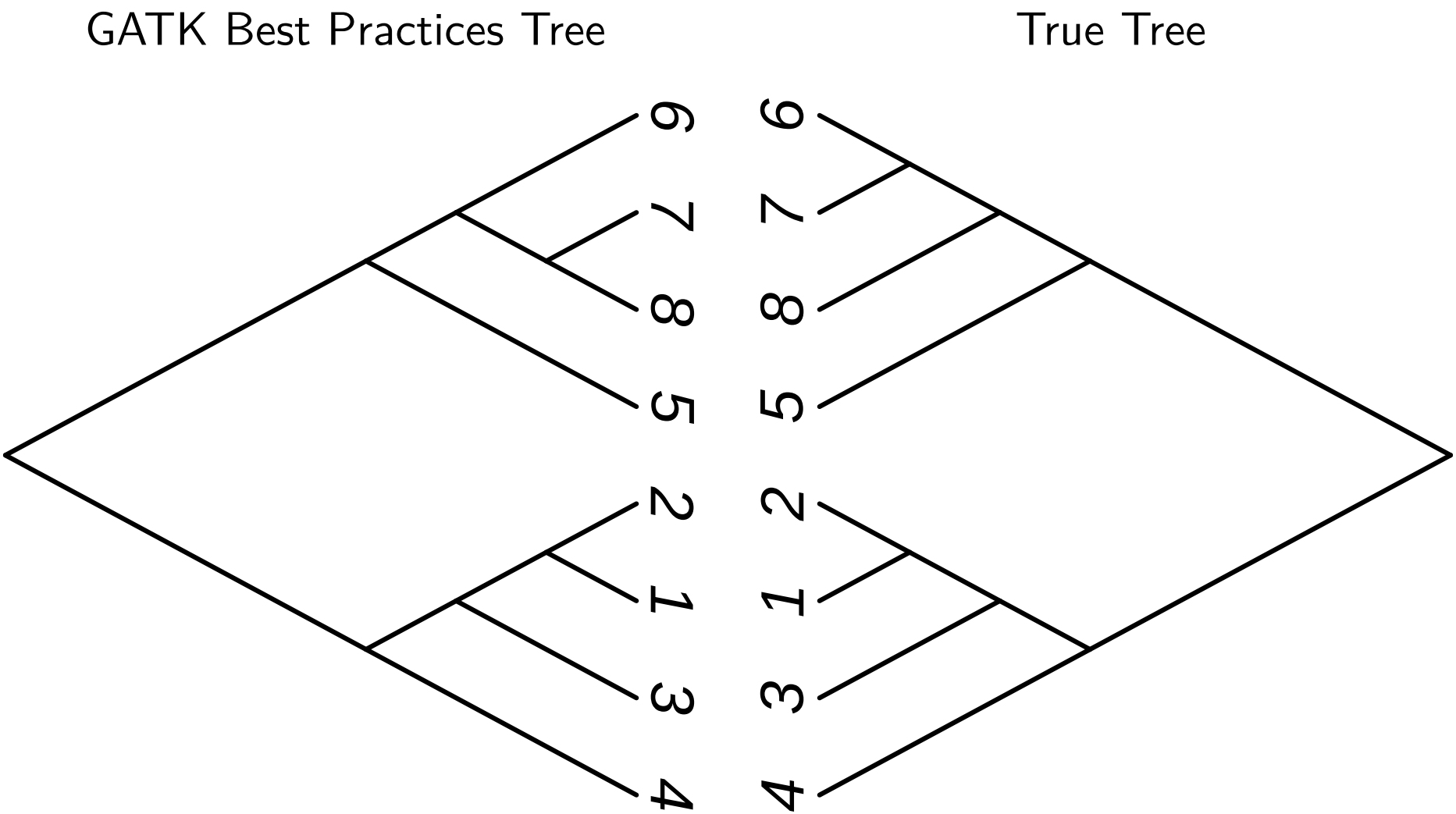
What's the mutation rate like? Is there evidence of hypermutation?

Study Methodology

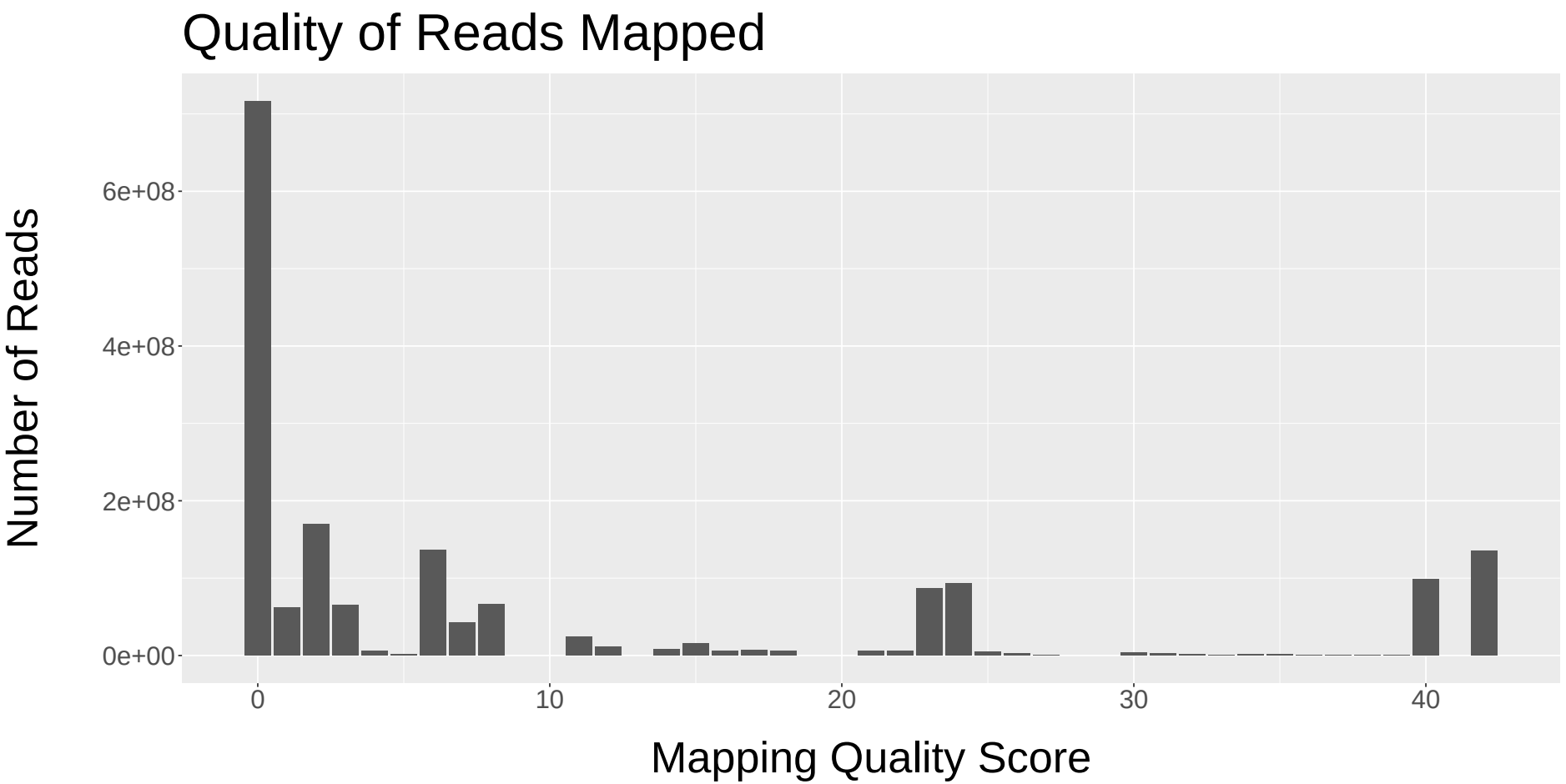
- Sequence 8 samples in triplicate
- ~10X coverage for each replicate
- Align sequence to genome of *Eucalyptus grandis*
- Use replicates to remove false positives



Mutation Pattern Approximately Matches Tree Structure

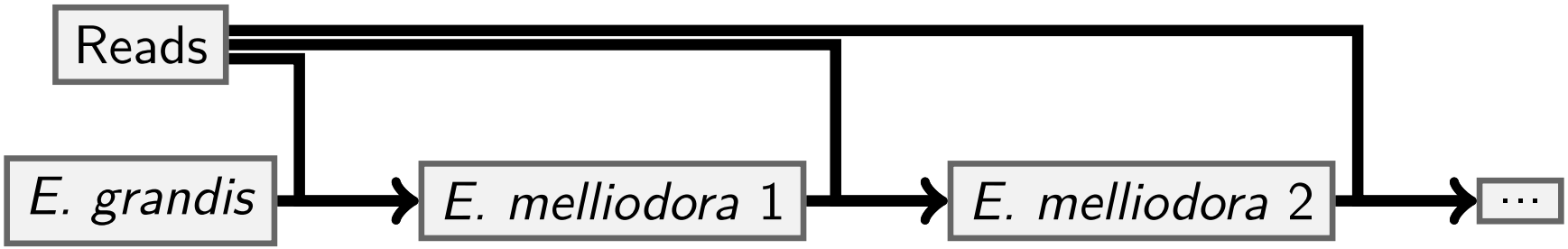


Most Reads Are Not Mapped to the *E. grandis* Reference

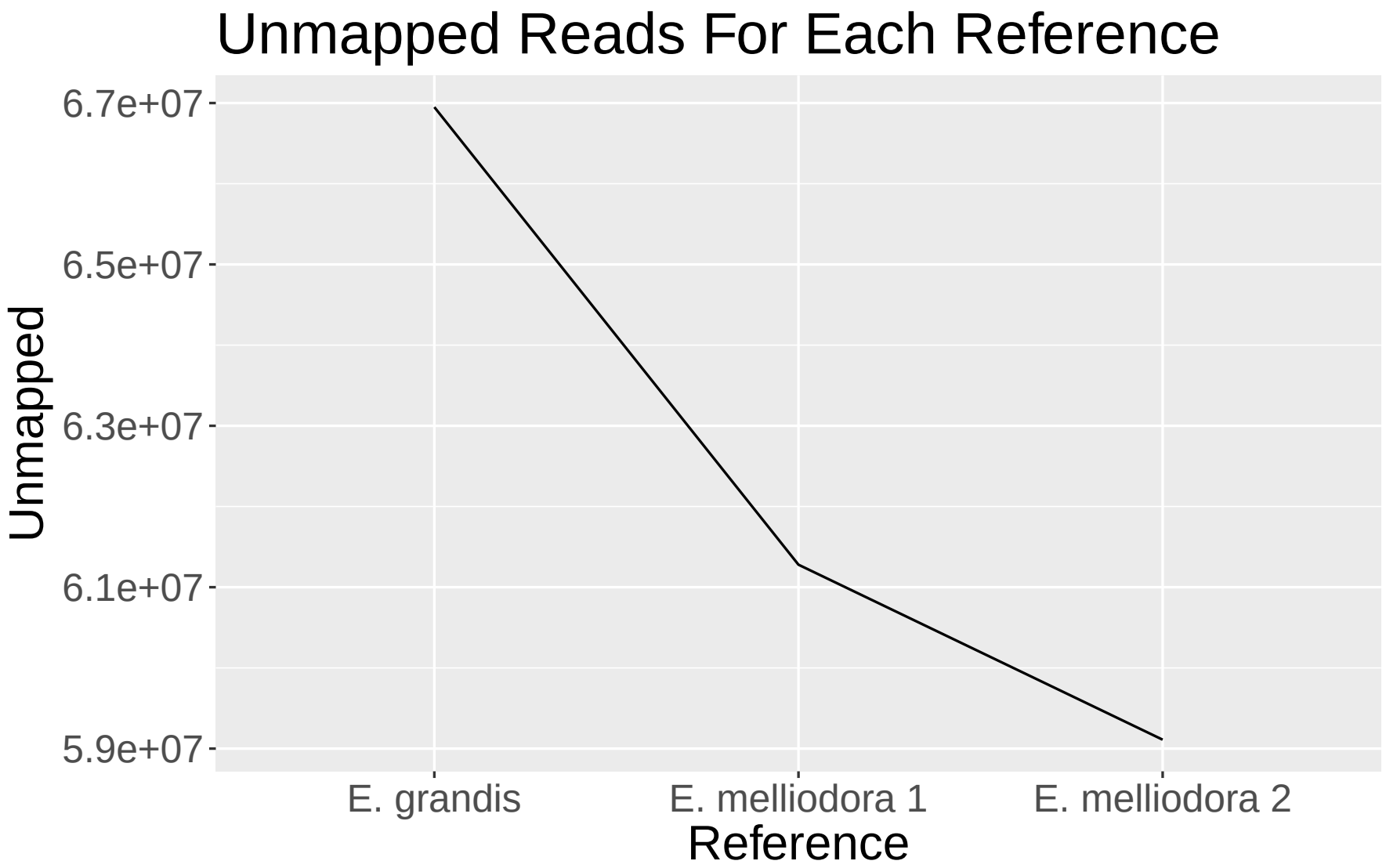


Approximating a Genome

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

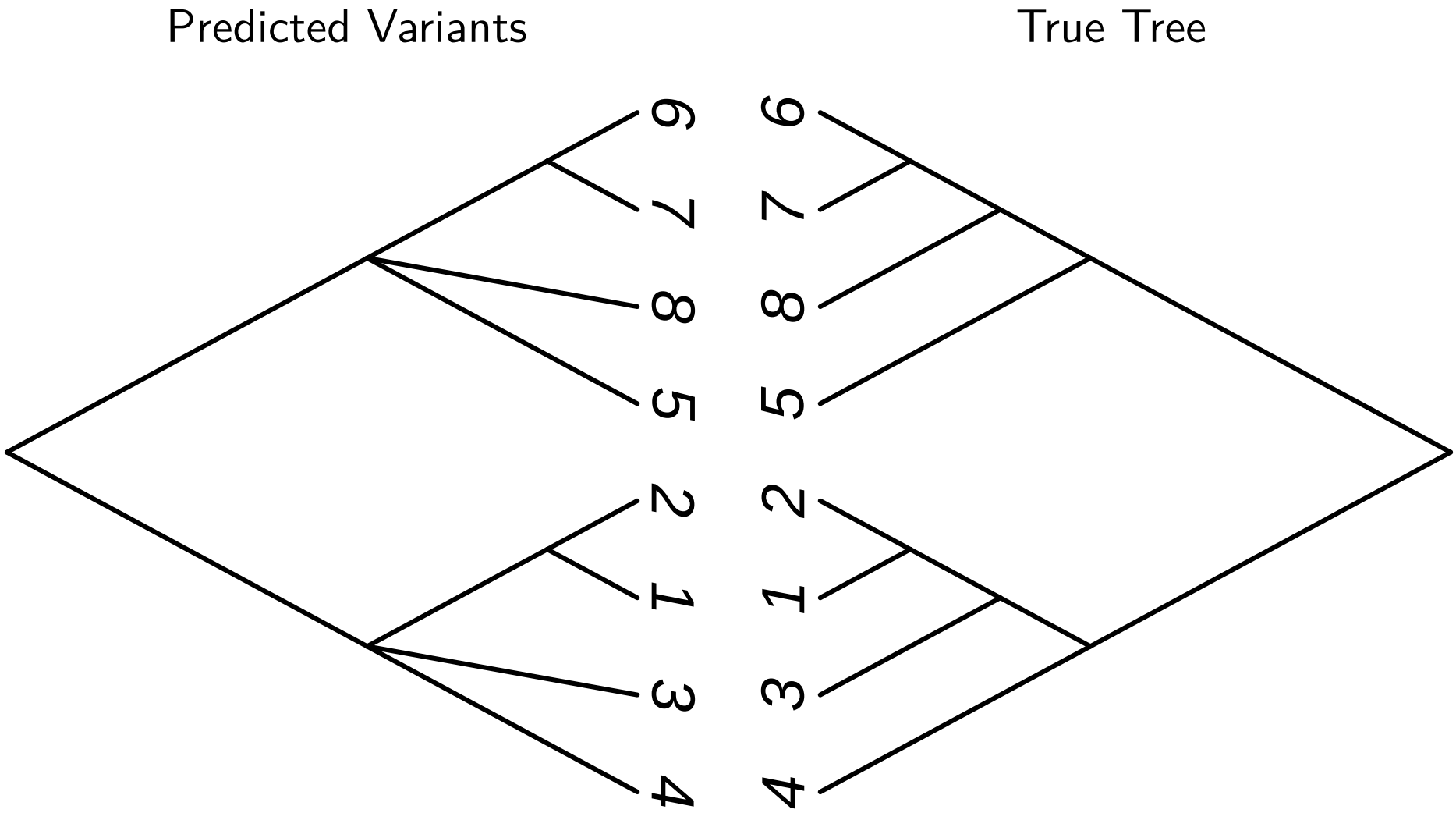


Filtering Variants

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



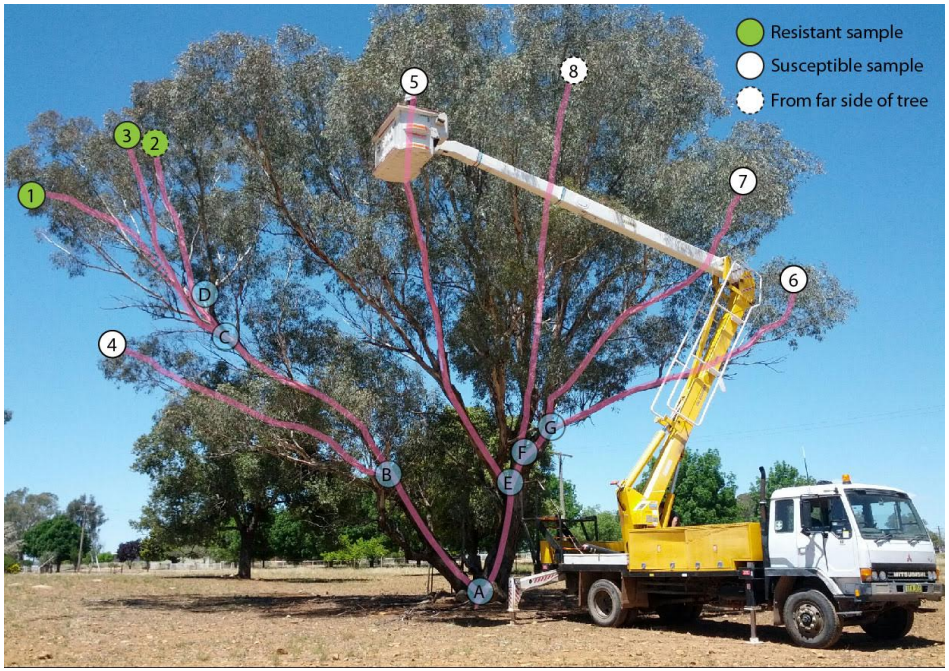
Using Tree Topology Gives Higher Recall Rate

- 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

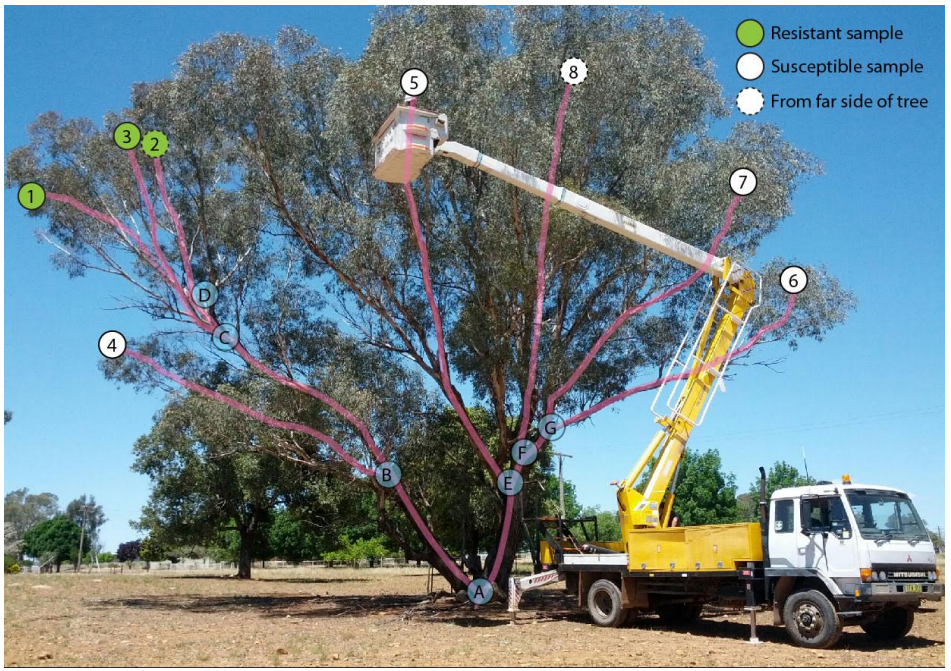
- Detected 90 mutations.
- 20 mutations in genes.
- Estimated recall of $\sim 30\%$.
- $90 \times \frac{1}{.3} = 300$ mutations.
- ~ 3.3 mutations per meter of length
- 2.7×10^{-9} mutations per base per meter
- Somatic mutations account for ~ 55 mutations per leaf tip.



Extrapolating Beyond the Range of the Data

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.

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.

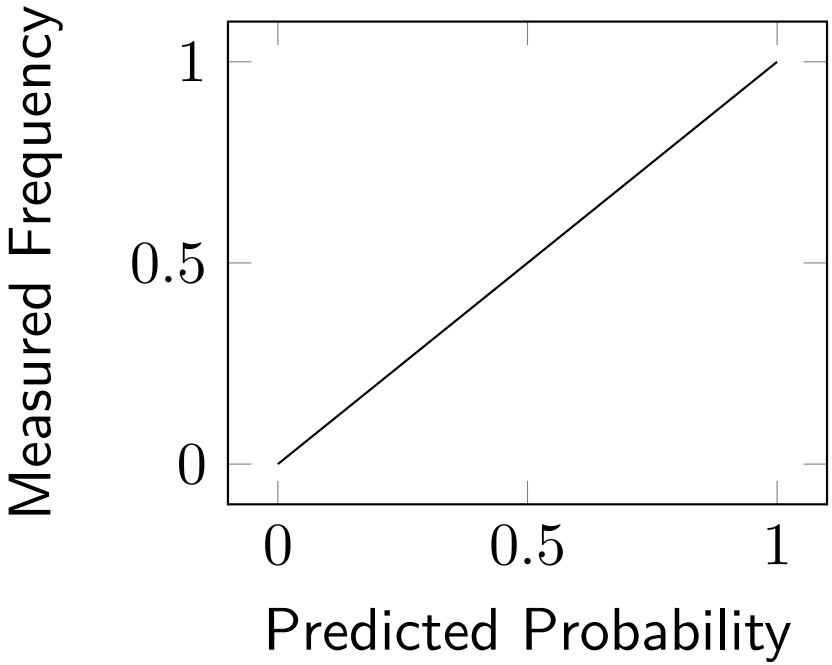
$$P(error) = 10^{\frac{-Q}{10}}$$

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

Quality Score	$P(error)$
1	0.8
2	0.6
3	0.5
4	0.4
5	0.3
6	0.3
7	0.2
8	0.2
9	0.1
10	**0.1**
20	**0.01**
30	**0.001**
40	**0.0001**

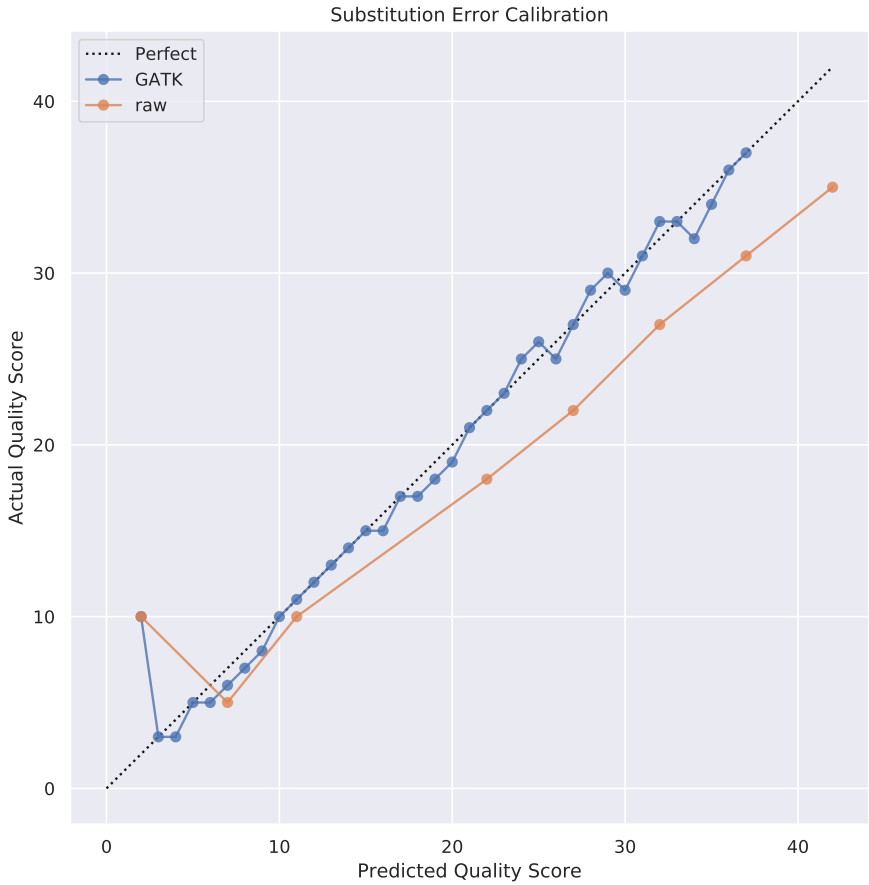
Quality scores are predictions

- 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?

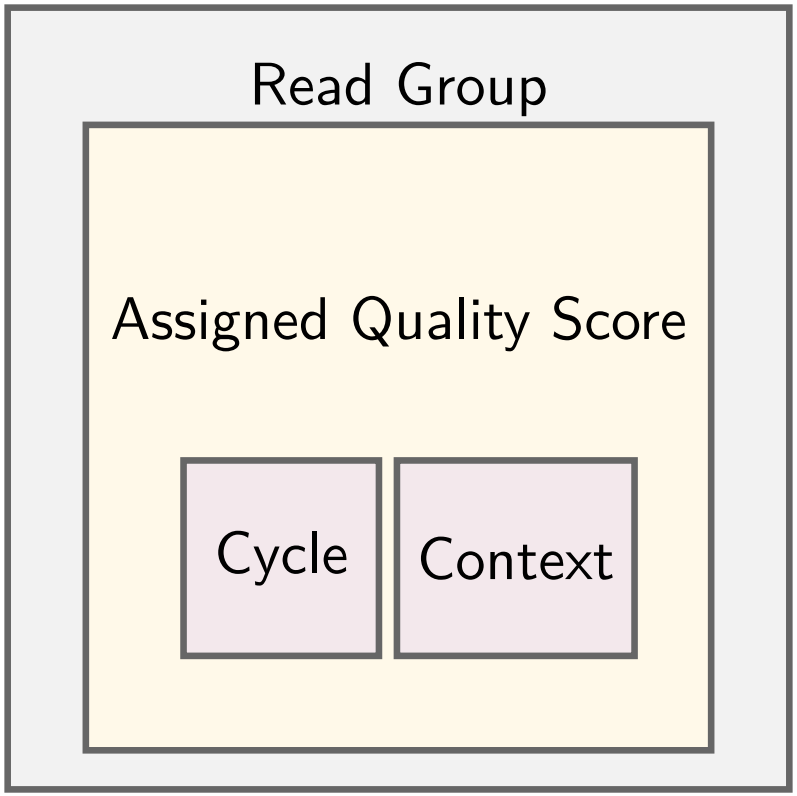


Quality scores aren't well-calibrated

- If quality scores *were* well-calibrated, it would be easy 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.

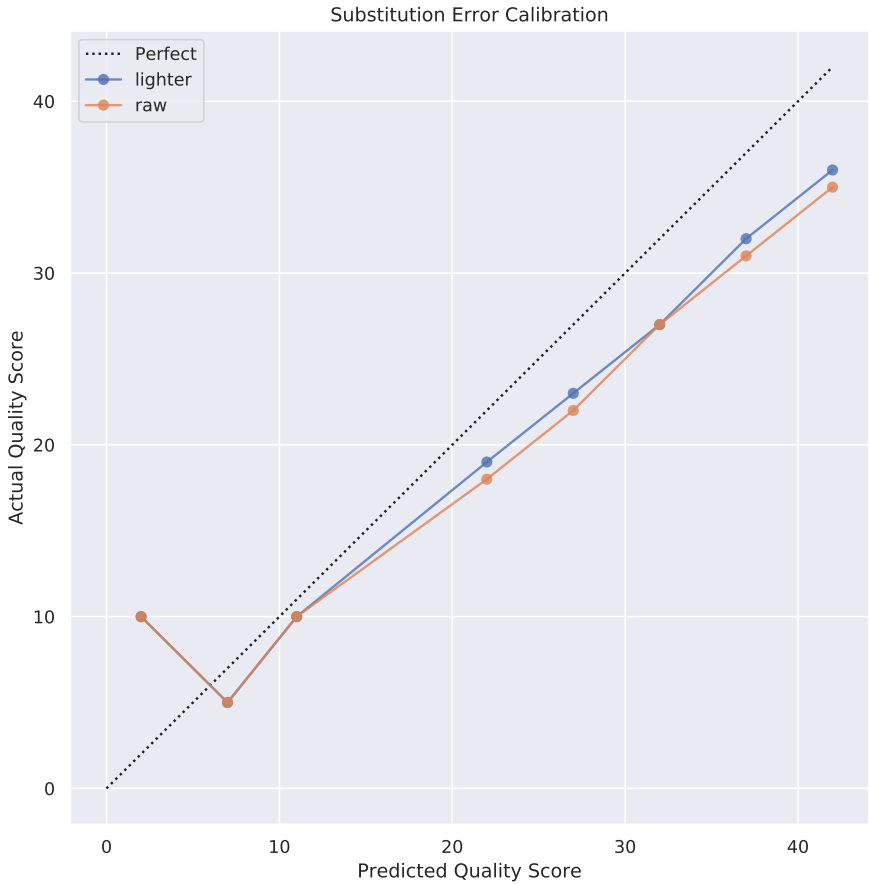


BQSR uses a linear model to determine how much to adjust each quality score



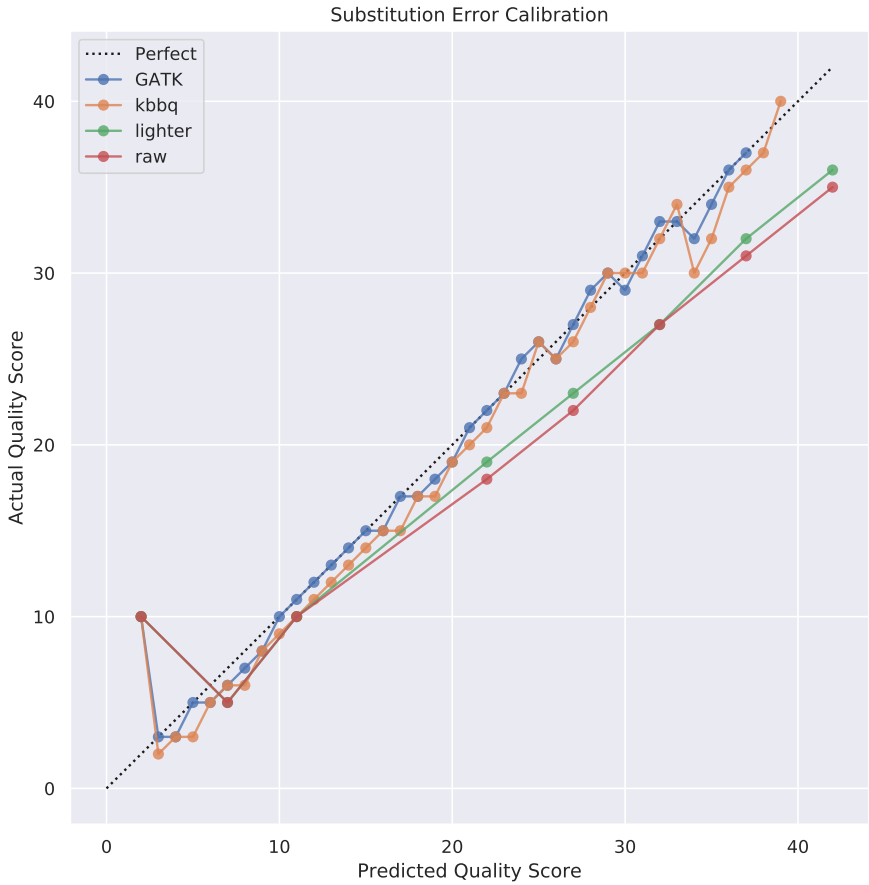
Error correctors can find some errors without a reference

- 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.



K-mer-Based Base Quality score recalibration

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



Future Plans

- Evaluate GATK's robustness to false-negative and false-positive rates
- Evaluate performance of other error correctors
- Evaluate downstream impact on quality of variant calls

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

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