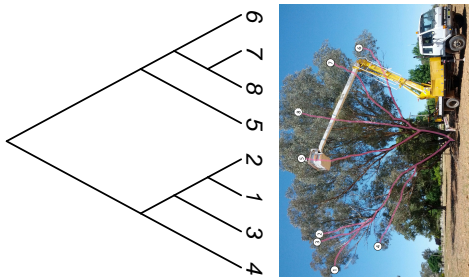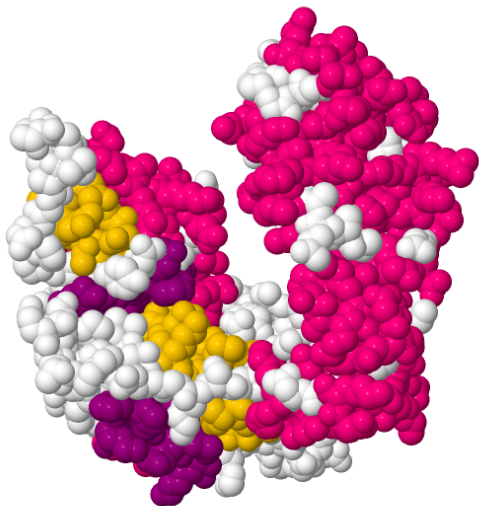# Calculating a non-model *Eucalyptus* Individual's Somatic Mutation Rate

Adam Orr    @AdamJOrr

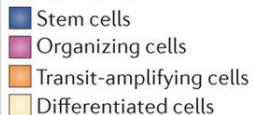9/14/18

# Somatic Mutations Occur During Replication Even Without Exposure to Mutagens
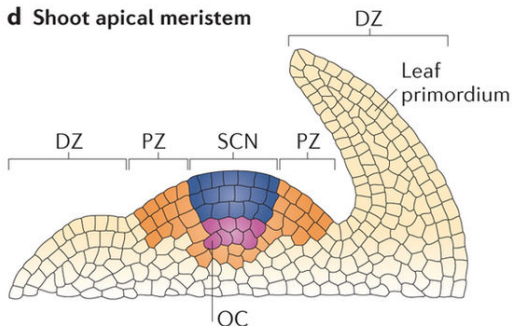


- DNA Polymerase $\beta$
- Mutation rate $\sim 10^{-9}$

PDB: 7ICG

# Plants Grow Directionally

Stem cells
Organizing cells
Transit-amplifying cells
Differentiated cells

**d** Shoot apical meristem

DZ

Leaf primordium

DZ  PZ  SCN  PZ

OC

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

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

# Do Plants Evolve Differently?

NEWS AND VIEWS

**Somatic mutation**

## Do plants evolve differently?

*from William J. Sutherland and Andrew R. Watkinson*

IT has recently been stressed that single plants may consist of a mosaic of geneti-

ornamental plant with variegated leaves knows that it is essential to remove bran-

trees. This model also shows that somatic mutation may be an important source of variation within trees and tree populations. It seems likely that this mechanism is unimportant for species with relatively few meristems (such as peas and maize) but more important for long-lived and extensive species. Thus, individual clones of the aspen (*Populus tremuloides*) can cover

The relative importance of somatic and gametic mutations in plants cannot be assessed until the necessary measurements are made. But it is clear that somatic mutation could be important in many plant species.

Plants are weird.

- Segregated germline?
- Dedifferentiation
- Selfing
- Mutational stress response?

# 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
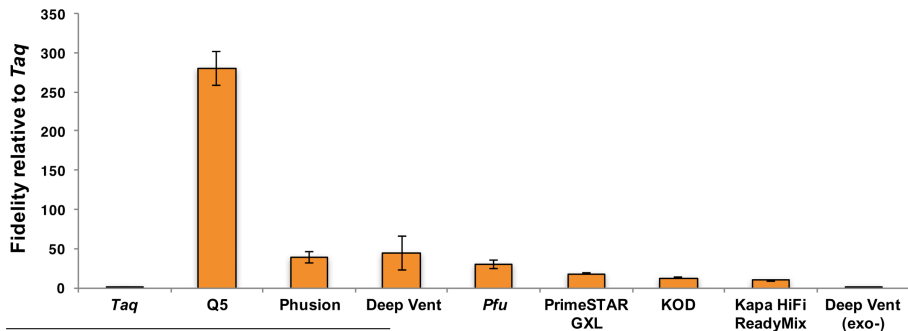


[0] Potapov V, Ong JL (2017) Examining Sources of Error in PCR by Single-Molecule Sequencing

# 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

# Project Goals

### Is it possible?

Can we detect mutations with sufficient accuracy to reconstruct the physical structure of a tree?

### The Relevant Measurements

What's the mutation rate like? Is there evidence of hypermutation?
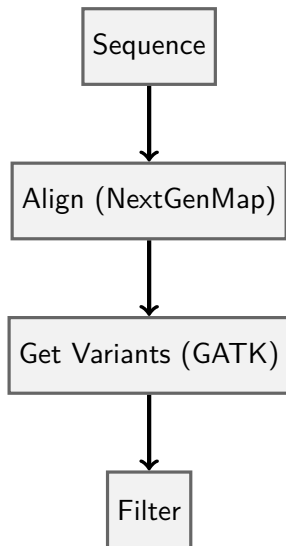
# A Genetic Mosaic



- Edwards identified as mosaic in 1993[1]
- Sheep pen in Yeoval, New South Wales
- Differential oil production gives protection from Christmas beetles
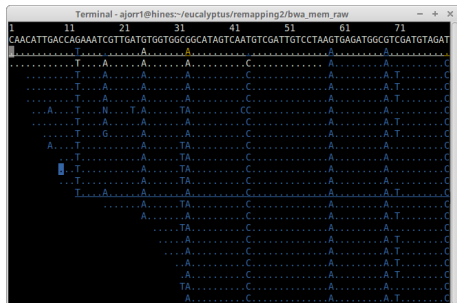
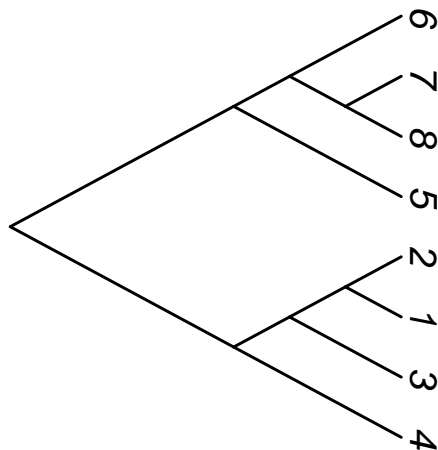[1] *Edwards PB, Wanjura WJ, Brown WV. Oecologia 1993, 95:551–557.*

# 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

```
┌──────────────────┐
│    Sequence      │
└──────────────────┘
          │
          ▼
┌──────────────────┐
│ Align (NextGenMap)│
└──────────────────┘
          │
          ▼
┌──────────────────┐
│ Get Variants (GATK)│
└──────────────────┘
          │
          ▼
    ┌──────────┐
    │  Filter  │
    └──────────┘
```

# 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

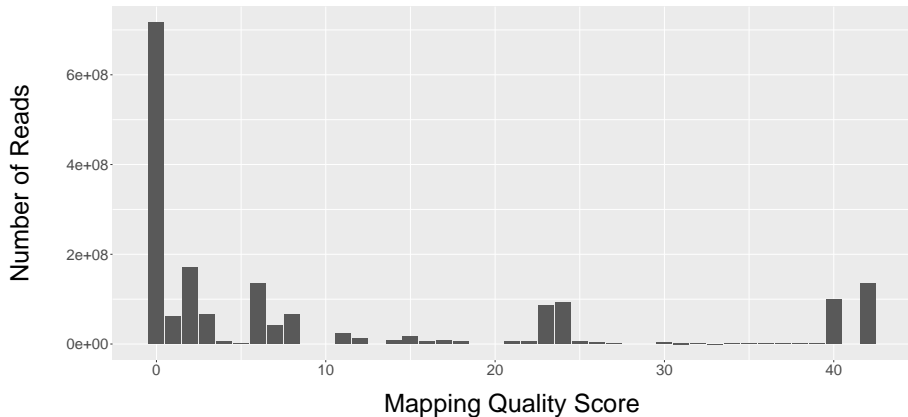GATK Best Practices Tree                                    True Tree
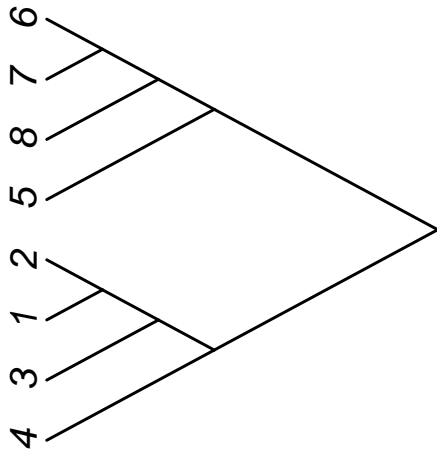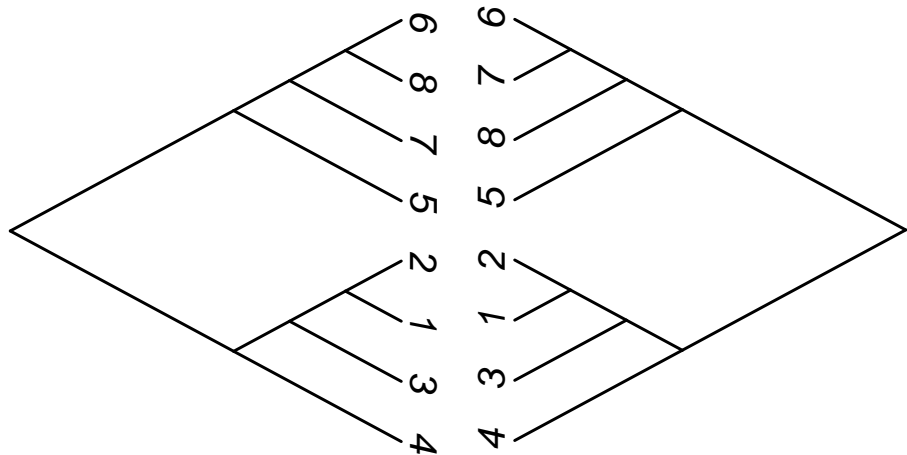
# Most Reads Are Not Mapped to the *E. grandis* Reference



Quality of Reads Mapped

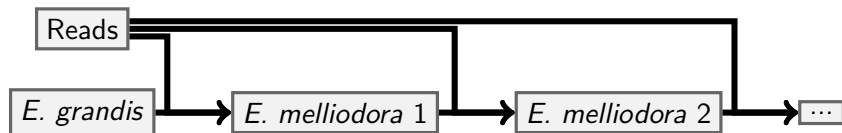# A Reference-Free Method Performs Similarly
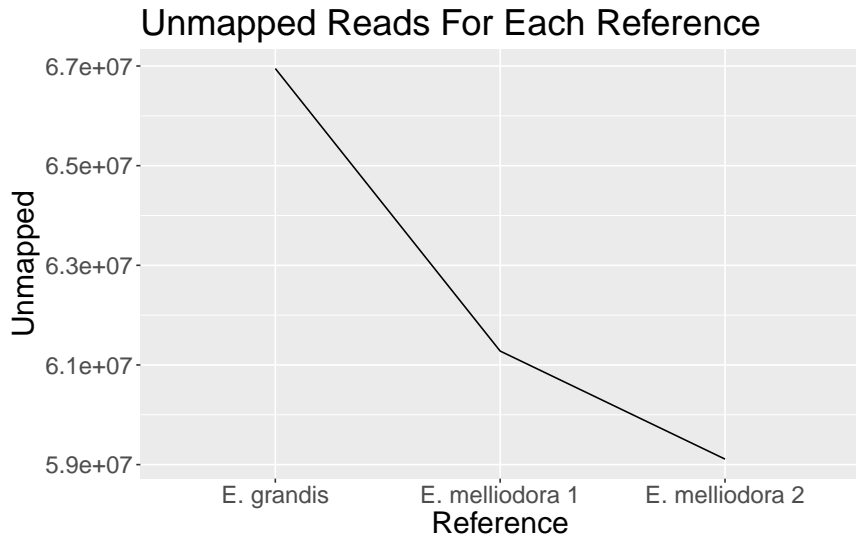


DiscoSNP++ Tree

True Tree

# Approximating a Genome

Use *E. melliodora* genome as a starting place, then generate a new reference and map to that reference.
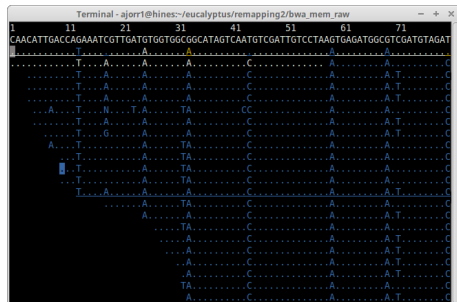
# Our New Reference Has Fewer Unmapped Reads



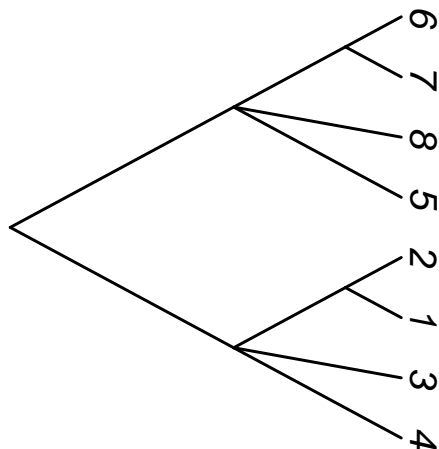Unmapped Reads For Each Reference

# Filtering Variants

Remove variants likely from alignment errors:

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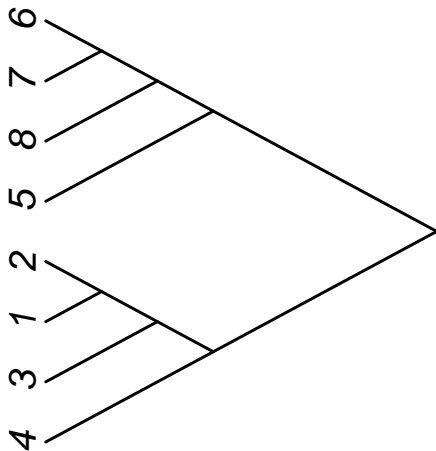- within 50 bases of an indel.
- in repeat regions

# Removing Variants in Repeat Regions Improves Tree Topology



Predicted Variants

True Tree

# Using Tree Topology Gives Higher Recall Rate

- *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% |

# Randomizing Trees to Estimate False Positive Rate

- Does using a phylogeny increase our false positive rate?
- Simulate 100 trees maximally distant from the true tree.
- Does our pipeline detect variation where there shouldn't be any?
- **All** sites detected were variable sites.

# Mutation Rates

- Detected $\sim 91$ mutations.

- 20 mutations in genes.

- Estimated recall of $\sim 30\%$.

- $91 \times \frac{1}{.3} = 303$ mutations.

- $6 \times 10^{-7}$ mutations per site per tree

- $\sim 4$ mutations per meter of length

- Somatic mutations account for $\sim 65$ mutations per leaf tip.
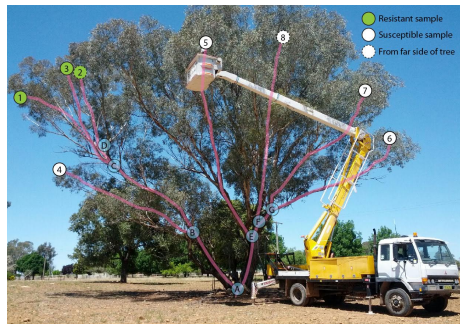
# Acknowledgements

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

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