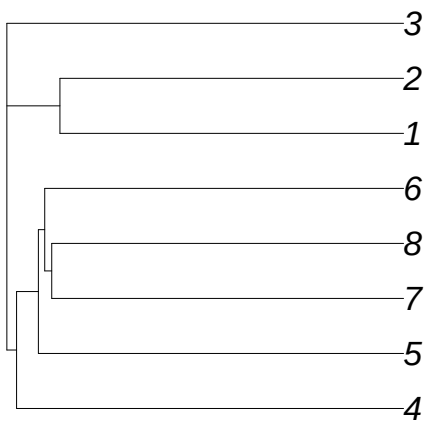
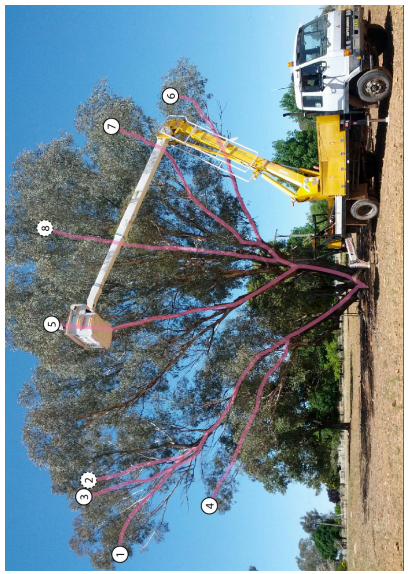
Detection of somatic mutations in *Eucalyptus melliodora*

Adam Orr

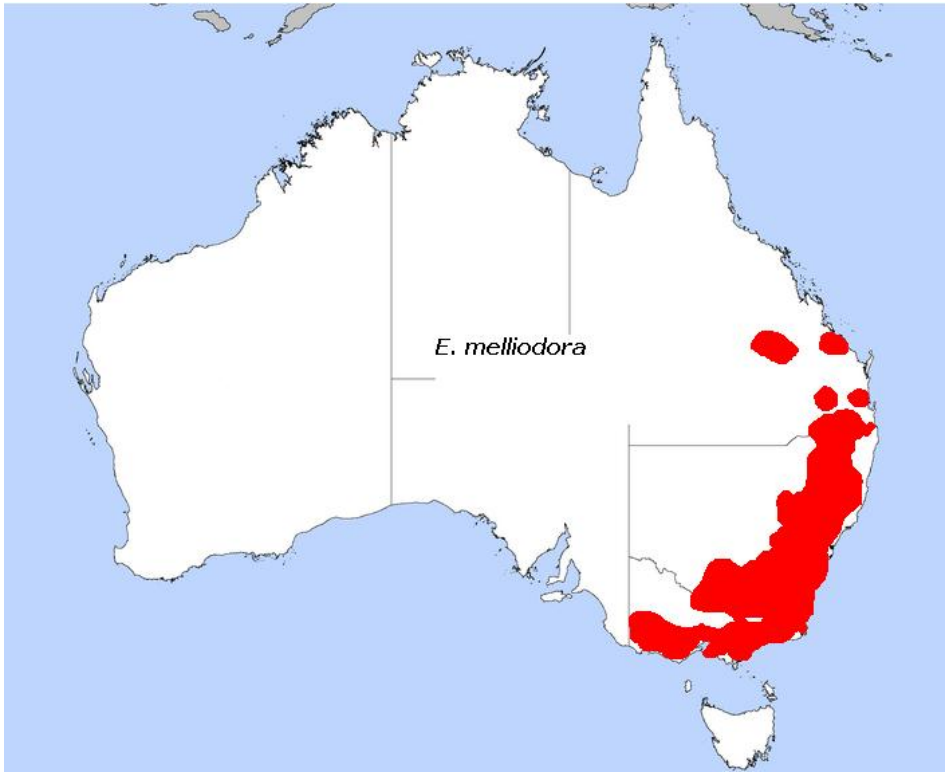
4/8/16





The Yellow Box Tree: *Eucalyptus melliodora*

- Produces **5 times** more nectar than smaller trees.
- Food source for bees
- Strong wood used for bridges



[1]

[1] https://commons.wikimedia.org/wiki/File:E._melliodora.JPG

A Genetic Mosaic

- Edwards identified as mosaic in 1993[2]
- Sheep pen in Yeoval, New South Wales
- Differential oil production gives protection from Christmas beetles
- Is this mutation a controlled process?

[2] *Edwards PB, Wanjura WJ, Brown WV. Oecologia 1993, 95:551557.*

Somatic Mutations are Commercially Interesting

Definition

A **somatic mutation** is a mutation that occurs in non-germline cells

- Nectarines arose from a somatic mutation on a peach tree
- In botany, this is called a **sport**
- Limited understanding of how plants grow

[3]

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

Broad Implications

- How do somatic mutations spread? How can we study this?
- Cancers and other somatic diseases.
- A tree as a system for studying somatic mutation.
- The tree has a built-in point of reference

What mutation is causing the herbivore resistance phenotype?

Somatic mutations are hard to find

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

- Errors accumulate during PCR prior to sequencing - then propagate
- Errors accumulate in amplification steps during sequencing
- Technical error from sequencer

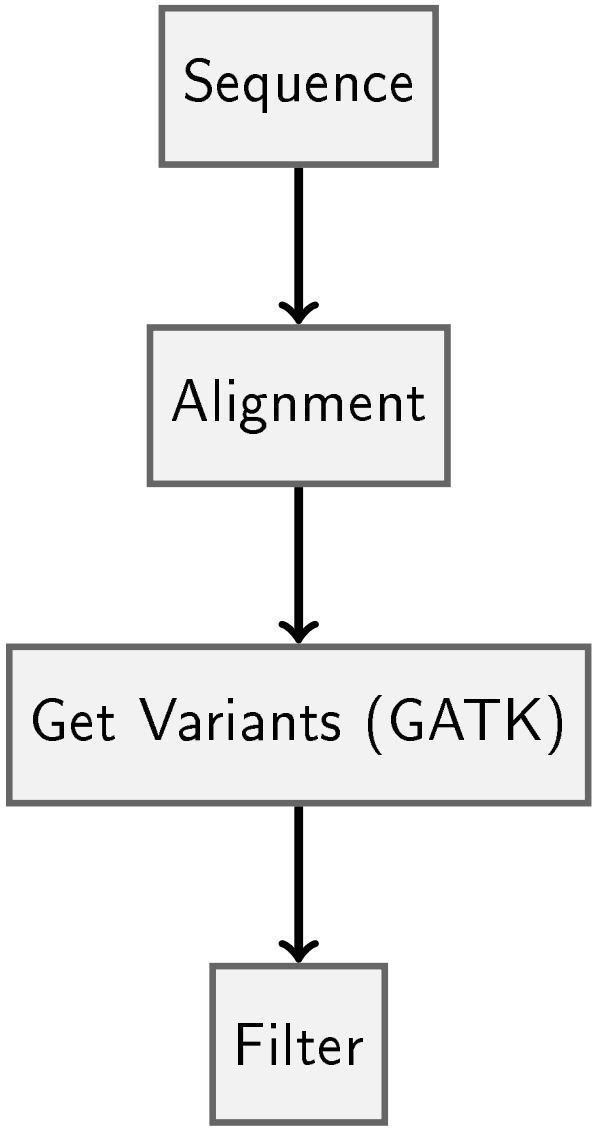
Sequencer error alone is $\sim 10^{-2}$ while mutation rate after error-checking is $\sim 10^{-10}$

Study Methodology

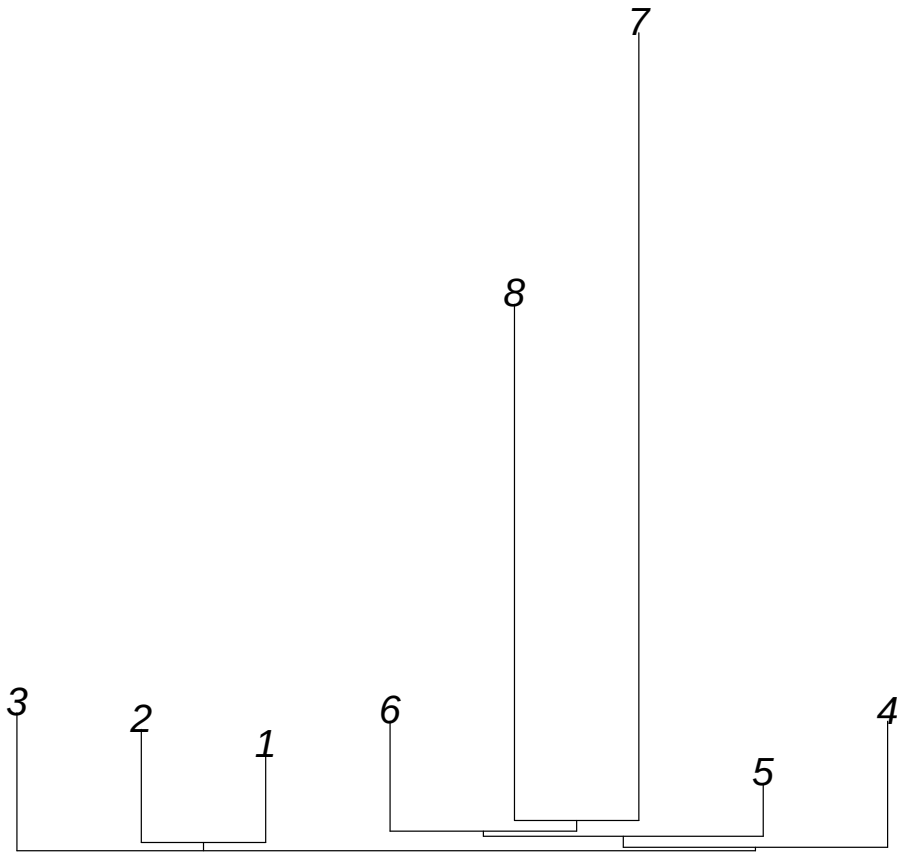
Definition

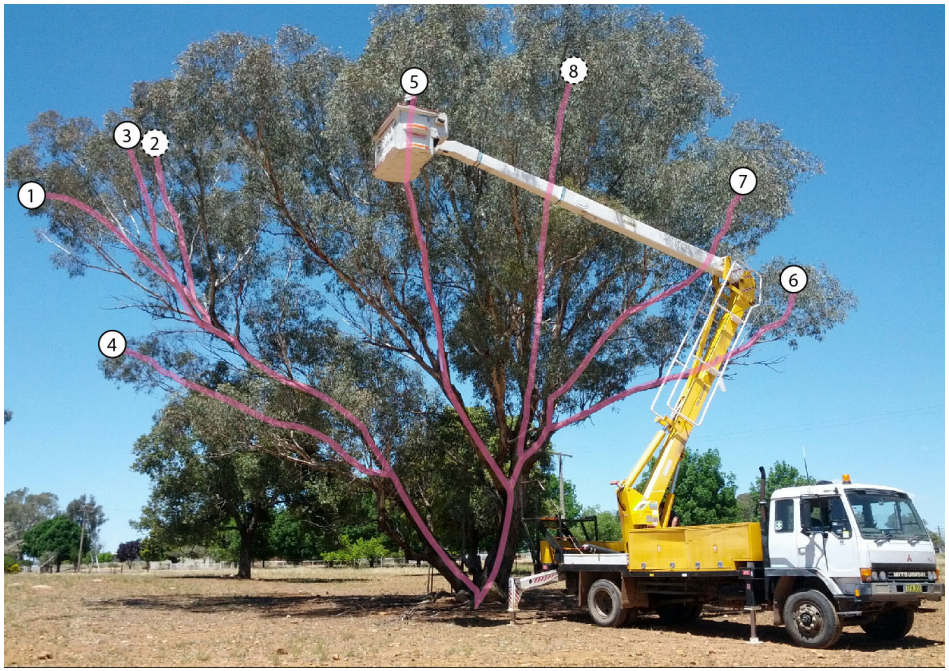
Coverage: Average number of times a single base is sequenced.

- Sequence 8 samples in triplicate
- Ultra-deep coverage for each replicate (∼30X)
- Align sequence to genome of *Eucalyptus grandis*
- Use replicates to remove false positives

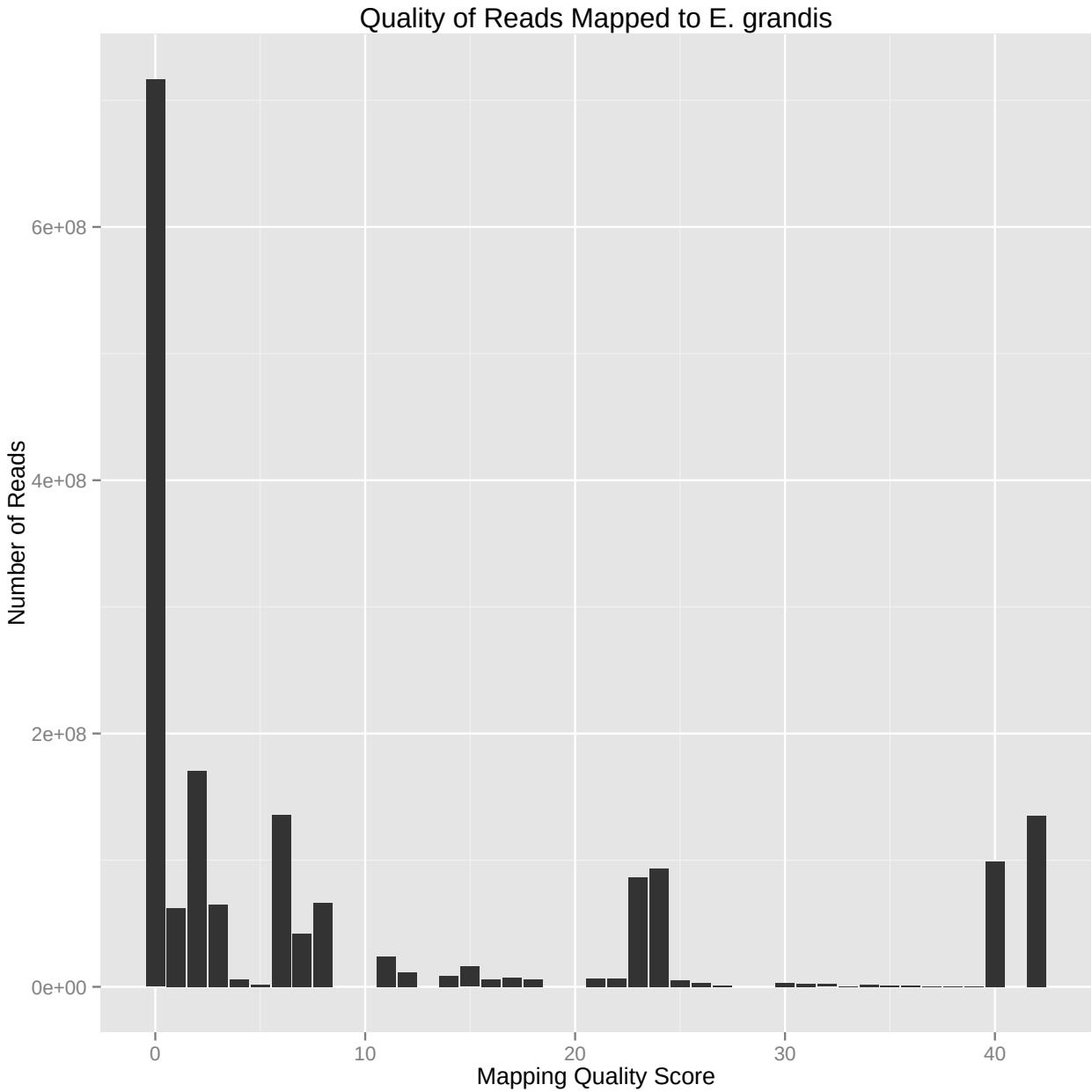


Mutation Pattern Approximately Matches Tree Structure

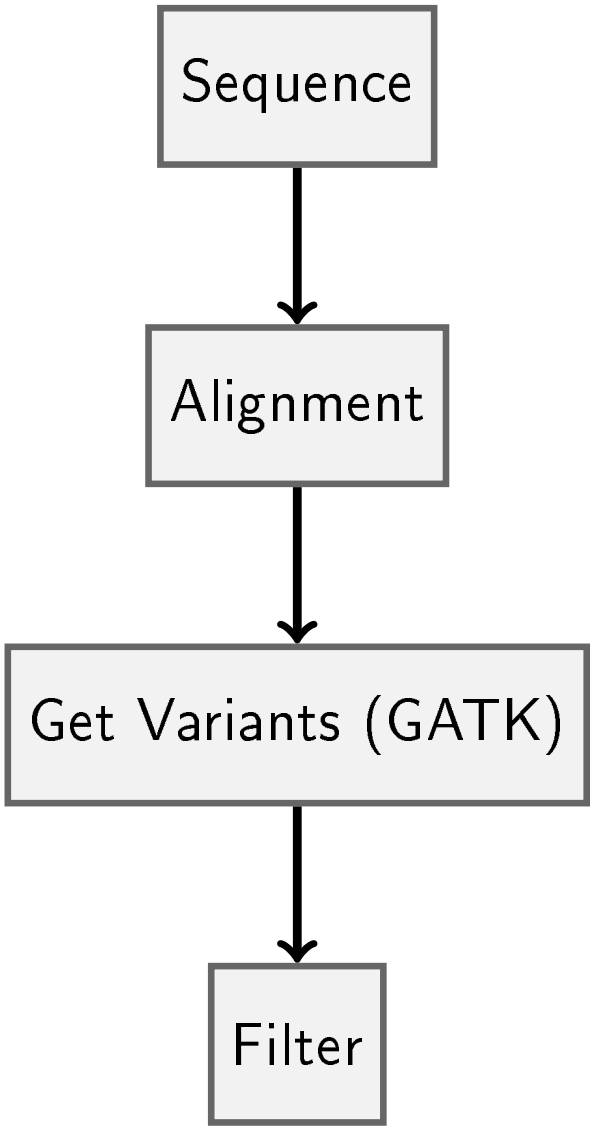


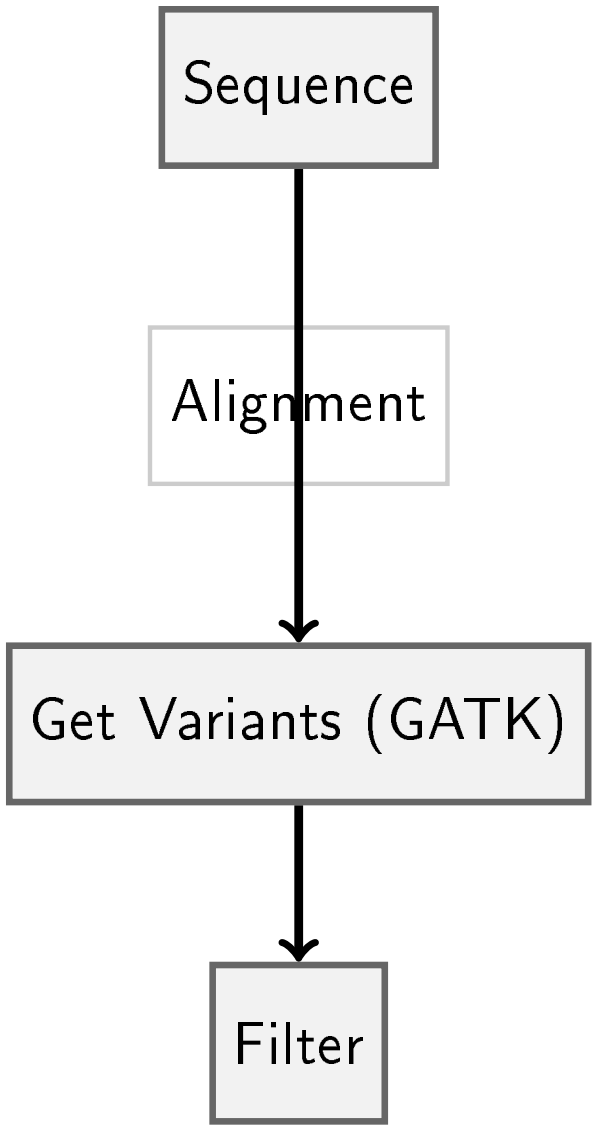


Most Reads Are Not Mapped to the *E. grandis* Reference



A Reference-Free Method





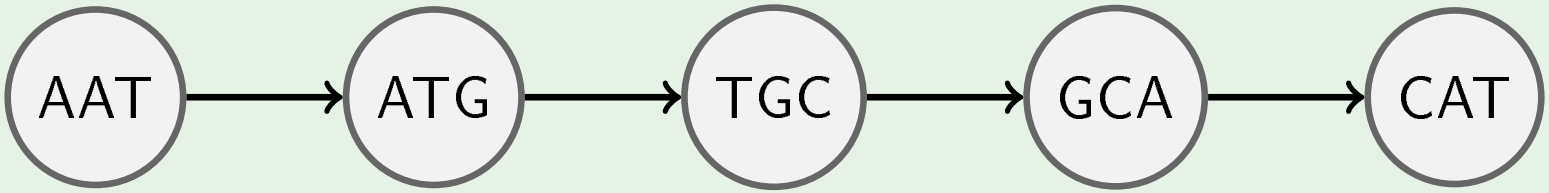
The De Bruijn Graph

Definition

A **De Bruijn Graph** is a graph where the nodes represent symbols and edges represent overlaps between those symbols.

Example

Consider the sequence **AATGCAT** and split it with **kmer** size 3

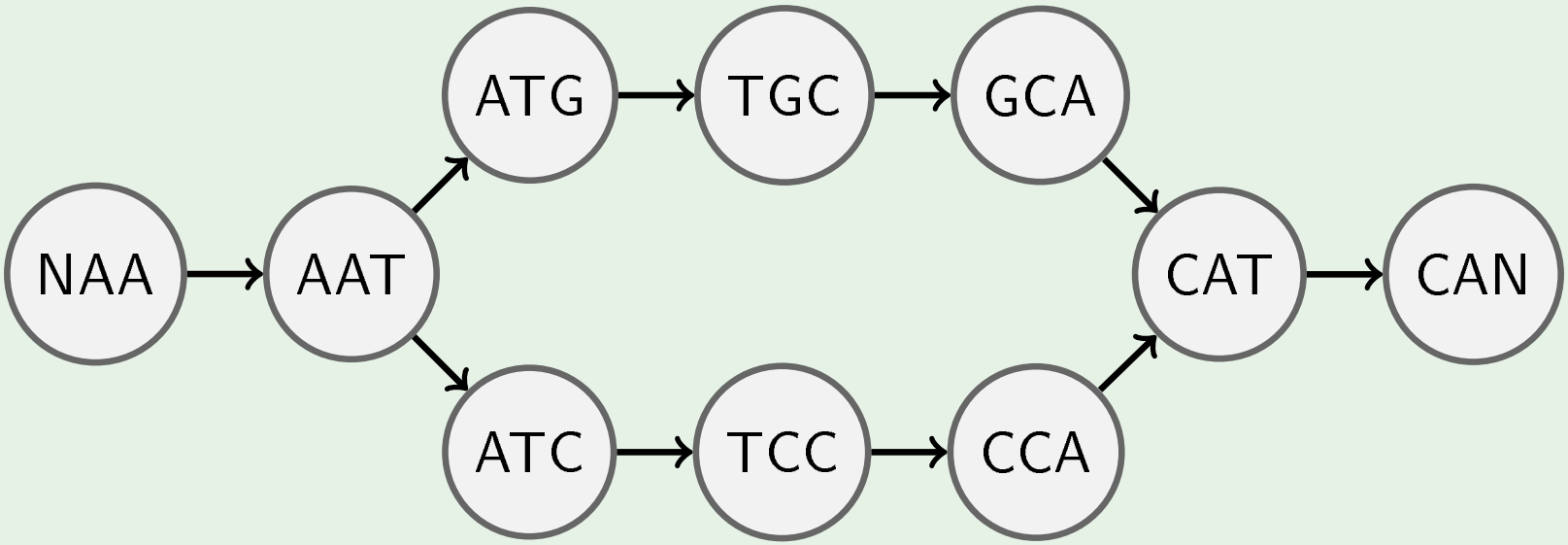


Calling Variants Using Bubbles

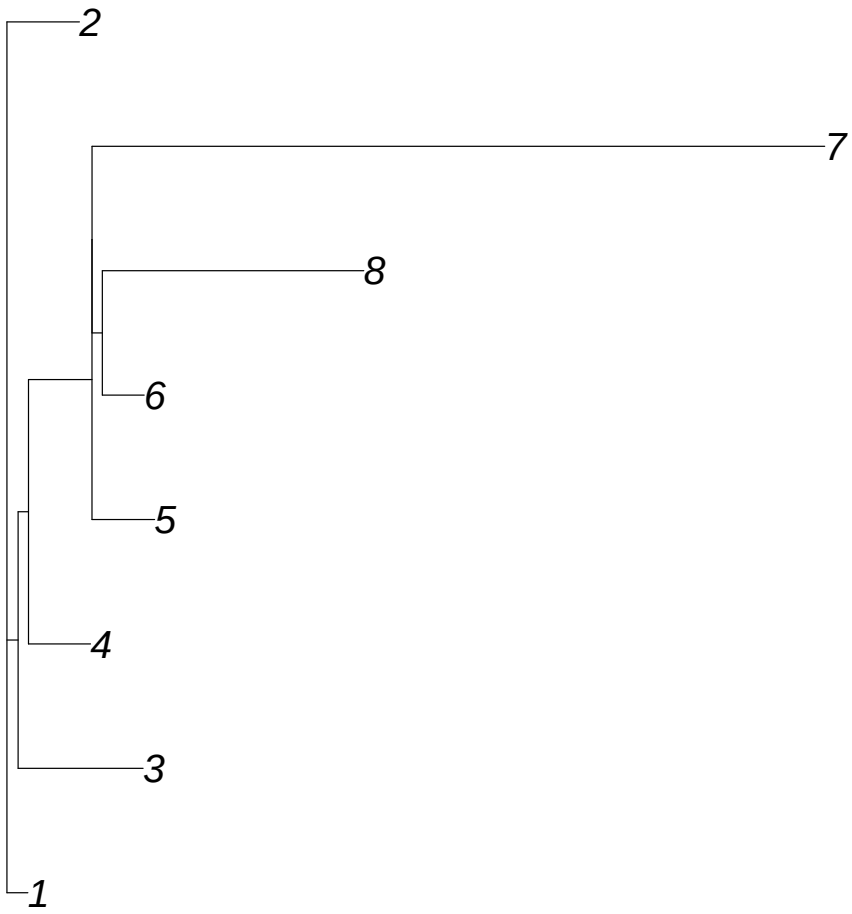
A difference in one base will cause a **bubble** to form in the graph.

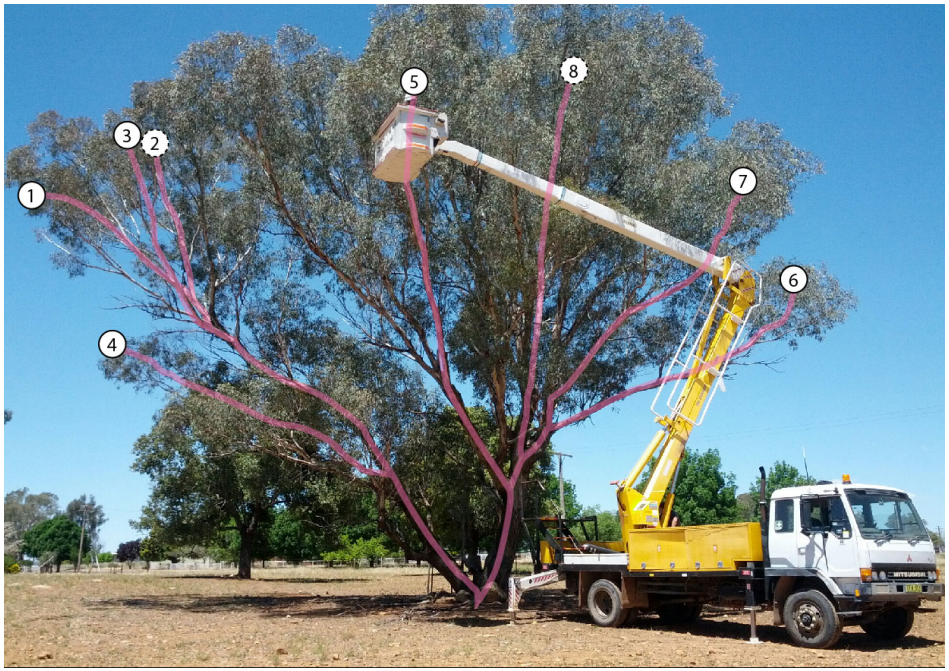
Example

Consider the sequence **AATGCAT** and split it with **kmer** size 3. There is a sample with a G to C mutation!



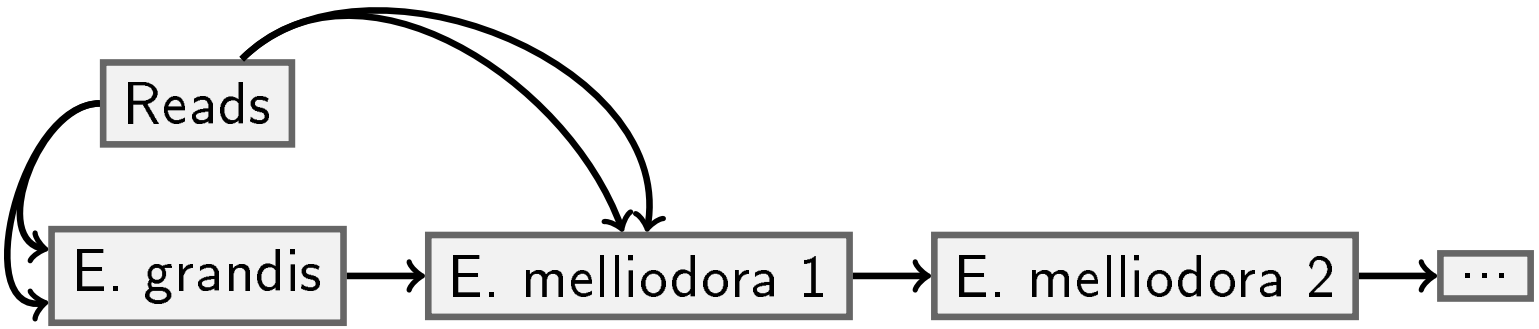
The Reference-Free Method Performs Similarly



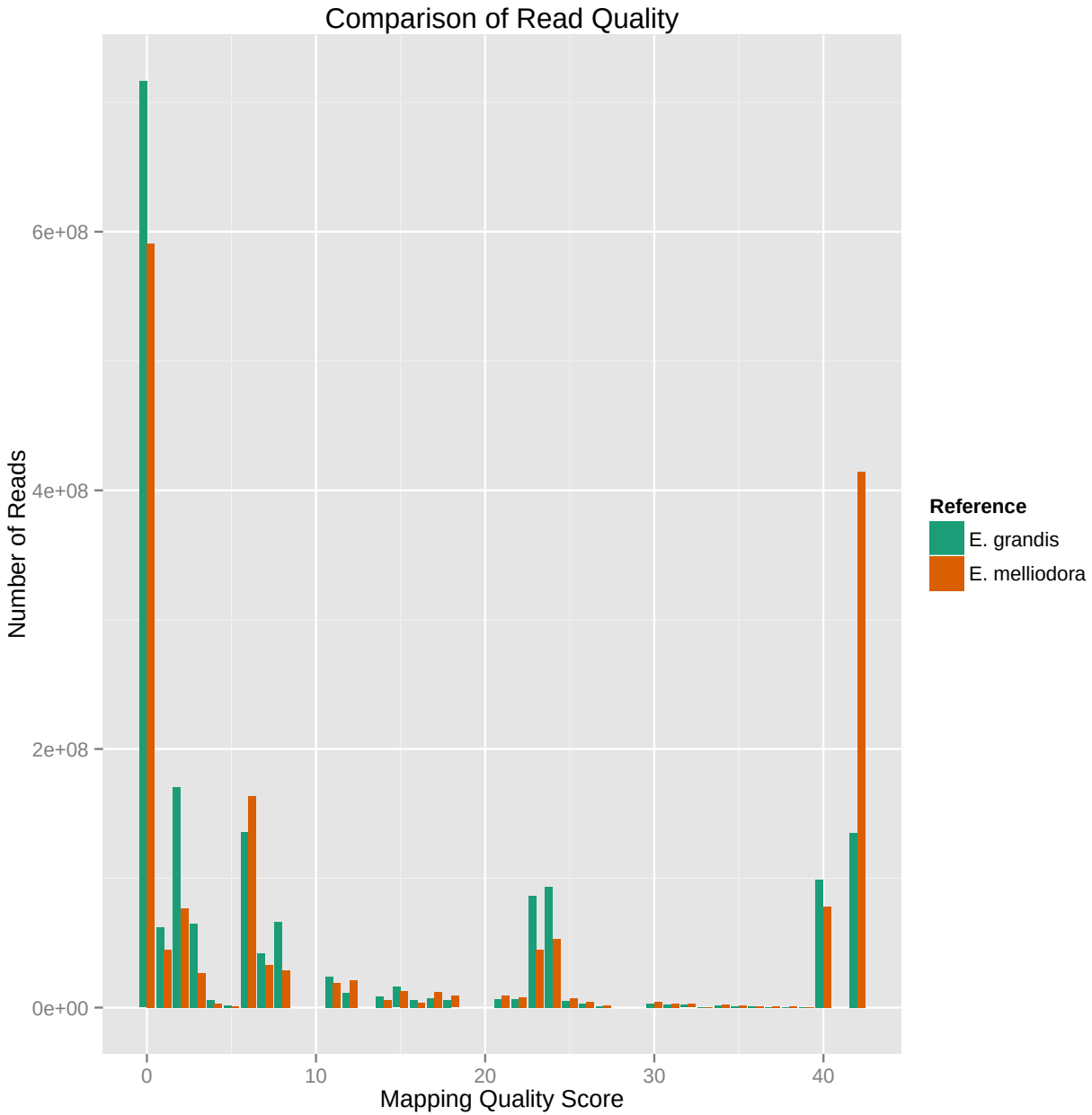


If you want something done right...

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



Our New Reference Improves Mapping Quality



Next Steps

- Continue improving the reference we've created.
- Filter out repetitive elements that make mapping difficult
- Once we are confident in our data, make a prediction about the herbivore resistance
- Validate

Conclusions

- A reference-free method performs similarly to a standard pipeline
- Aligning to a reference, then using that alignment as a reference for another alignment can improve mapping qualities.

Acknowledgements

- Reed Cartwright
- Human and Comparative Genomics Laboratory
- Robert Lanfear, Macquarie University