Detection of somatic mutations in *Eucalyptus melliodora*

Adam Orr

3/4/16

About Me

- Graduated from ASU with dual major in Math and Molecular Bio
- Computational biology in Cartwright lab
- Interest in mutation, disease, and evolution
- Honors thesis: Database of approximate gene "ages" using homology and species divergence

Motivation

Remember, a phylogeny is a hypothesis!

We use simulations because we can't know the truth

Knowing the truth will allow us to evaluate phylogenetic tools

Definition

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

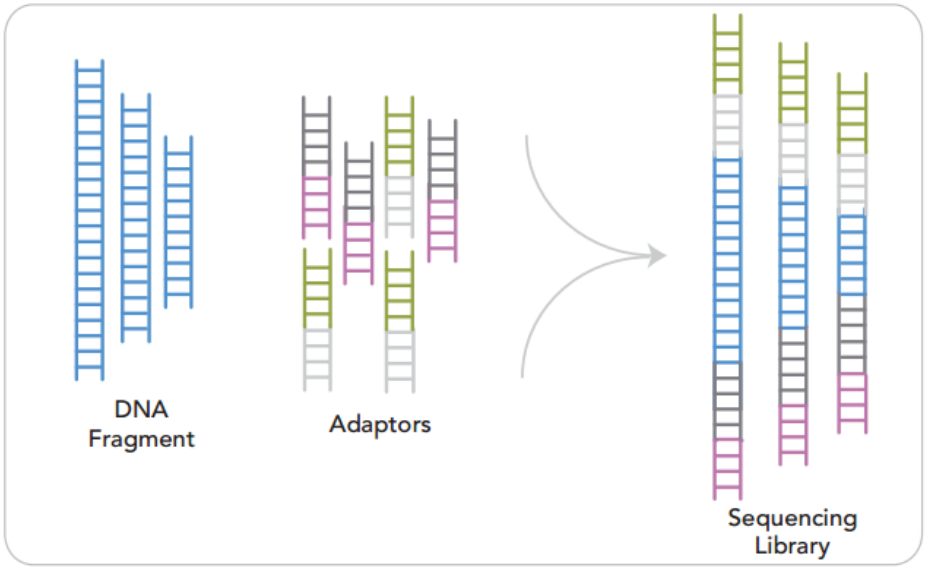
- How do somatic mutations spread? How can we study this?
- Cancers are heterogeneous
- Other somatic diseases

A Primer on Next Gen Sequencing

- Send to the core lab
- Get data

Illumina Library Preparation

- The advantage of Illumina: massive parallelization
- Amplify your DNA
- **Physically** shear your DNA
- Add adaptors and inserts to your DNA fragments

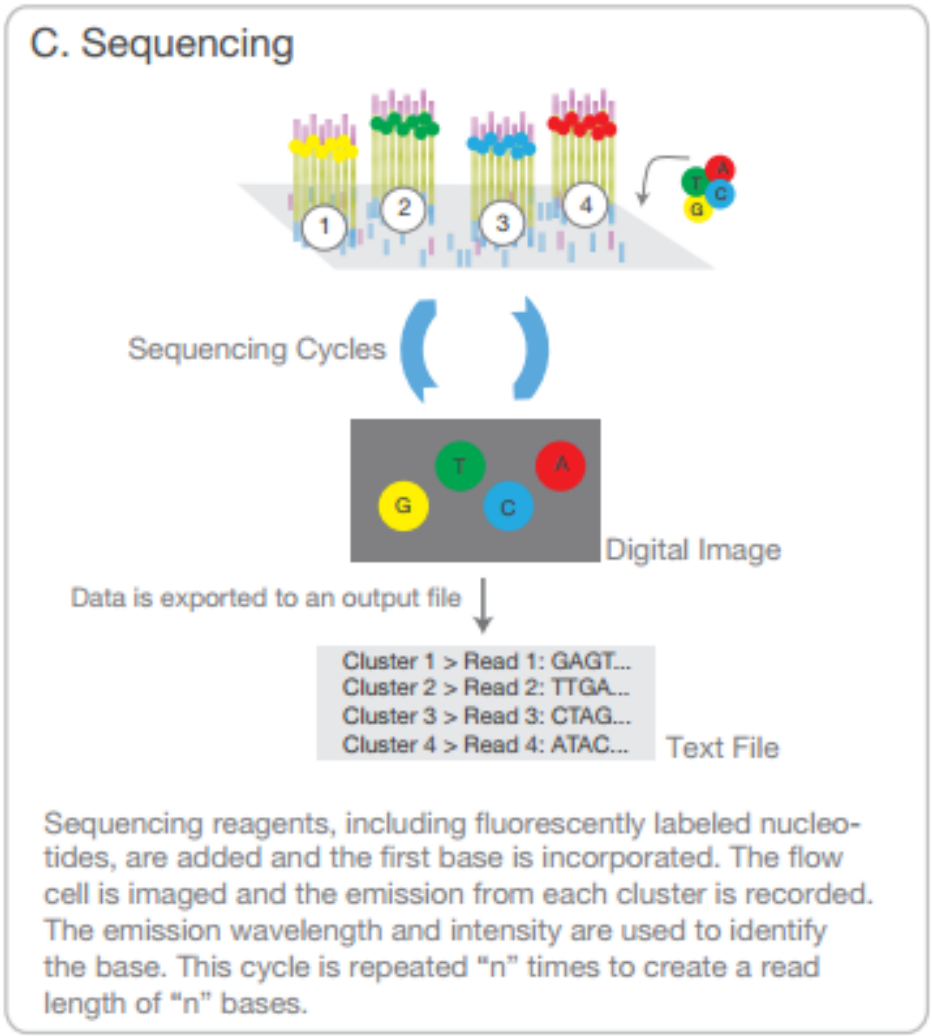


Illumina Sequencing

- Adaptor is bound to chip
- Bridge amplification
- Both strands are bound to chip
- Flood with fluorescent nucleotides
- Rinse and repeat

Follow Me on Instagram

Illumina sequencing is a problem of optics. Advances in Illumina sequencing are all fundamentally advances in camera technology.

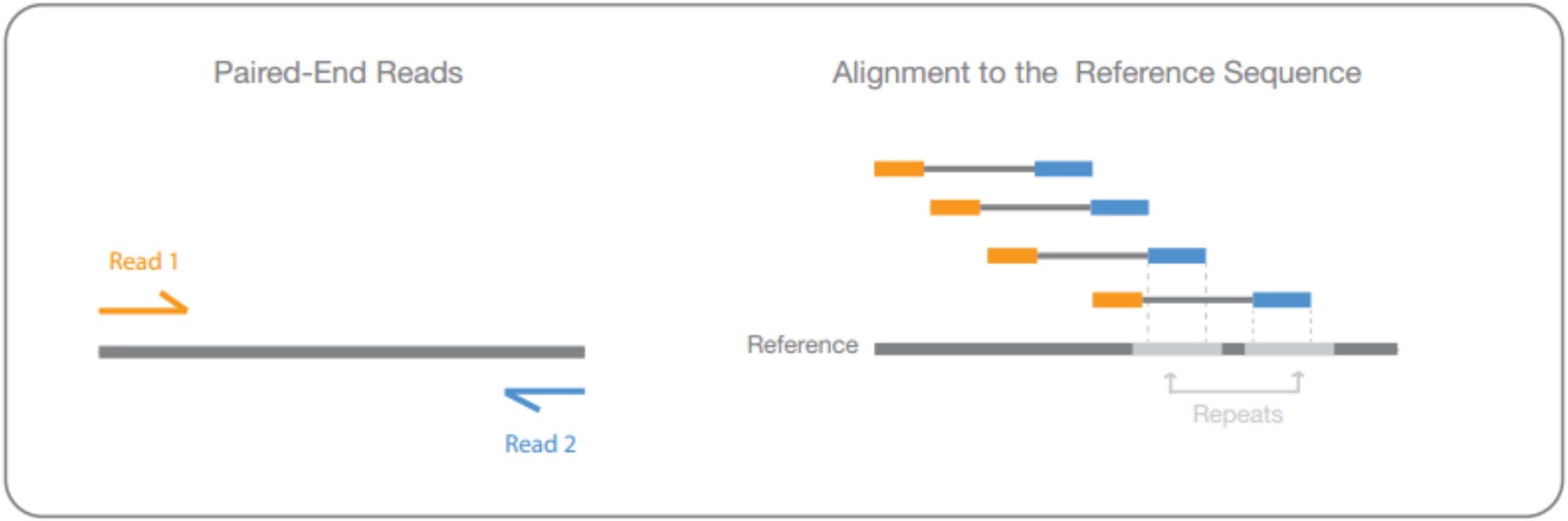


Alignment

Definition

A **read** is a short (30bp) segment of sequence determined by the sequencer.

- Line up reads to a reference
- Some software takes advantage of paired-end technology to improve alignment

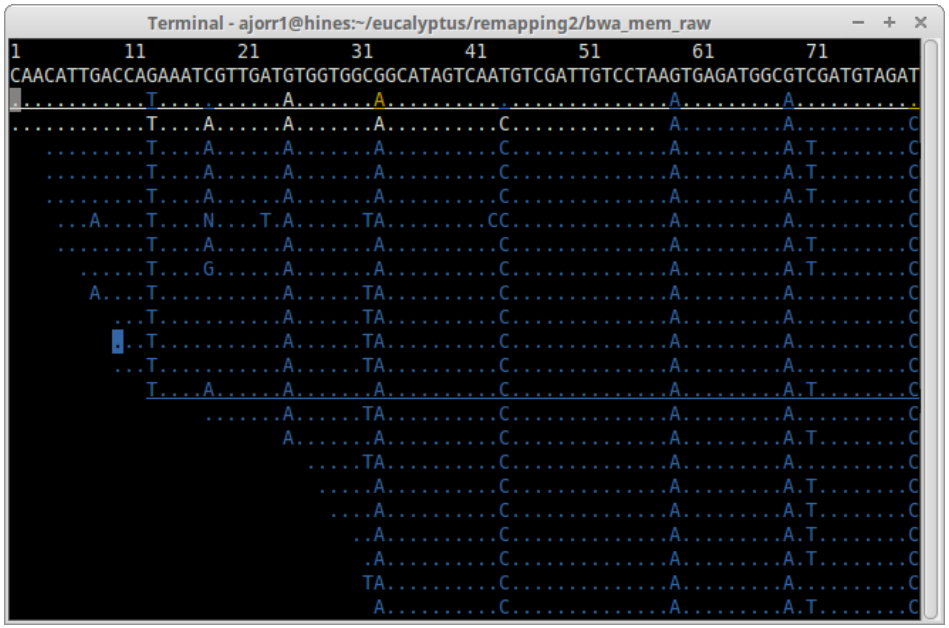


Variant Calling

Definition

Coverage is the number of reads aligned at a particular position. Large variations in coverage is a symptom of poor alignment or library prep.

- Account for sequencing error
- Variation in sequenced population
- Ploidy
- Could be Complex algorithm, may be as simple as checking if the base agrees with the reference.



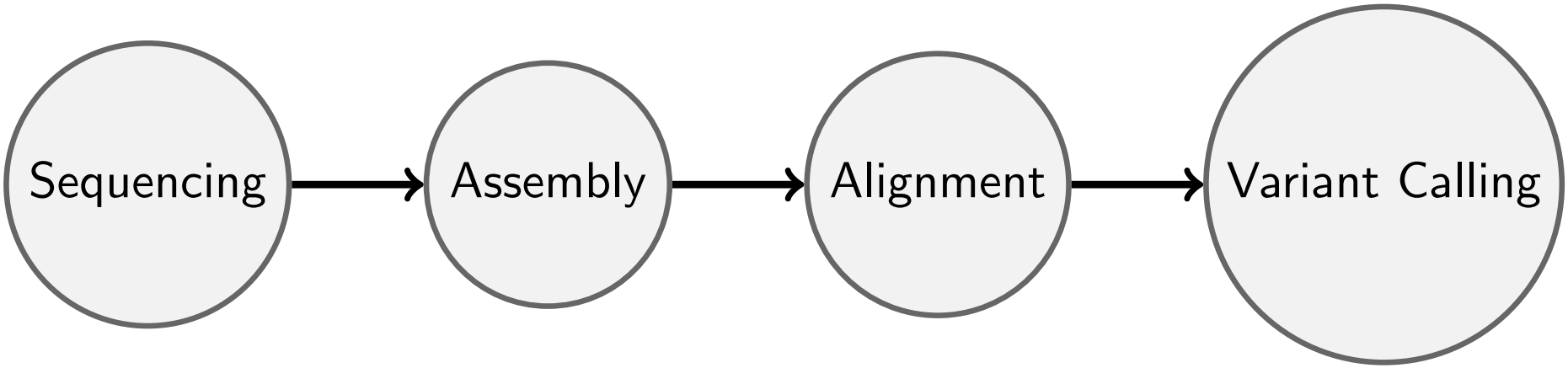
Assembly

Definition

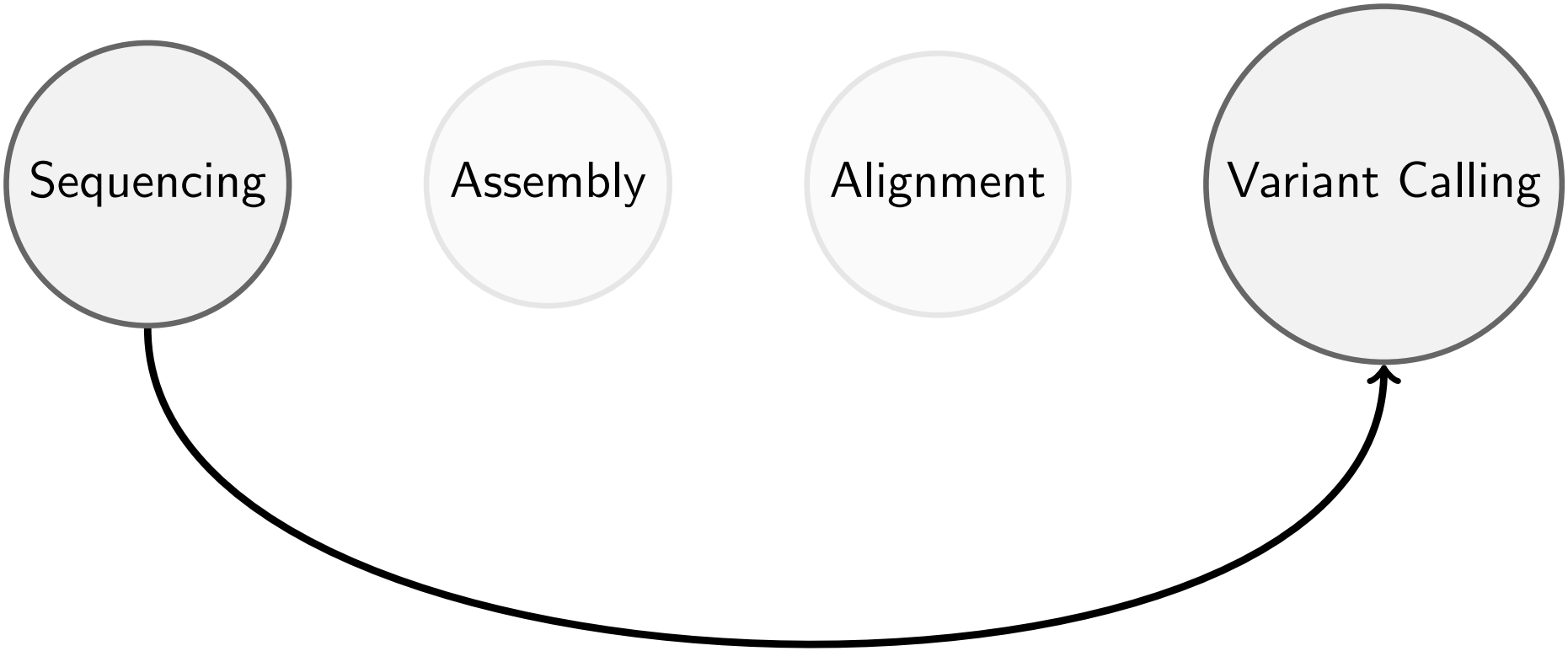
A **contig** is a contiguous stretch of sequence that has been assembled from reads.

- Look for overlaps between sequence
- Long contig lengths is a sign of a good assembly
- Very computationally expensive and time-consuming
- Difficult to get right
- Contigs are **not** chromosomes
- Sometimes transcript data can be used to join contigs

The Pipeline



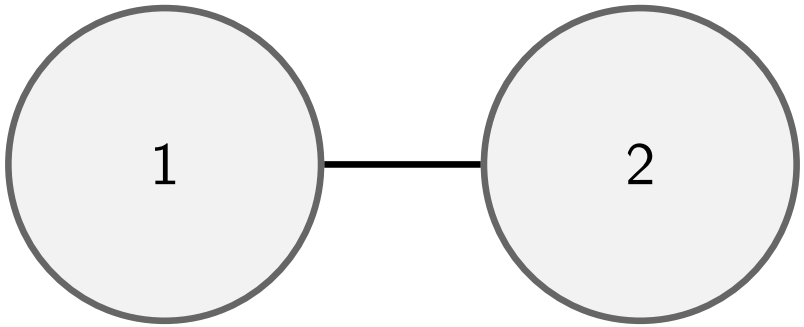
What if there was a better way?



What is a graph?

Definition

A **graph** is a set of nodes and a set of edges, G = (V,E)



$V = \{1 , 2\}$
$E = \{(1,2)\}$

The De Bruijn Graph

Definition

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

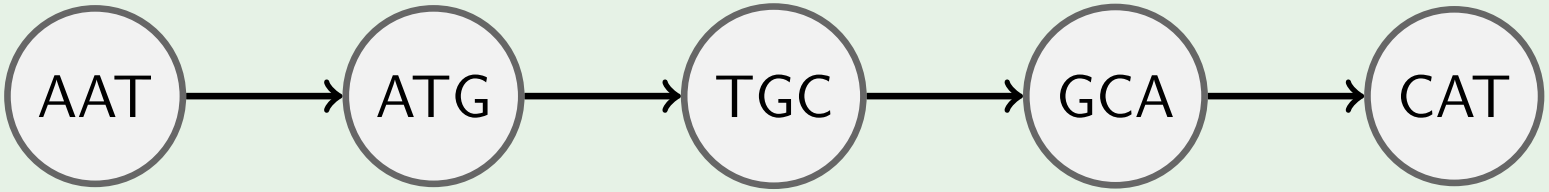
De Bruijn Graphs are **Eulerian**

Definition

A **Eulerian** graph is a graph in which each edge is visited once.

Example

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

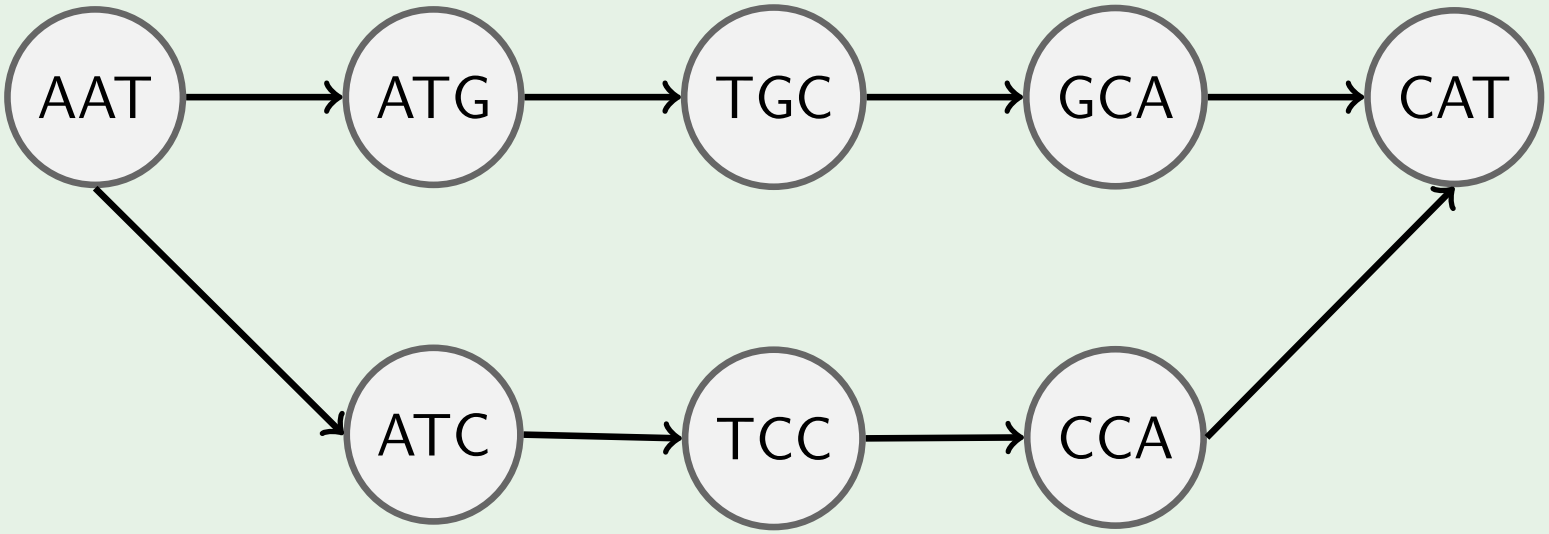


Calling Variants Using Bubbles

What happens when your pooled sample contains a mutant?

Example

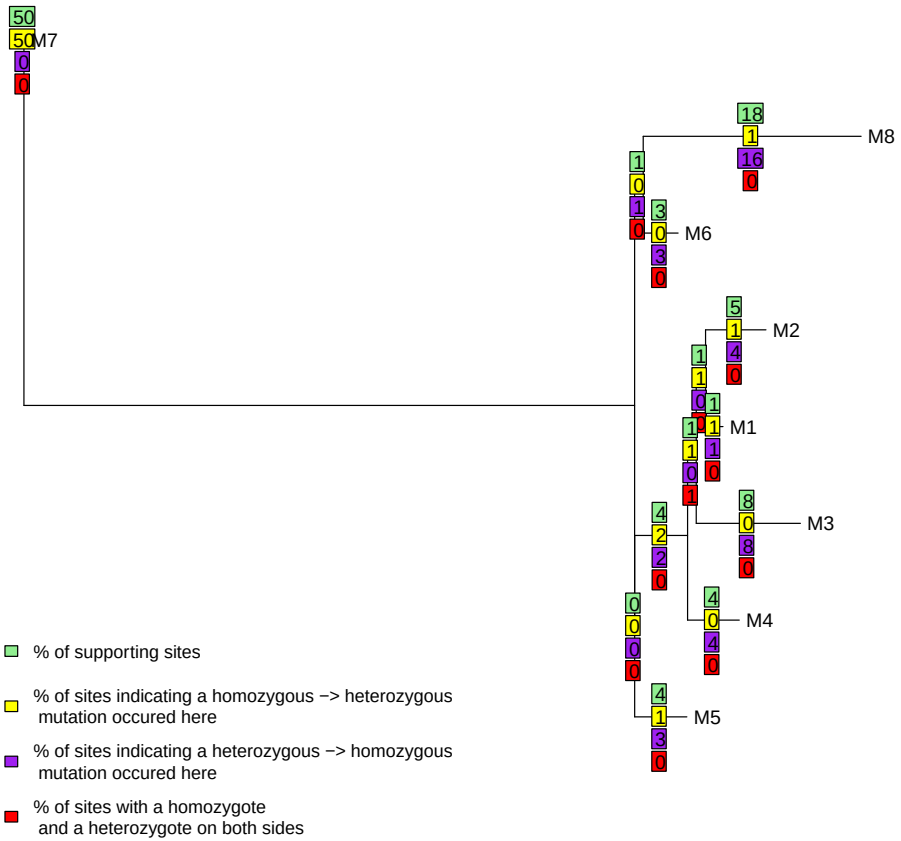
Consider the sequence **AATGCAT** and split it with **kmer** size 3. There is a sample with a G to C mutation!

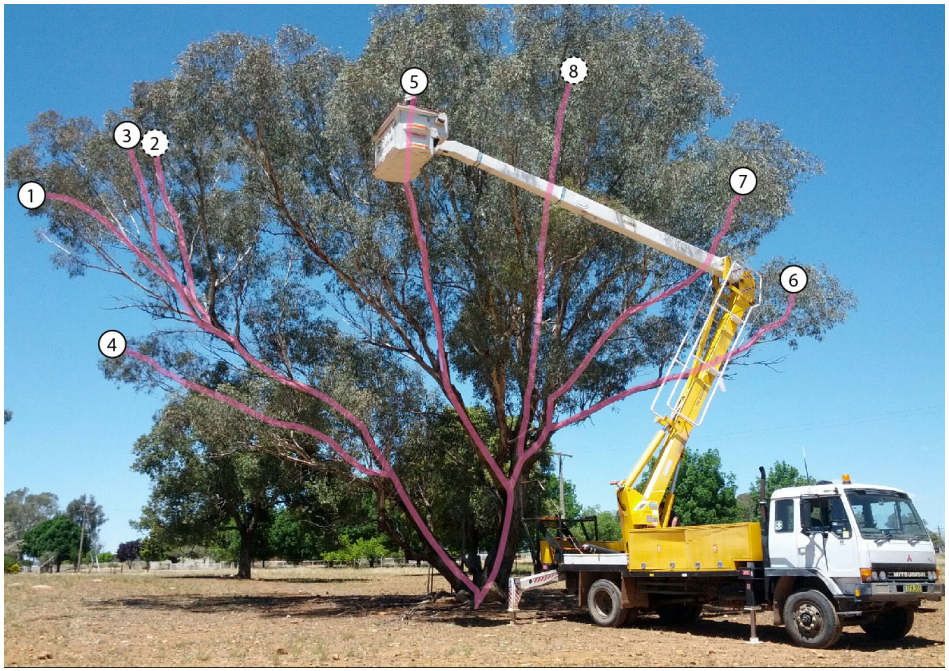


This is a **bubble**. Bubbles of the same size as the kmer size are indicative of a mutation.

Back to *Eucalyptus*

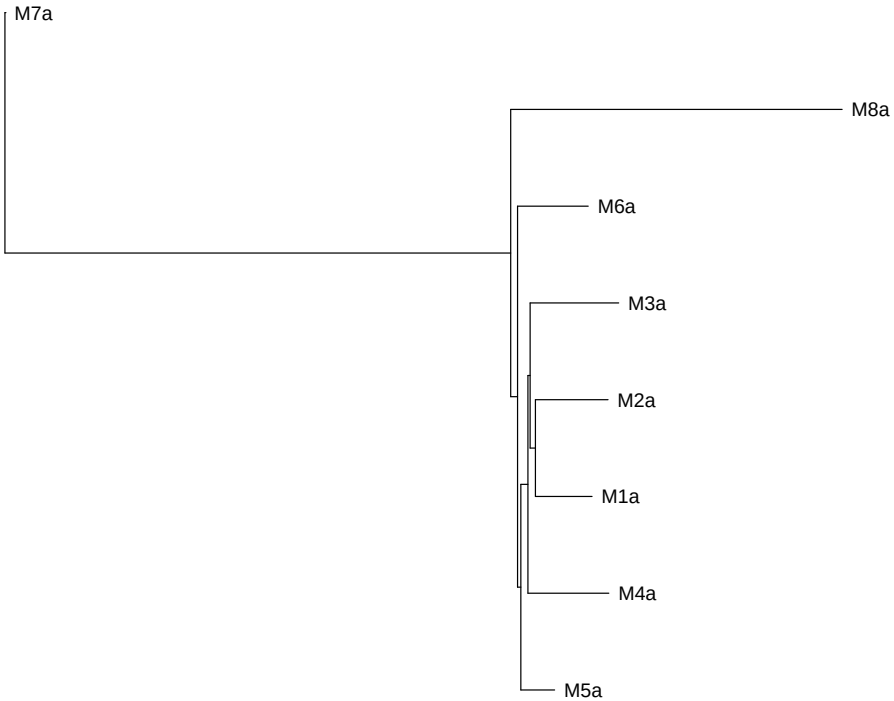
- Reference-free De Bruijn variant caller **DiscoSNP++**
- 8 samples, each with 3 replicates
- Protect against false positive by forcing replicates to agree

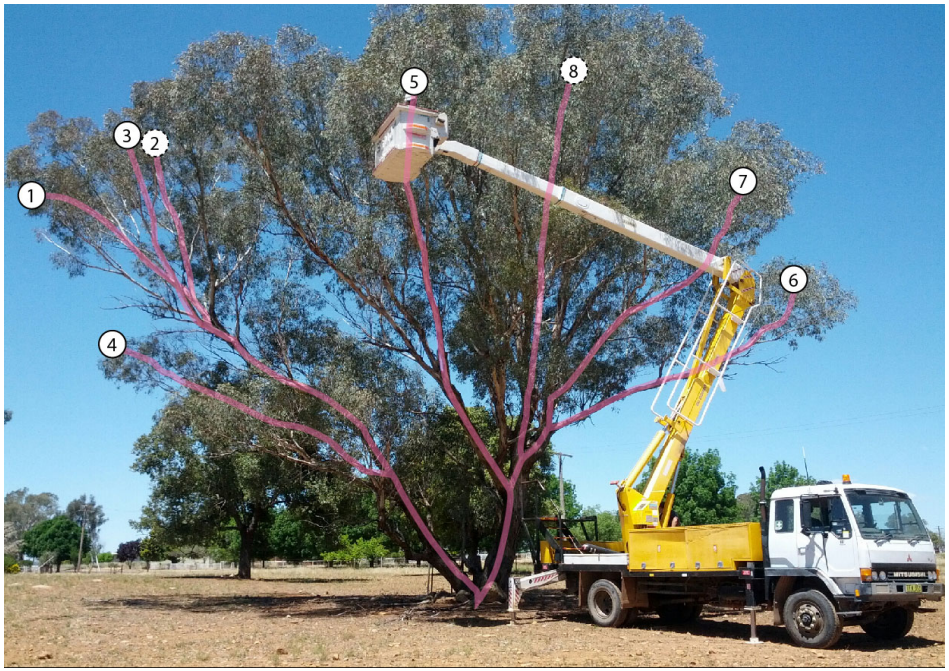




The Genome Analysis Toolkit

- Used a traditional variant-calling pipeline: GATK best practices workflow
- Reference genome of a close relative: *Eucalyptus Grandis*





What's going on here?

Branch 7 is consistently in the wrong place. Why? What's going on with that branch?

- Branch 7 is the longest branch
- May do with assumptions about zygosity
- Recursively improving alignment with various tools

Acknowledgements

- Advisor Reed Cartwright
- Collaborator Robert Lanfear