Using Rooted Triplets to Infer Experimental Error in Cancer Data

Jeff Gaither [1]
Laura Kubatko [2] Julia Chifman [3] Kate Hartmann [4]

[1]Mathematical Biosciences Institute
The Ohio State University
[2]Departments of Statistics and Evolution, Ecology and Organismal Biology, The Ohio State University [3]Department of Mathematics and Statistics, American University [4]The Ohio State University College of Medicine

Workshop in Analytic and Probabilistic Combinatorics
Banff International Research Station
October 26, 2016

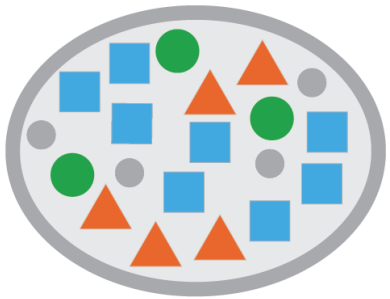
Background

Goal: infer the order of mutations in cancer

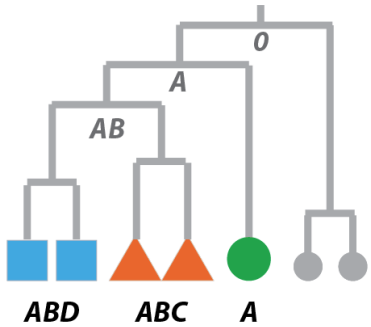
Background

Goal: infer the order of mutations in cancer

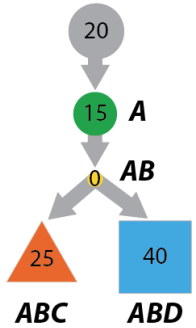
Poly-clonal tumor at sampling



Classical phylogenetic tree

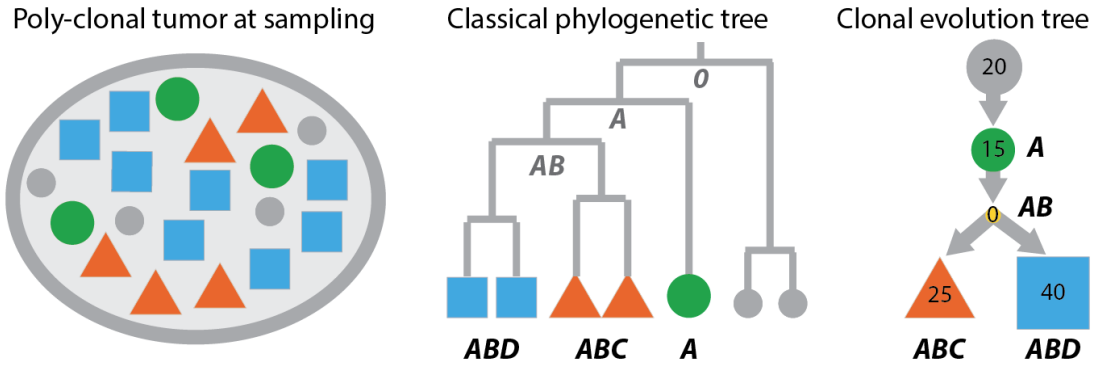


Clonal evolution tree



Background

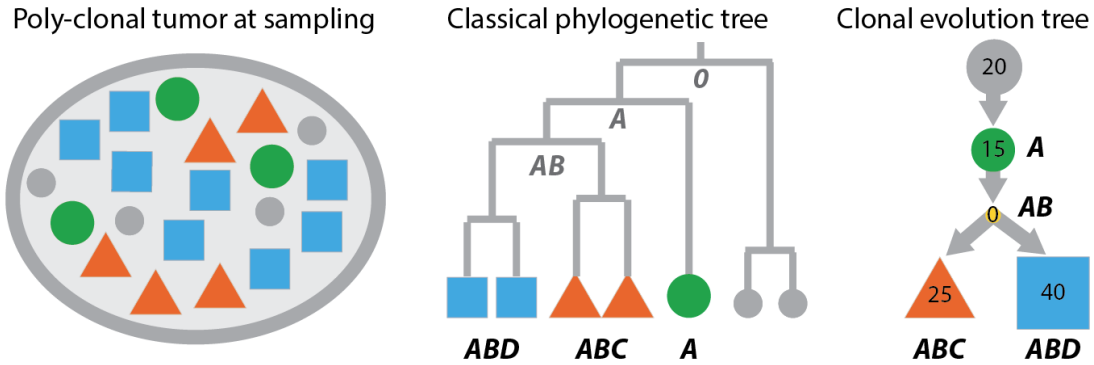
Goal: infer the order of mutations in cancer



Why: gives info on how disease will progress (Ortmann, 2015)

Background

Goal: infer the order of mutations in cancer

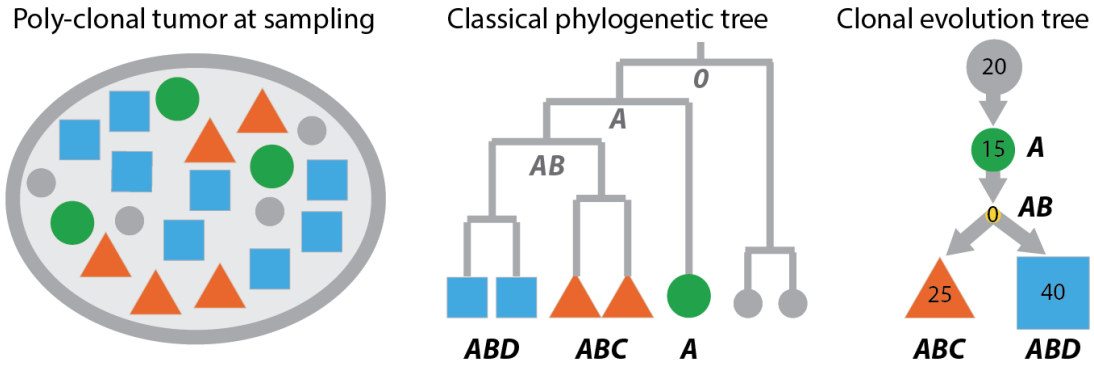


Why: gives info on how disease will progress (Ortmann, 2015)

Challenge: must use single-cell data, which is fraught with error

Background

Goal: infer the order of mutations in cancer



Why: gives info on how disease will progress (Ortmann, 2015)

Challenge: must use single-cell data, which is fraught with error

Result: formulated method to infer and fix false negatives

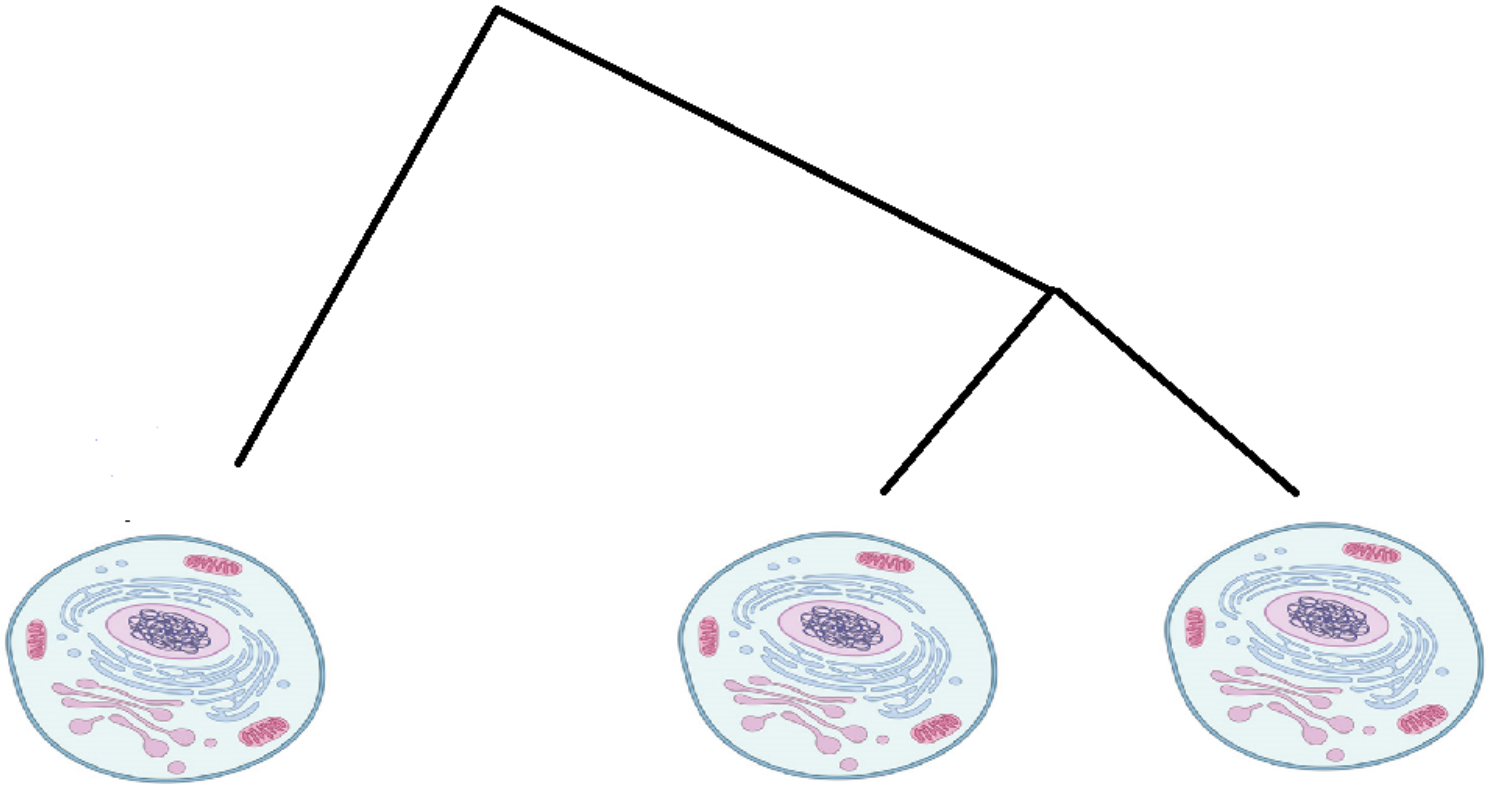
Basic idea

Suppose we have three cells, indexed $c = 1 \ldots 3$, and mutations indexed by j. Let

$$M_{c,j} = \begin{cases} 1 & \text{cell } c \text{ has mutation } j \\ 0 & \text{otherwise.} \end{cases}$$

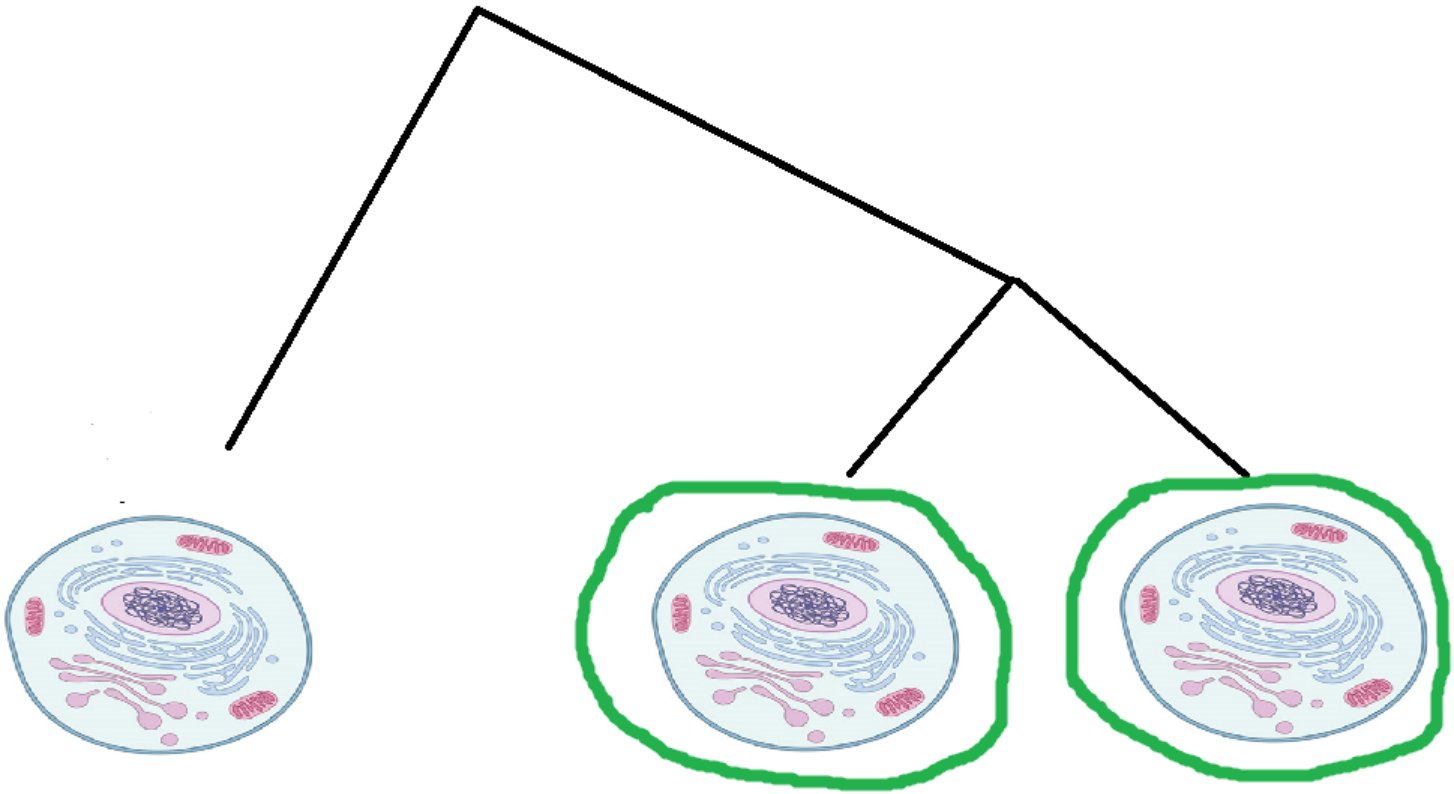
Actual cell lineage

Suppose this is our actual lineage



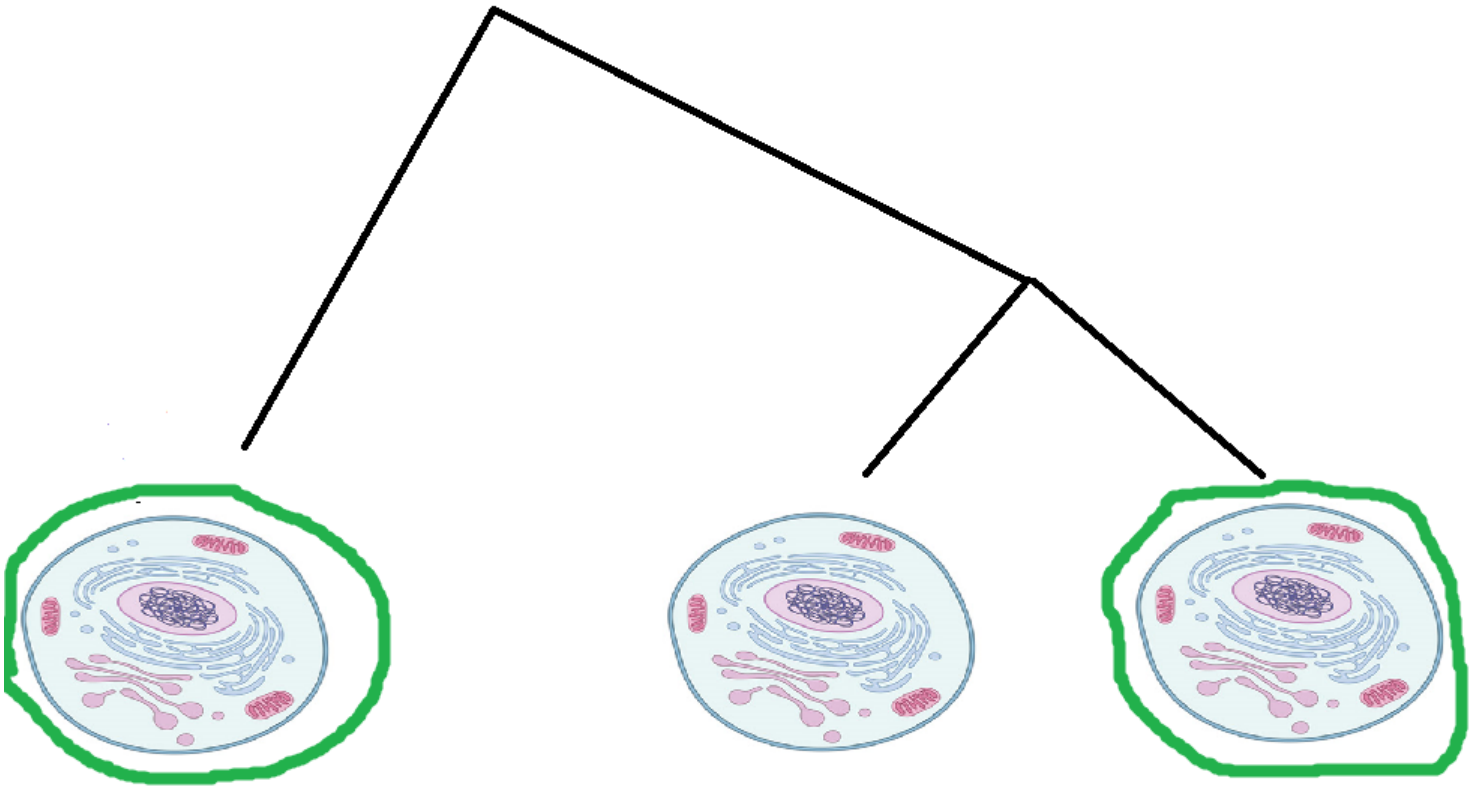
Mutation-pattern #1

$$(M_{1,j}, M_{2,j}, M_{3,j}) = (0, 1, 1)$$



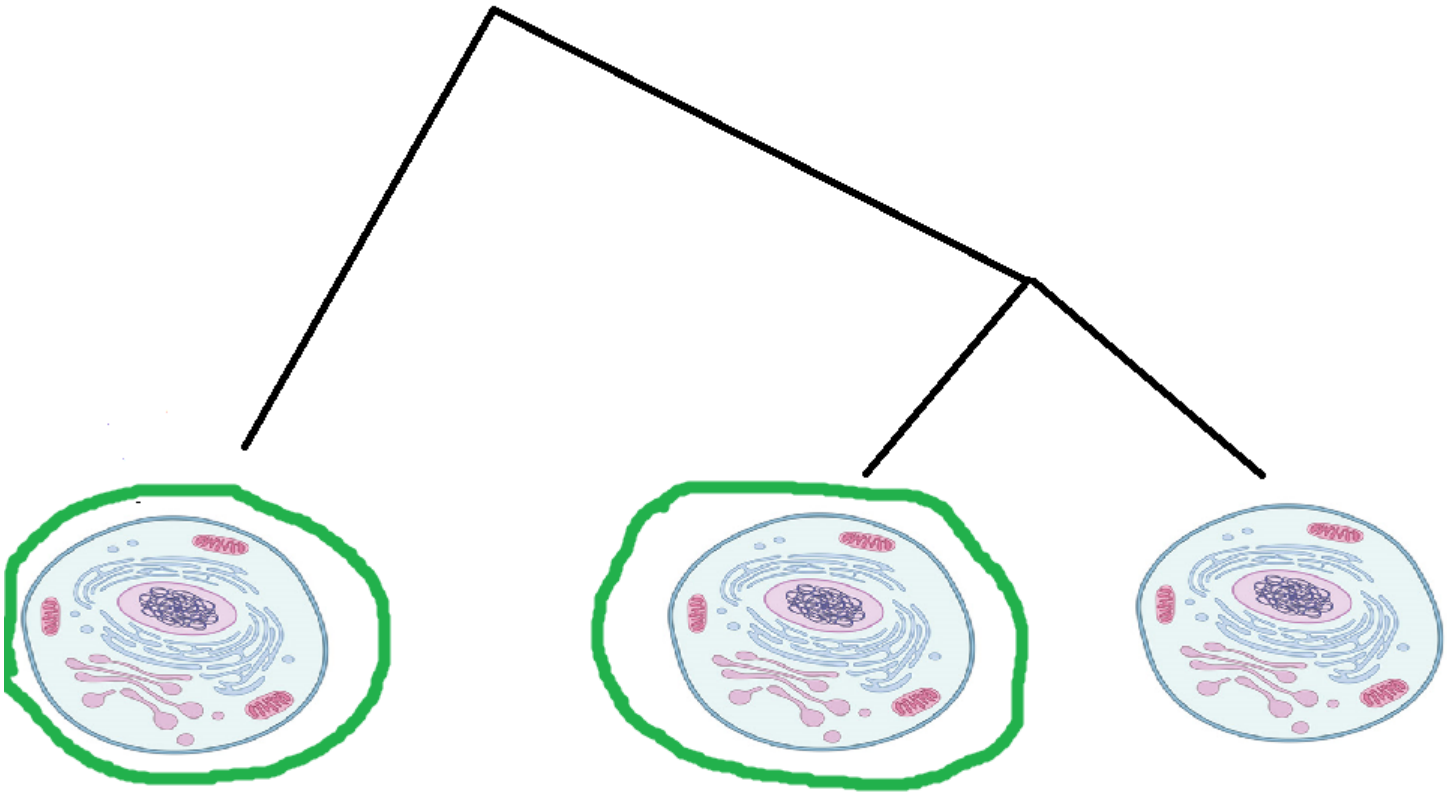
Mutation-pattern #2

$$(M_{1,j}, M_{2,j}, M_{3,j}) = (1, 0, 1)$$



Mutation-pattern #3

$$(M_{1,j}, M_{2,j}, M_{3,j}) = (1, 1, 0)$$



Basic idea

Suppose we have three cells, indexed $c = 1 \ldots 3$, and mutations indexed by j. Let

$$M_{c,j} = \begin{cases} 1 & \text{cell } c \text{ has mutation } j \\ 0 & \text{otherwise.} \end{cases}$$

Then if we see

$$\begin{aligned} (M_{1,j}, M_{2,j}, M_{3,j}) &= (1,1,0) && \text{for tons of } j\text{'s} \\ &= (1,0,1) && \text{for only a few } j\text{'s} \\ &= (0,1,1) && \text{for about as many } j\text{'s as } (1,0,1) \end{aligned}$$

then we conclude that the latter two cases are due to error, and change them to $(1,1,1)$.

Does it work?

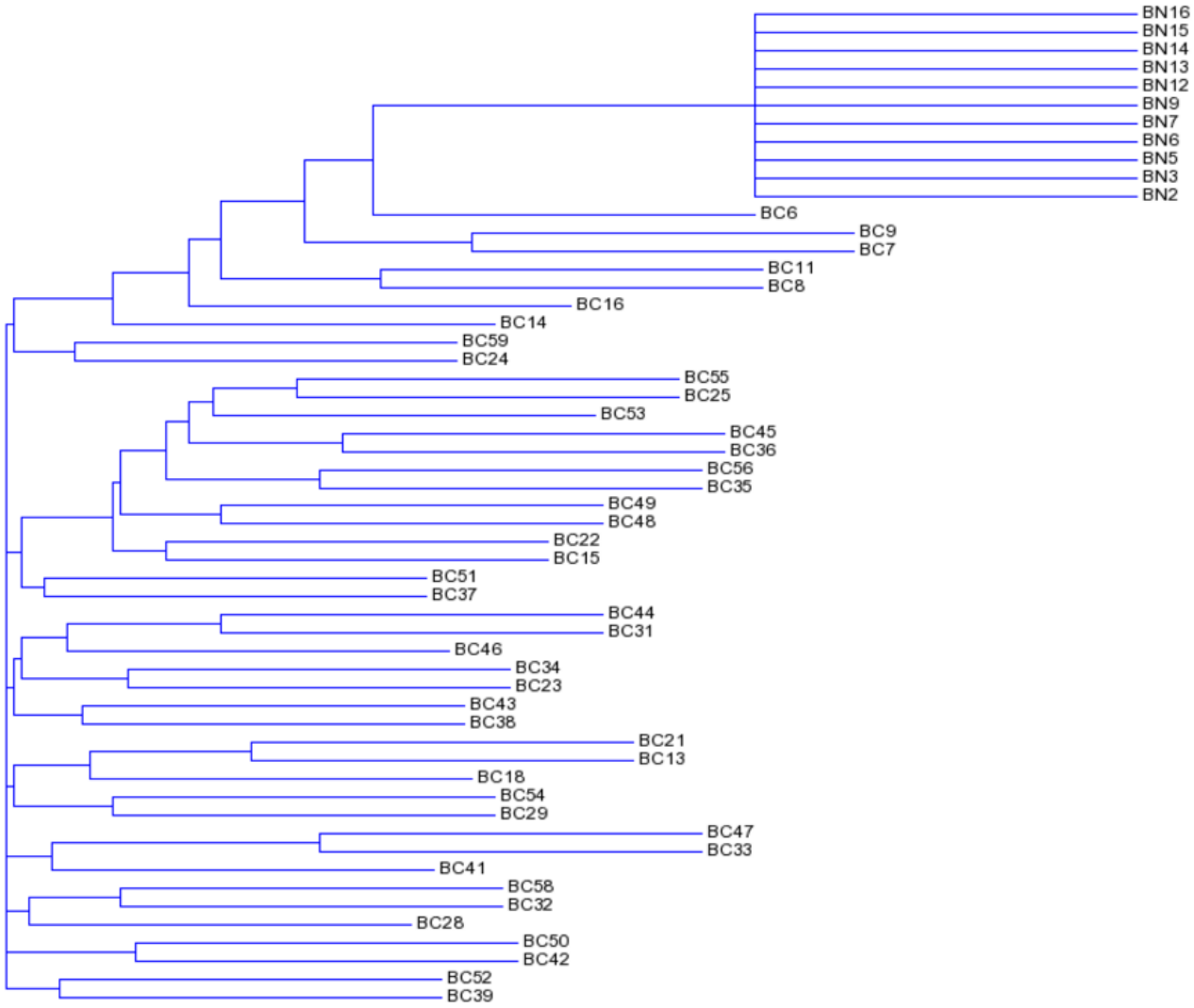
Implement scheme using formal Likelihood ratio test

This actually works – gives us way more confidence in our branches

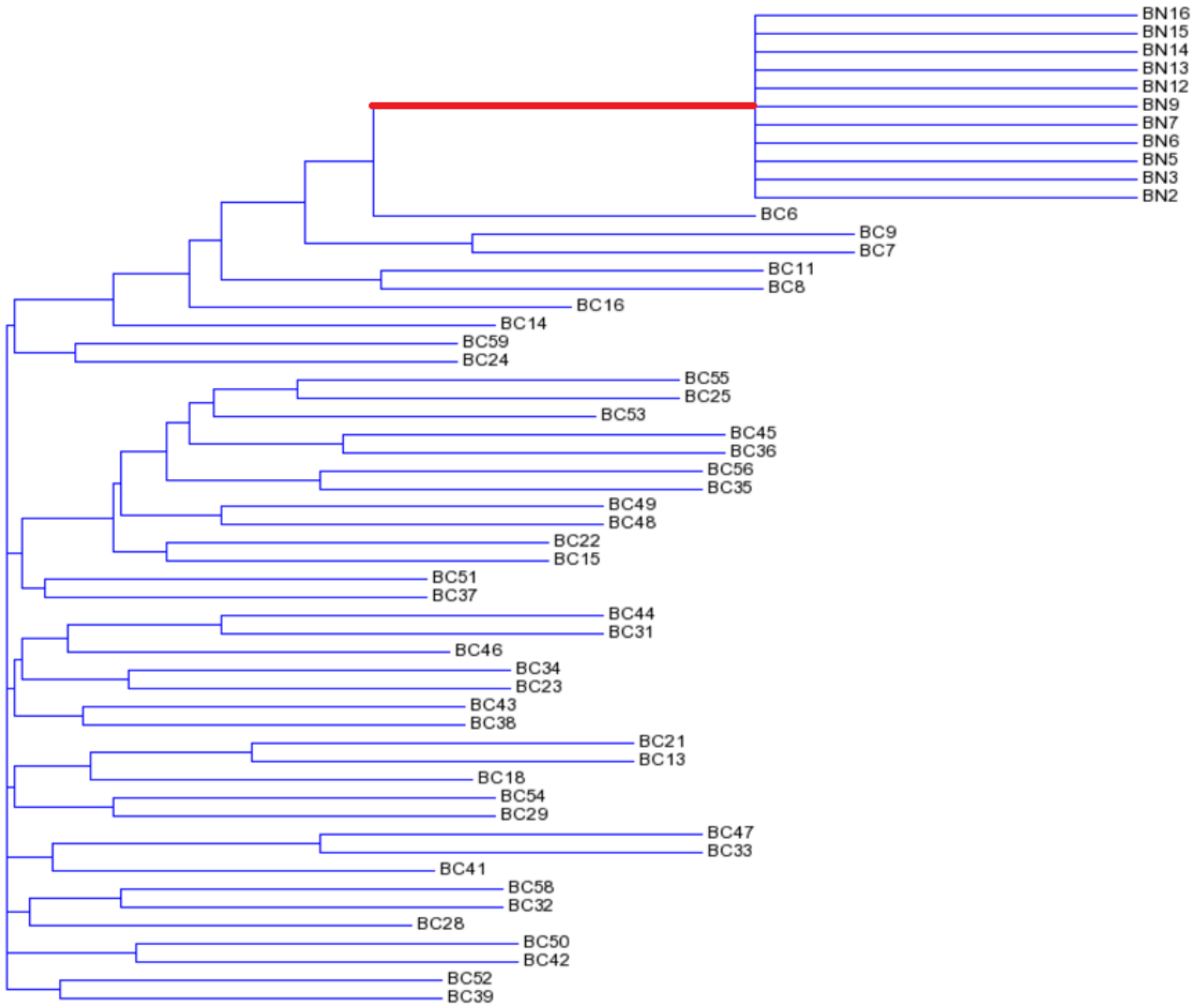
By “confidence,” we mean bootstrap support, i.e. we get same result if we sample the data with replacement

We use single cell data from bladder cancer study (Li et al, 2012)

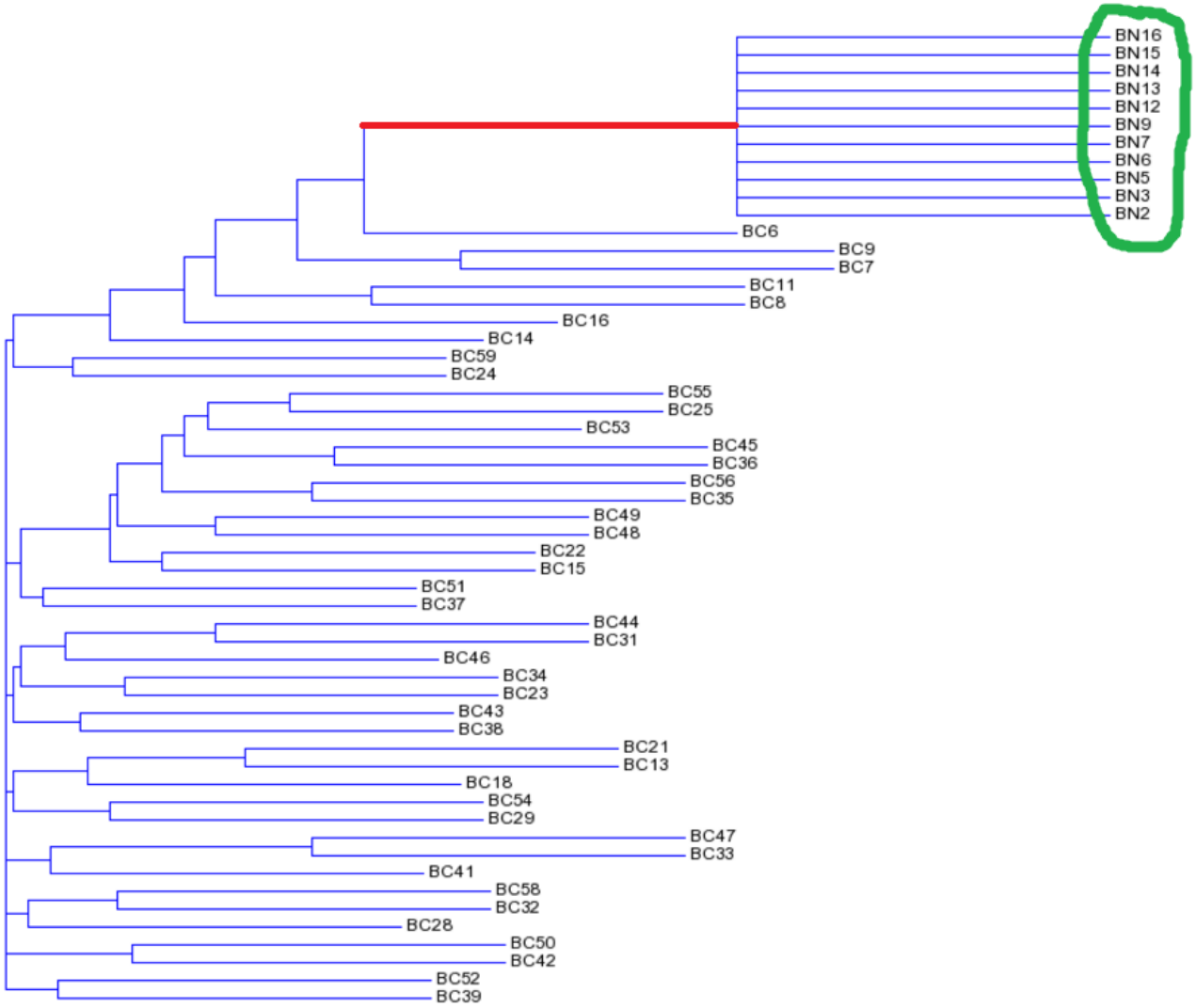
Tree built from unprocessed data



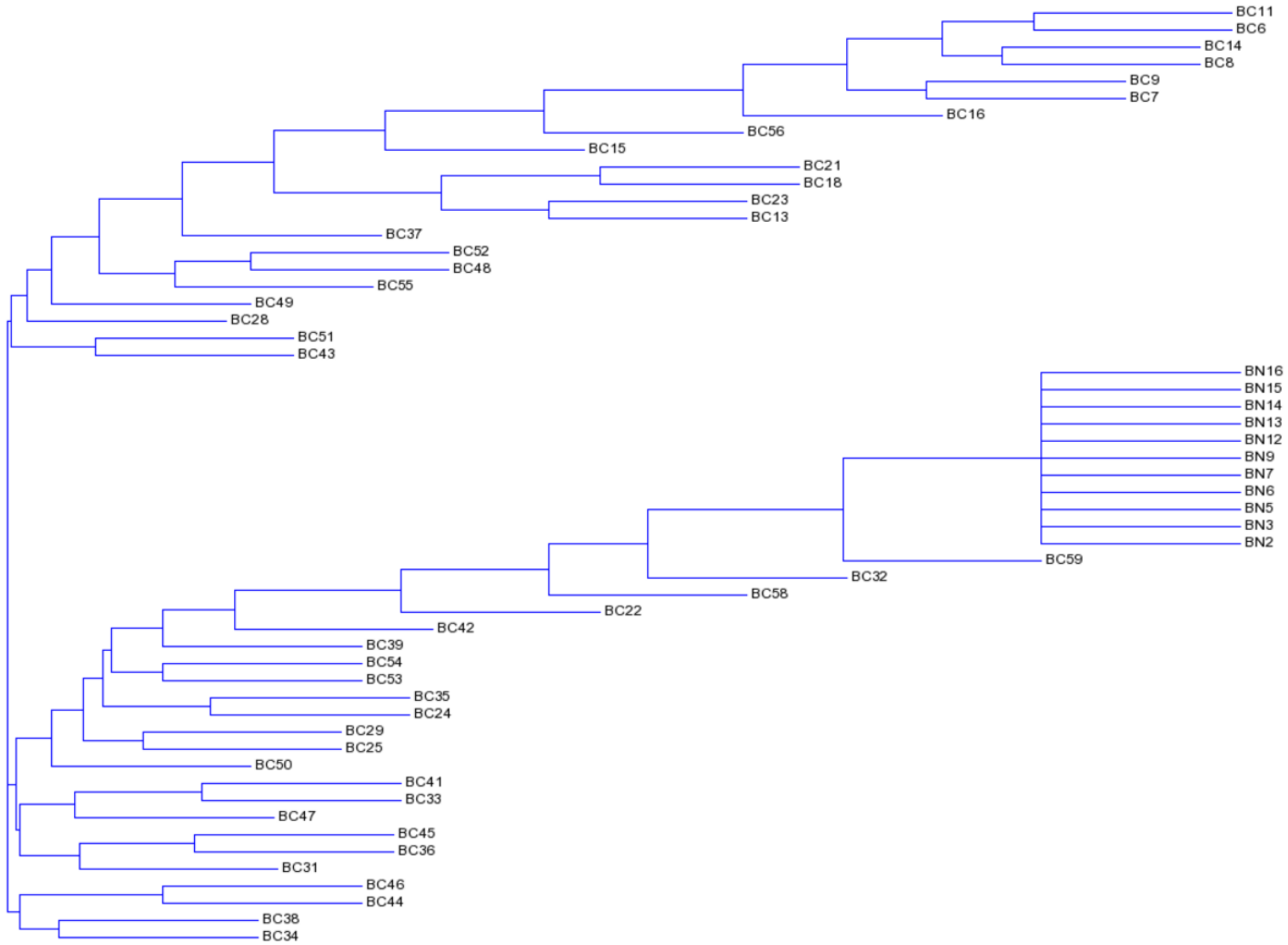
Only one branch has 80% support



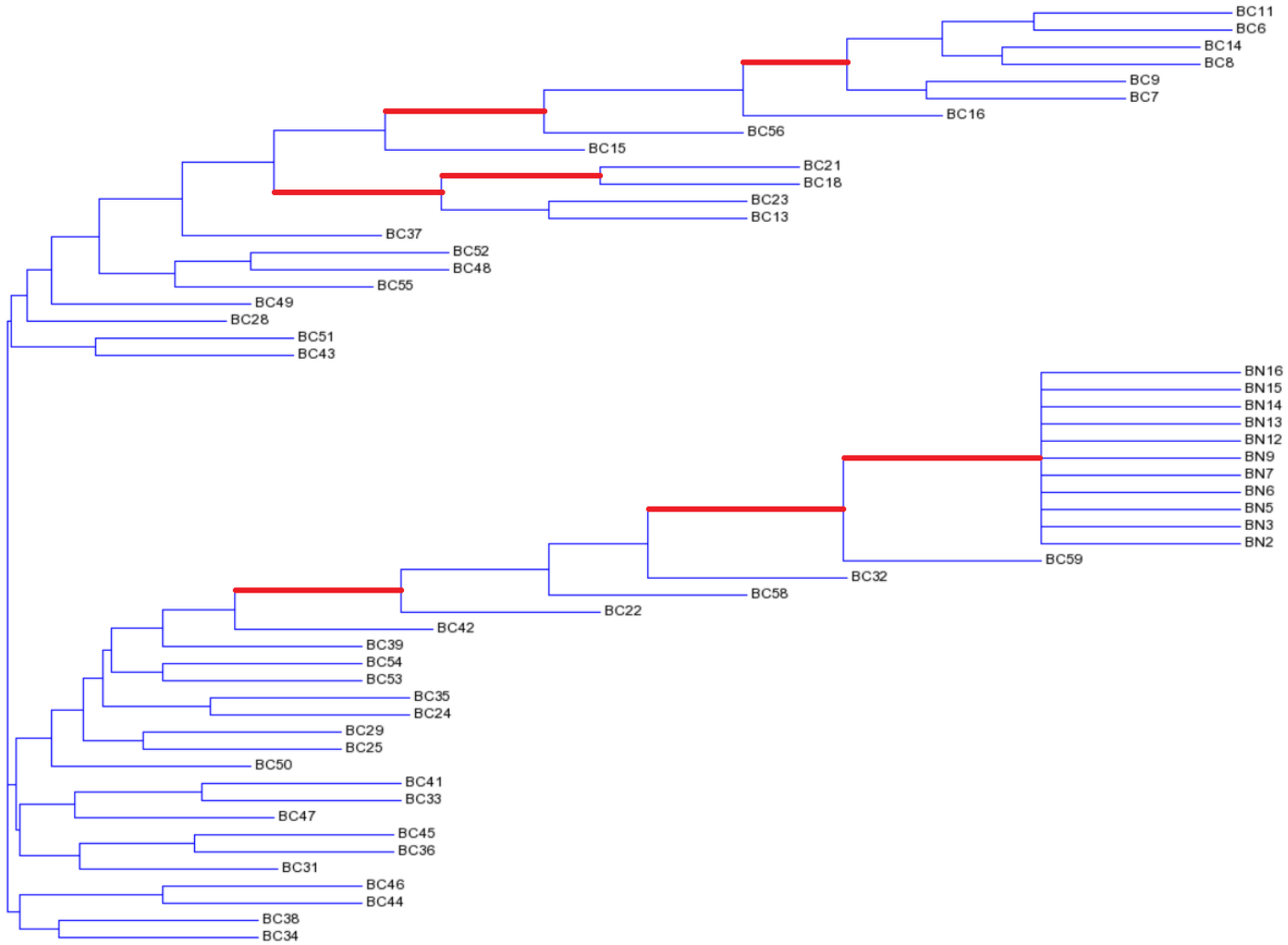
Only clone we ID is the noncancerous cells



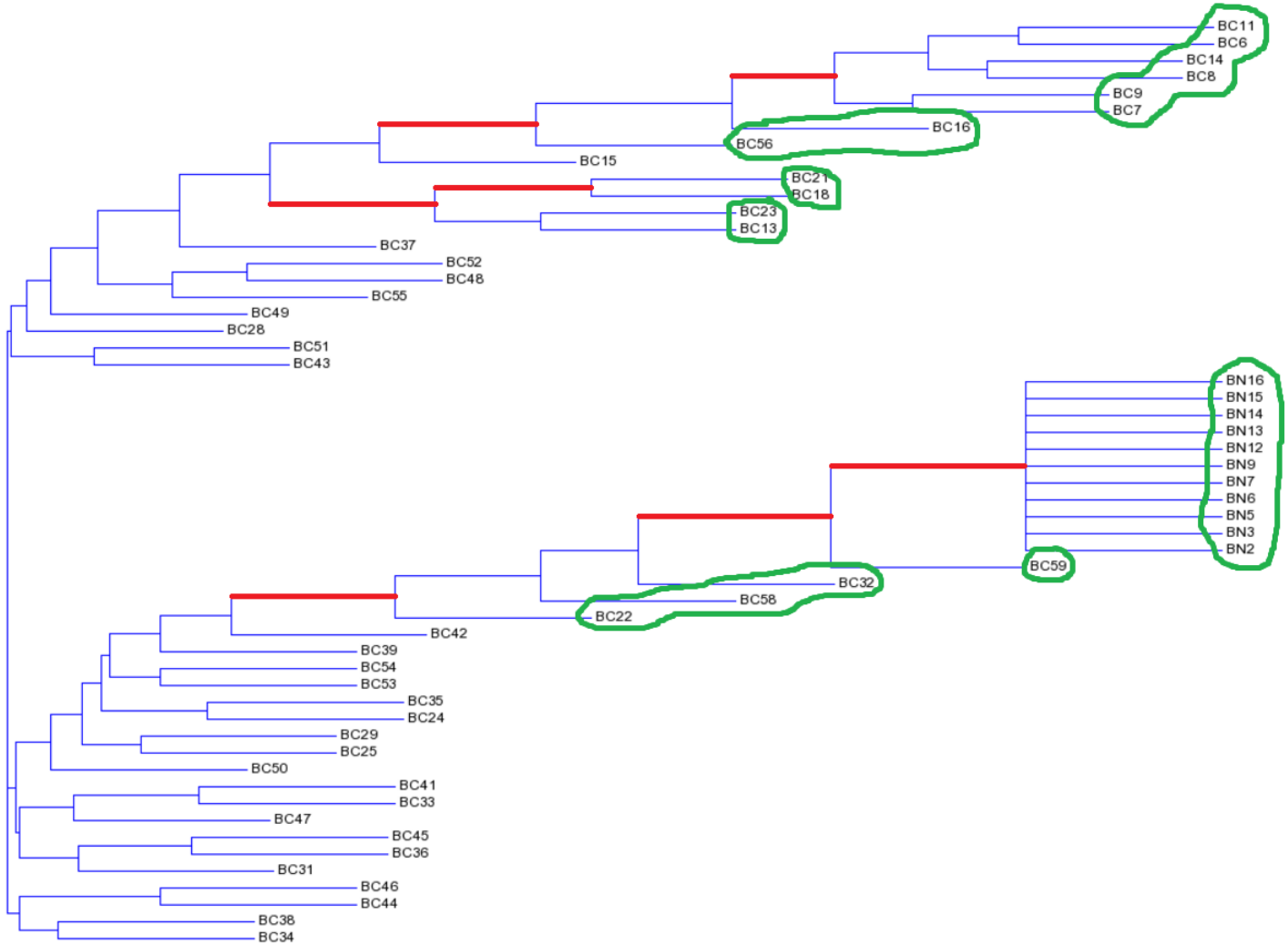
Apply method to data, get new tree



Several well-supported branches

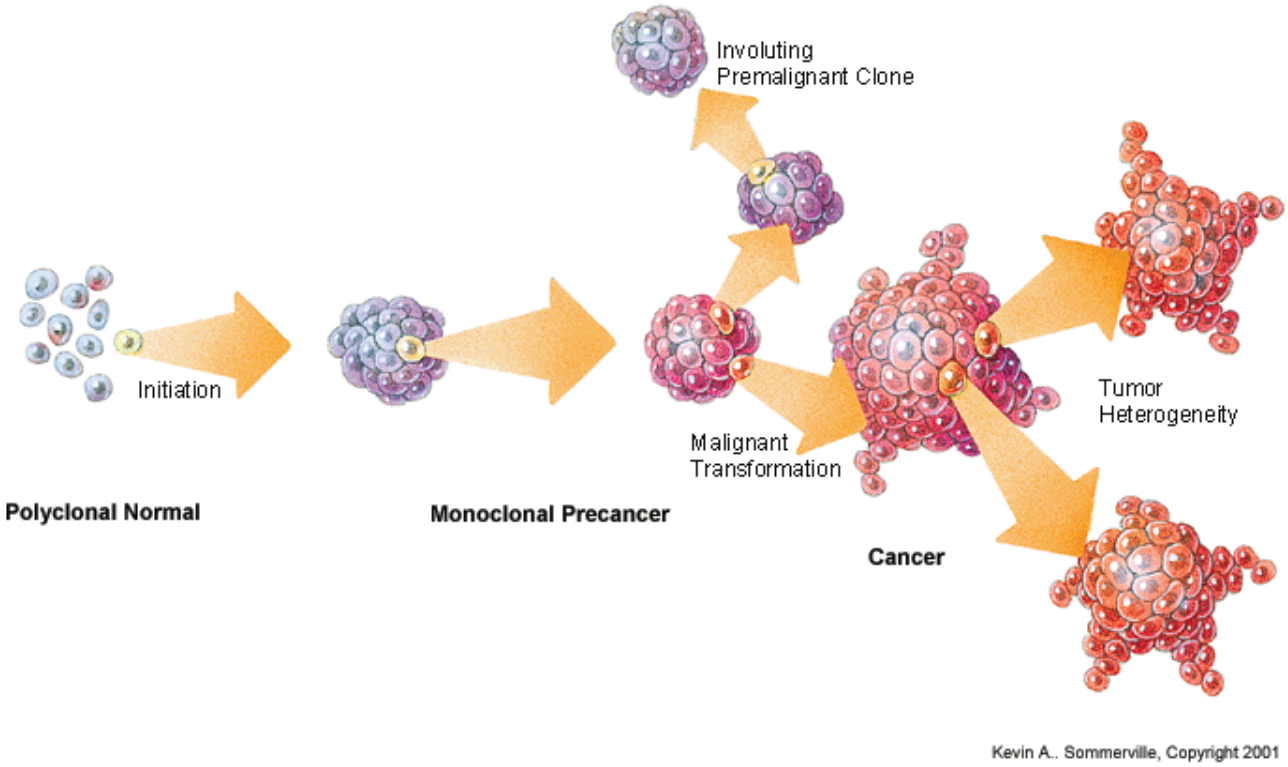


Thus, several well-supported clones



Background - Clones

Cancer evolves so fast in a body that different species (called *clones*) emerge



We'd like to characterize these clones, infer their order of appearance

Methodology: phylogenetics

We can track the evolutionary lineage of these clones using phylogenetics

Phylogenetics is the discipline that infers evolutionary history from genomic data

Many histories are *possible* – we try to infer the most likely history, which we express as a tree, given data

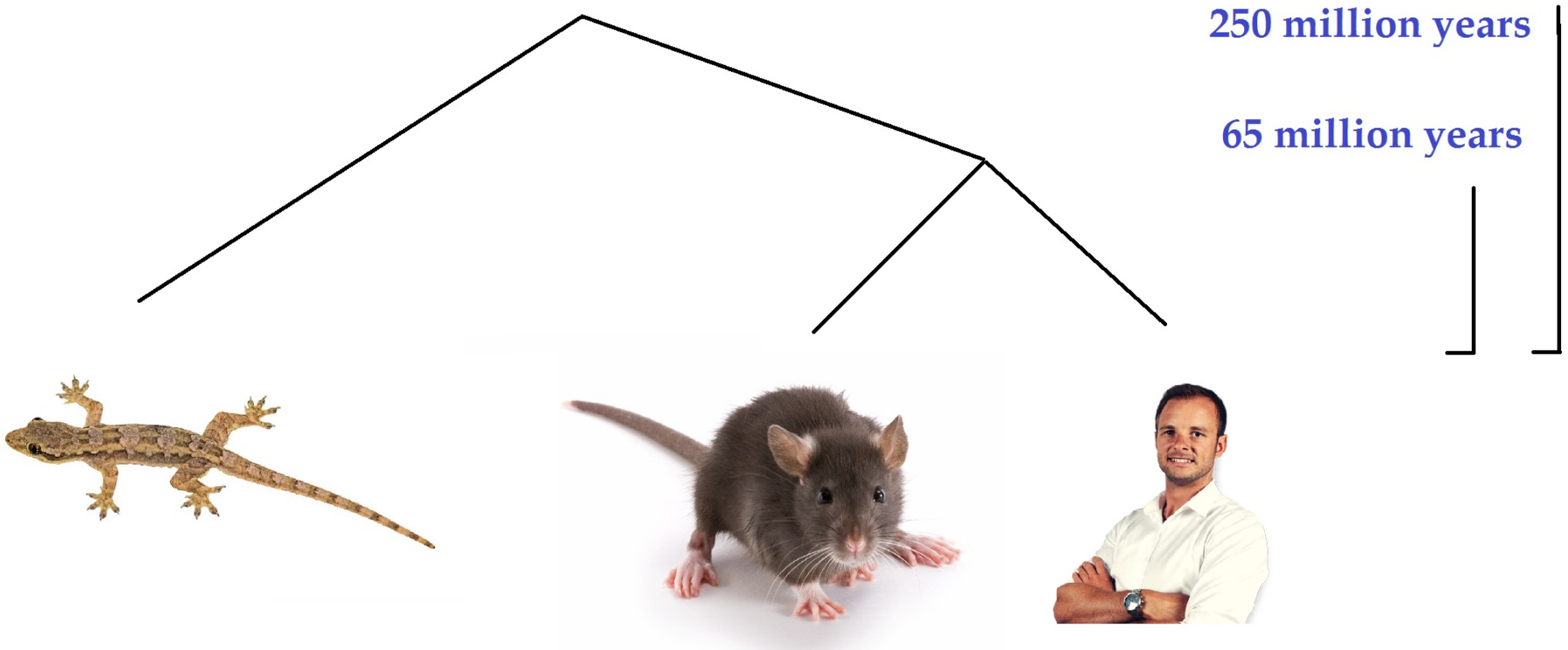
Data – Toy example of phylogenetic matrix

	Site:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	Species	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	Pre (Chimp)	C	T	T	G	A	G	A	A	A	A	T	T	C	T
2	Pme (Lizard)	T	C	T	A	A	A	A	G	A	T	T	A	T	A
3	Pma (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T
4	Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T
5	Pbe (Rodent)	T	T	T	A	A	G	A	A	A	A	T	T	T	A
6	Plo (Bird)	T	T	T	A	A	G	A	A	A	A	C	T	C	A
7	Pfr (Monkey)	C	T	T	A	A	G	A	A	G	A	T	T	C	T
8	Pkn (Monkey)	C	T	T	A	A	G	A	A	A	G	T	T	C	T
9	Pcy (Monkey)	C	T	C	A	T	G	A	A	A	A	T	T	C	T
10	Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	T	C	T
11	Pga(Bird)	T	T	T	A	A	G	A	A	A	A	T	T	T	T

How to infer species tree from data

Suppose true species tree looks like this

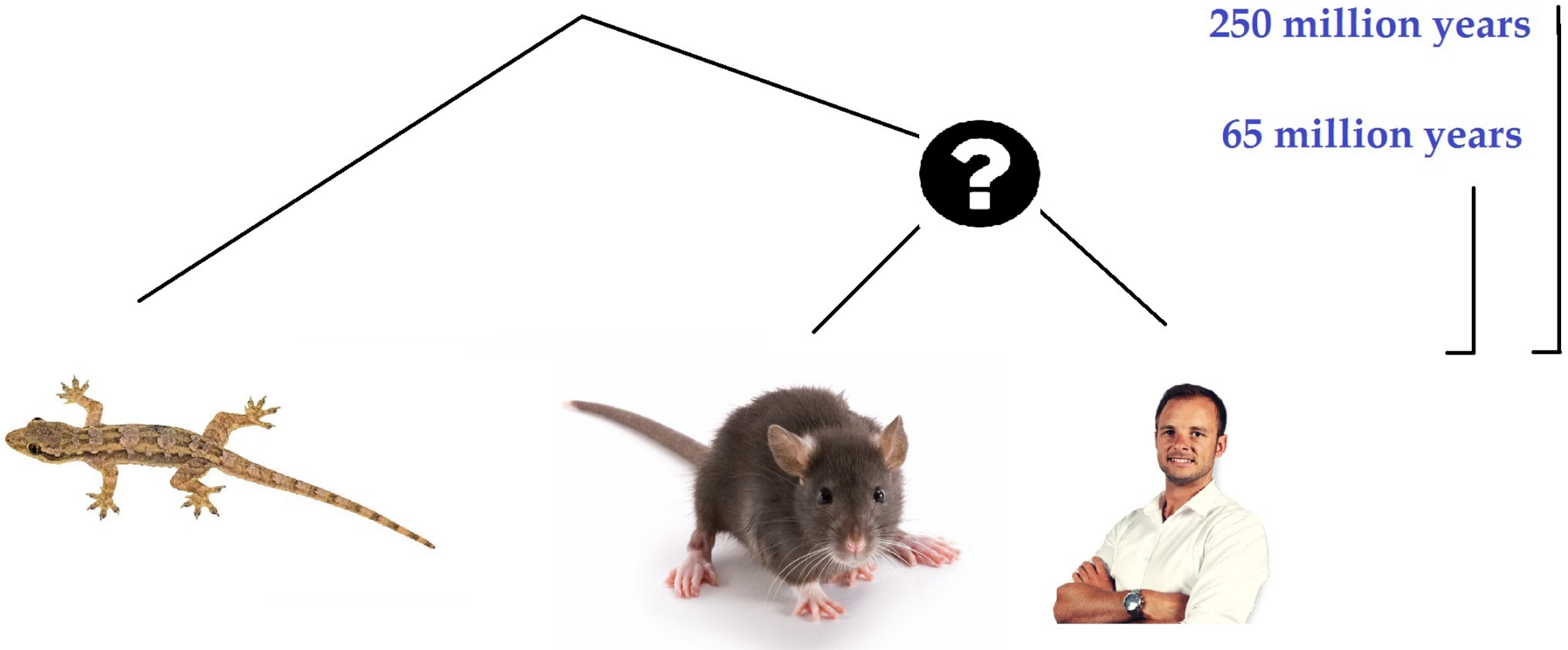
Last common ancestor of mice and humans lived 65 million years ago



How to infer species tree from data

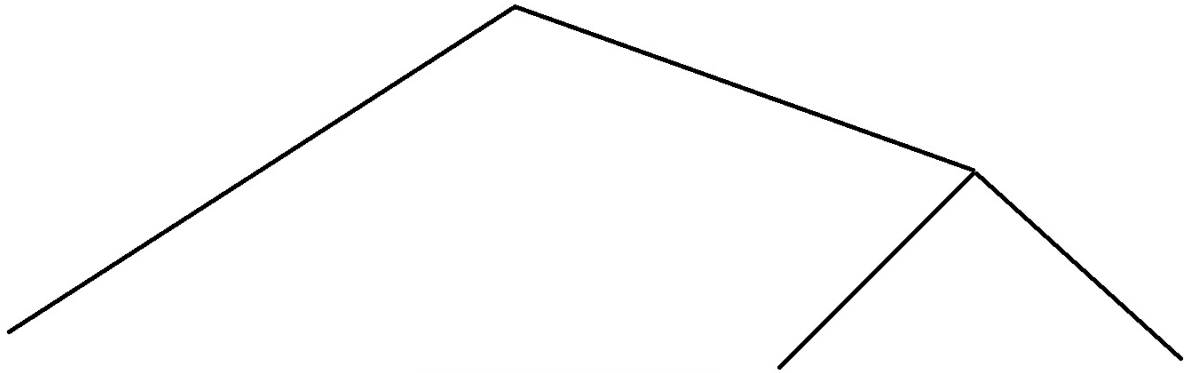
Suppose true species tree looks like this

Last common ancestor of that ancestor and lizards lived 250 million years ago



How to infer species tree from data

Our goal is to infer this tree, given data.



250 million years

65 million years

How to infer species tree from data

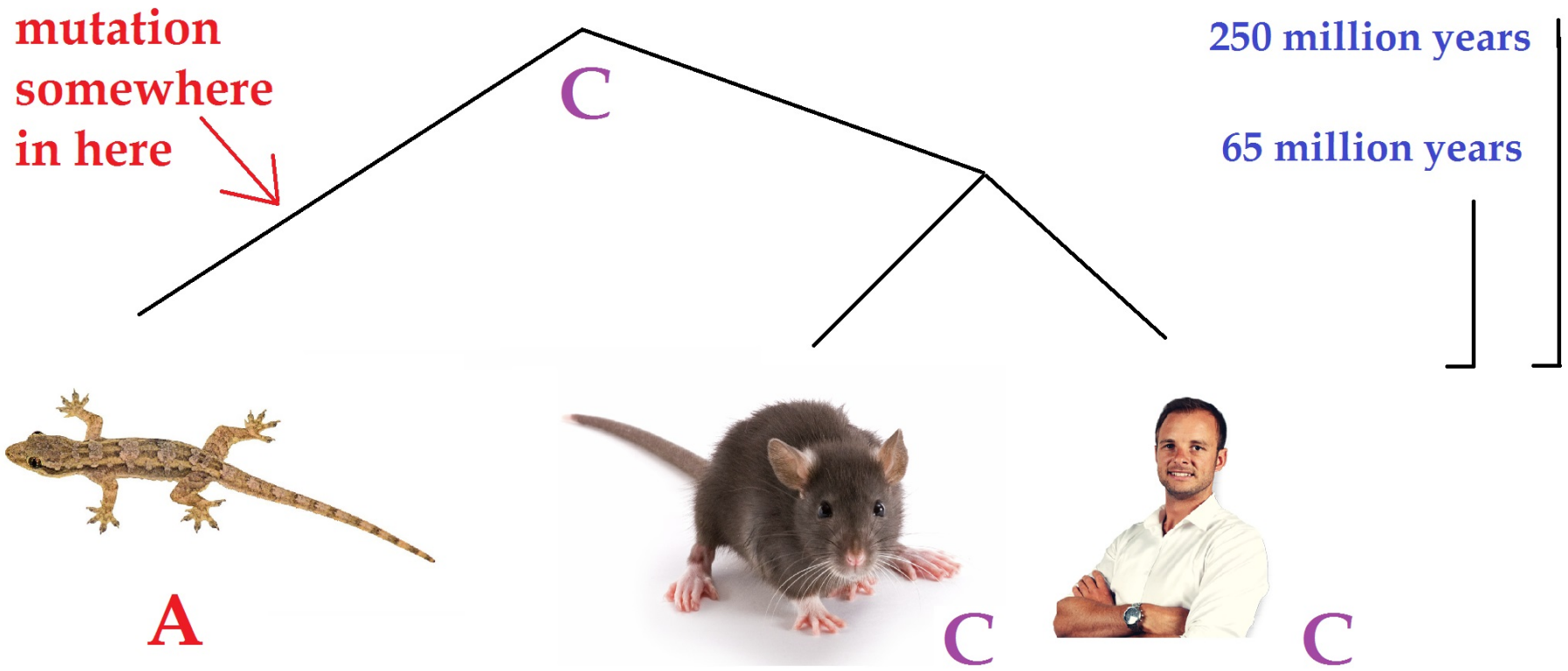
Let L_j, M_j and H_j be values of lizard, mouse and human DNA at site i
Then we expect to see

$$\begin{aligned}(M_j, L_j, H_j) &= (X, Y, Y) && \text{for tons of } j\text{'s} \\ &= (Y, X, Y) && \text{for only a few } j\text{'s} \\ &= (Y, Y, X) && \text{for about as many } j\text{'s as } (Y, X, Y)\end{aligned}$$

Let's consider (A, C, C) vs. (A, A, C).

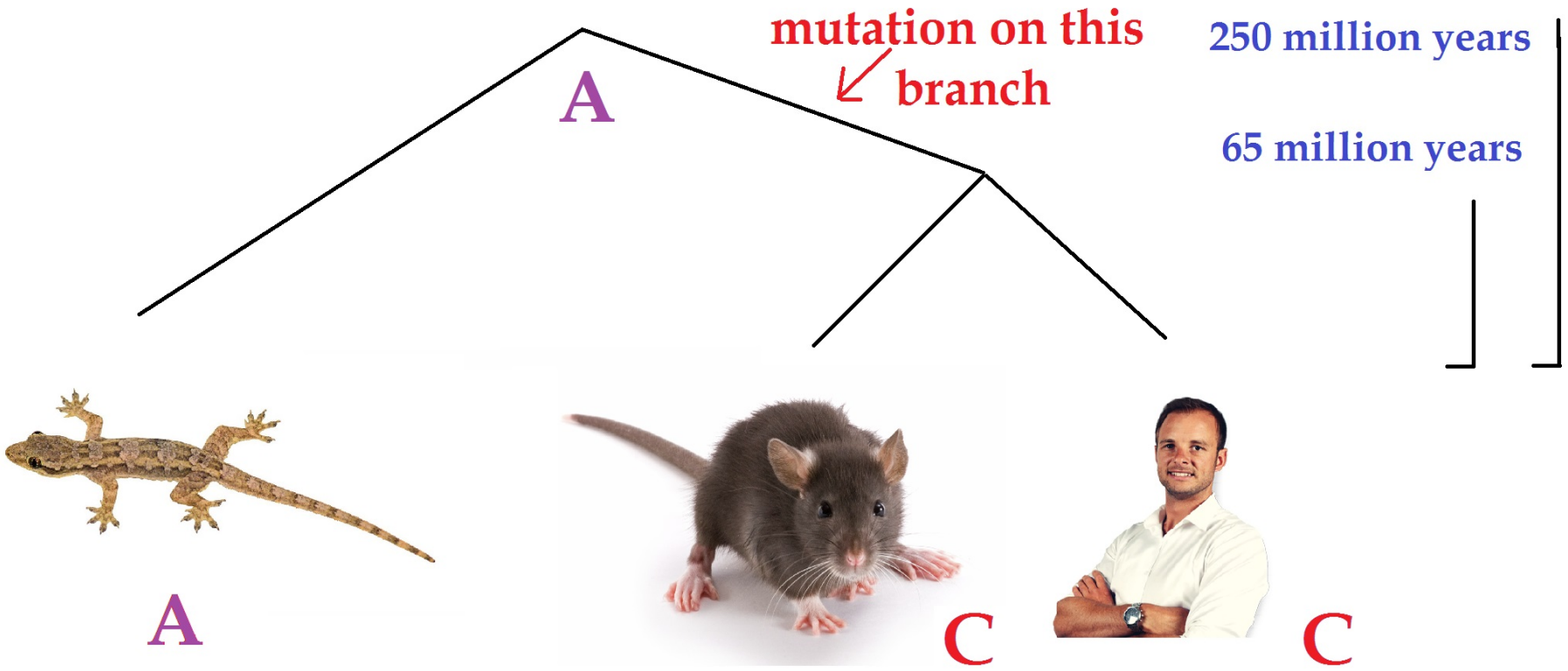
How to infer species tree from data

How could we get (A,C,C)? Mutation on long left branch, or upper portion of right



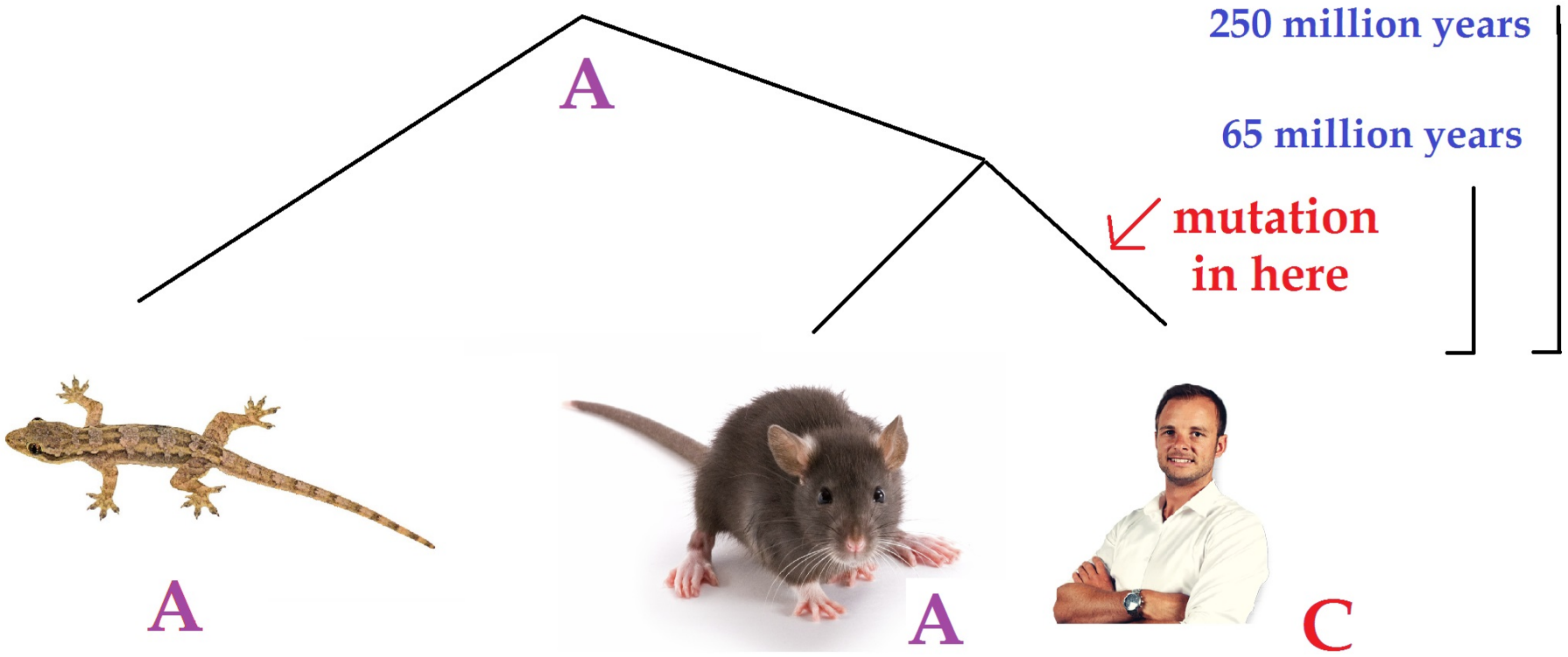
How to infer species tree from data

How could we get (A,C,C)? Mutation on long left branch, or upper portion of right



How to infer species tree from data

Whereas (A,A,C) could only arise from a mutation on the short rightmost branch



The big picture

So, if our genetic matrix

Phylogenetic matrix

	Site:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	Species	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	Pre (Chimp)	C	T	T	G	A	G	A	A	A	A	T	T	C	T
2	Pme (Lizard)	T	C	T	A	A	A	A	G	A	T	T	A	T	A
3	Pma (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T
4	Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T
5	Pbe (Rodent)	T	T	T	A	A	G	A	A	A	A	T	T	T	A
6	Plo (Bird)	T	T	T	A	A	G	A	A	A	A	C	T	C	A
7	Pfr (Monkey)	C	T	T	A	A	G	A	A	G	A	T	T	C	T
8	Pkn (Monkey)	C	T	T	A	A	G	A	A	A	G	T	T	C	T
9	Pcy (Monkey)	C	T	C	A	T	G	A	A	A	A	T	T	C	T
10	Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	T	C	T
11	Pga (Bird)	T	T	T	A	A	G	A	A	A	A	T	T	T	T

The big picture

So, if our genetic matrix shows more sites with

$$(L_i, M_i, H_i) = (X, Y, Y)$$

than

$$(L_i, M_i, H_i) = (X, X, Y), \quad \text{or}$$
$$(L_i, M_i, H_i) = (X, Y, X)$$

then we can conclude that humans and mice are closer kin to each other than either is to lizards.

Fantasyland

This basic paradigm would be perfect, IF

Fantasyland

This basic paradigm would be perfect, IF

1. A given site never mutated more than once

Fantasyland

This basic paradigm would be perfect, IF

1. A given site never mutated more than once
2. Mutation rates were invariant across branches

Fantasyland

This basic paradigm would be perfect, IF

1. A given site never mutated more than once
2. Mutation rates were invariant across branches
3. The sequence data was never wrong

Fantasyland

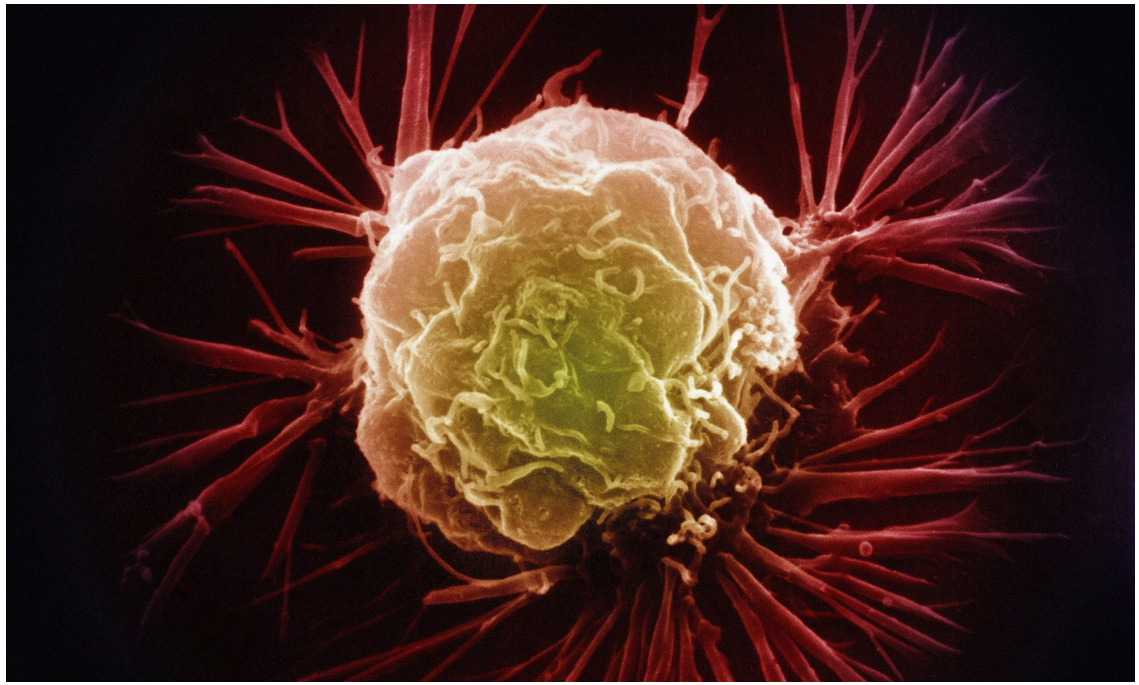
This basic paradigm would be perfect, IF

1. A given site never mutated more than once
2. Mutation rates were invariant across branches
3. The sequence data was never wrong

However, all these assumptions are violated, and they are violated especially frequently in cancer.

Cancer in a nutshell

Cancer begins when one cell rebels



This cell and its descendants lose all interest in *you*, start stealing resources and proliferating like crazy

Cancer genomics

So we get different species (*clones*) in a single tumor

Two techniques for analyzing genomes: *shotgun sequencing* and *single cell sequencing*

Shotgun sequencing – infer genetic character of region by taking consensus from many adjacent cells.

Cancer genomics

So we get different species (*clones*) in a single tumor

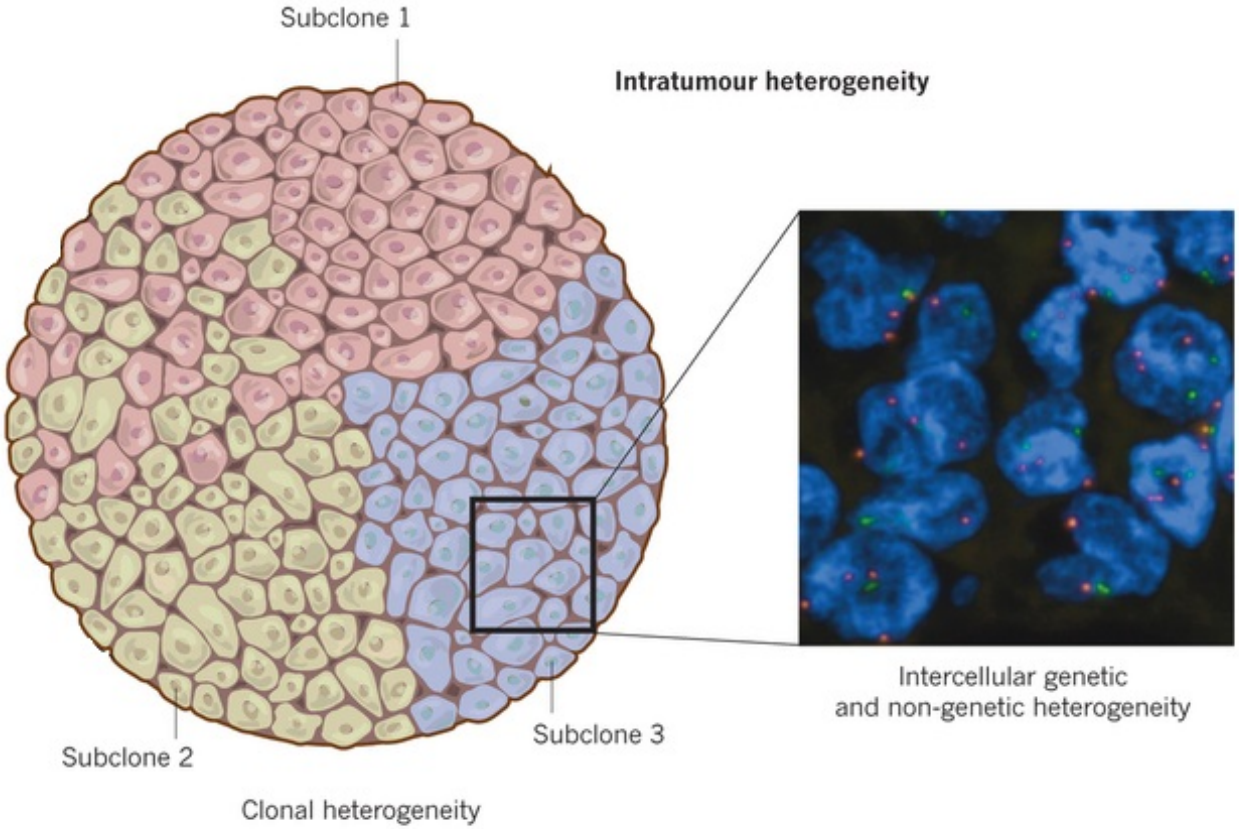
Two techniques for analyzing genomes: *shotgun sequencing* and *single cell sequencing*

Shotgun sequencing – infer genetic character of region by taking consensus from many adjacent cells.

When would shotgun sequencing be adequate? If nearby cancer cells had roughly same genetic makeup.

Shotgun sequencing adequate?

This totally isn't true – tumor heterogeneity. Cells right next to each other might have different genetic makeup.

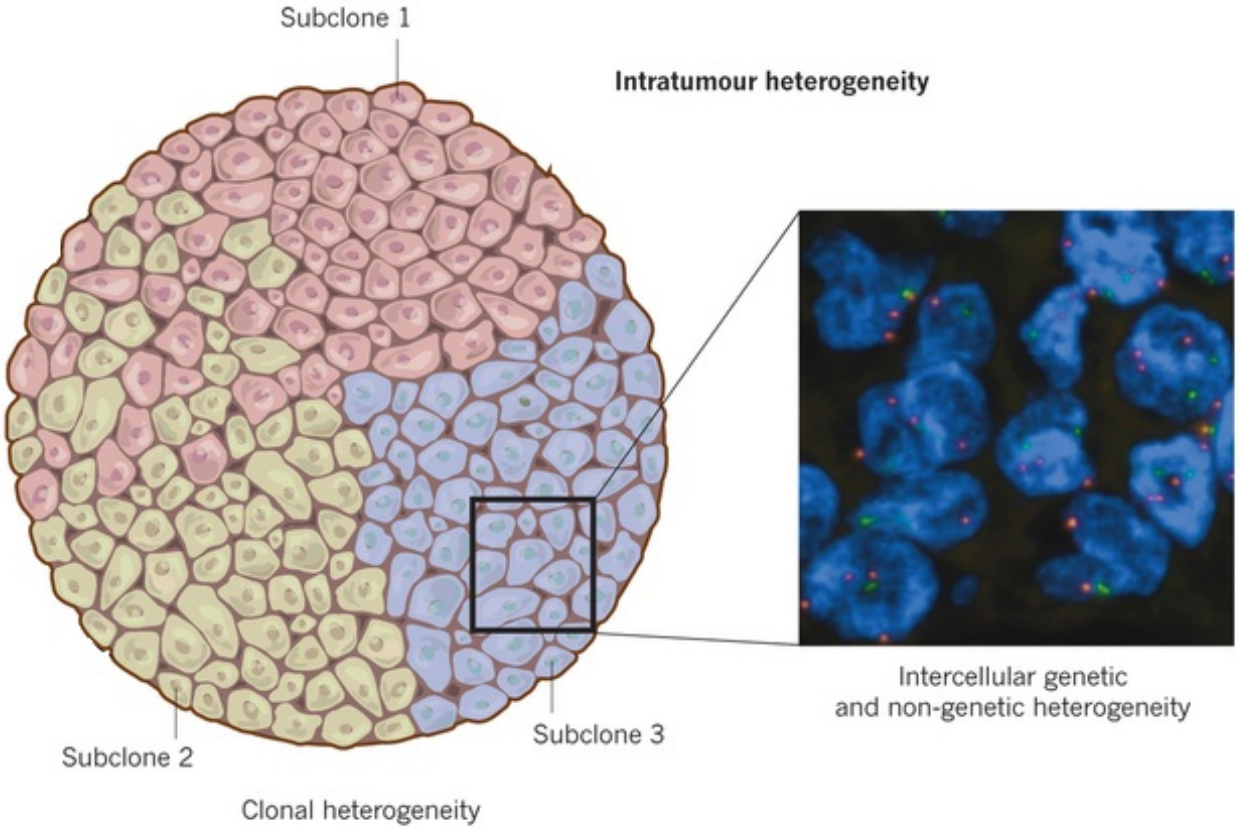


Battle

Because the clones are at war with each other

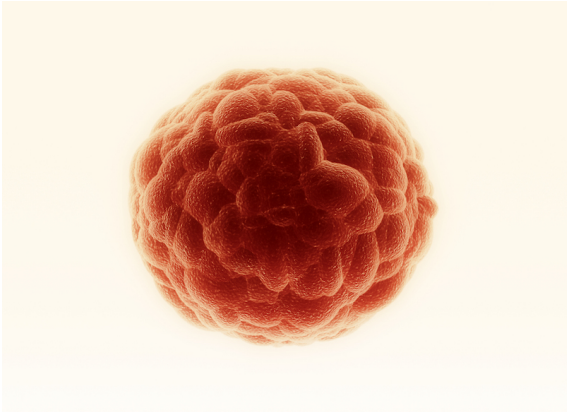
Clonal trees

So, shotgun sequencing can build clonal trees – take consensus as a clone, work out conflicts – but to actually infer order of particular mutations, we really need single-cell data.



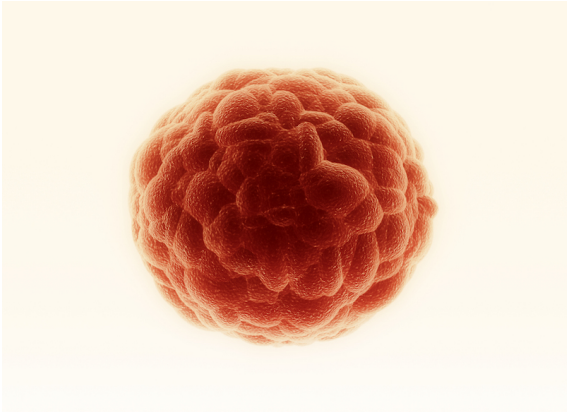
Single-cell data

Single cell-data consists of sampling the genome of a single cancer cell.



Single-cell data

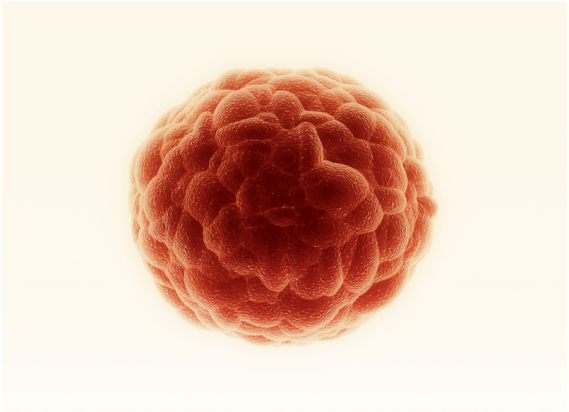
Single cell-data consists of sampling the genome of a single cancer cell.



Well duh, that's obviously better. Why doesn't everyone do that? Because it's

Single-cell data

Single cell-data consists of sampling the genome of a single cancer cell.



Well duh, that's obviously better. Why doesn't everyone do that? Because it's

- expensive
- extremely error-prone

Perils of Single-cell sequencing

We're not directly hindered by "expensive," since we're not experimentalists. But what does "extremely error-prone" mean? In effect, two things:

Perils of Single-cell sequencing

We're not directly hindered by "expensive," since we're not experimentalists. But what does "extremely error-prone" mean? In effect, two things:

1. We get a lot less data – not every site for every cell.
2.

Some single-cell data from a bladder cancer

	A	B	C	D	E	F	G	H	I	J	K
1		Li's somatic mutations, with mutated heterozygotes represented only as new base									
2											
3		BC-6	C	-	T	A	-	-	C	T	-
4		BC-7	G	-	C	-	T	-	C	T	C
5		BC-8	G	T	T	-	T	T	T	T	C
6		BC-9	-	-	C	G	T	C	C	T	C
7		BC-11	-	-	C	-	-	C	C	-	-
8		BC-13	C	-	C	-	T	C	-	T	-
9		BC-14	G	-	C	G	T	C	C	T	C
10		BC-15	-	-	C	-	T	C	C	T	-
11		BC-16	-	-	-	G	T	C	C	T	-
12		BC-18	-	-	-	G	-	C	C	T	C
13		BC-21	-	-	C	-	-	T	C	T	C
14		BC-22	-	-	-	G	T	T	-	-	C
15		BC-23	-	-	C	G	-	C	C	T	C
16		BC-24	-	-	C	G	-	C	C	-	G
17		BC-25	-	-	C	G	-	-	C	T	C
18		BC-28	-	T	C	G	-	T	C	T	C
19		BC-29	-	-	-	-	T	-	C	T	C

Perils of Single-cell sequencing

We're not directly hindered by "expensive," since we're not experimentalists. But what does "extremely error-prone" mean? In effect, two things:

1. We get a lot less data – not every site for every cell.

Perils of Single-cell sequencing

We're not directly hindered by "expensive," since we're not experimentalists. But what does "extremely error-prone" mean? In effect, two things:

1. We get a lot less data – not every site for every cell.
2. We have to deal with the problem of *allele dropout*, the rate for which can be as high as 40%.

Allele dropout

Humans are diploid organisms – each site is repped by two chromosomes (one from each parent)

A–G–G–A–T–T–A–C
A–G–G–G–T–T–T–C

A–G–G–G–T–T–A–C
A–G–G–G–T–T–A–C

We can't measure the chromosomes separately - only their consensus.

Allele dropout

Allele dropout occurs when experimentation misses or destroys nucleotides in one of the chromosome, which are called alleles

A–G–G–A–T–T–A–C

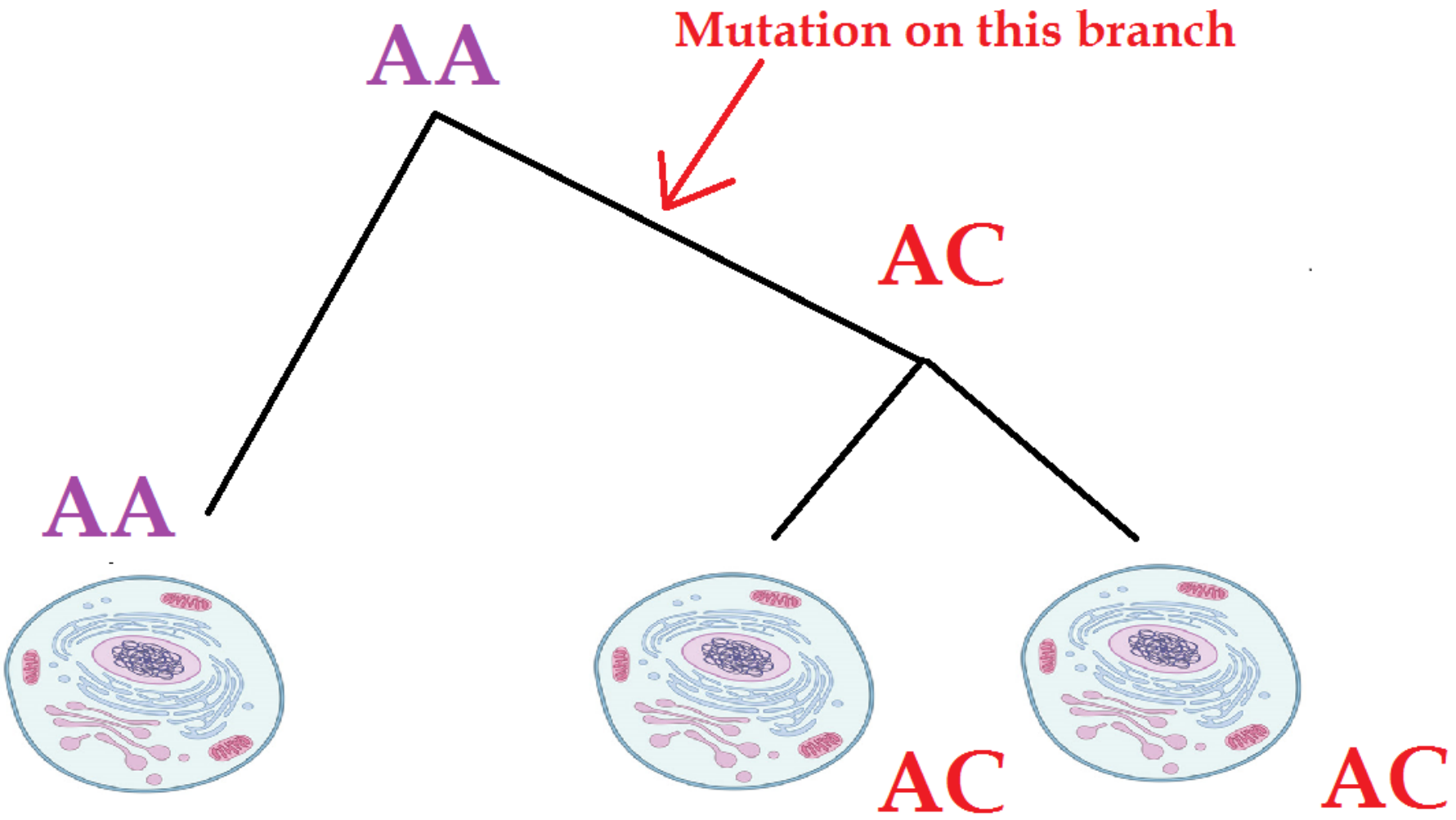
A–G–G–G–T–T–A–C

A–G–G–G–T–T–A–C

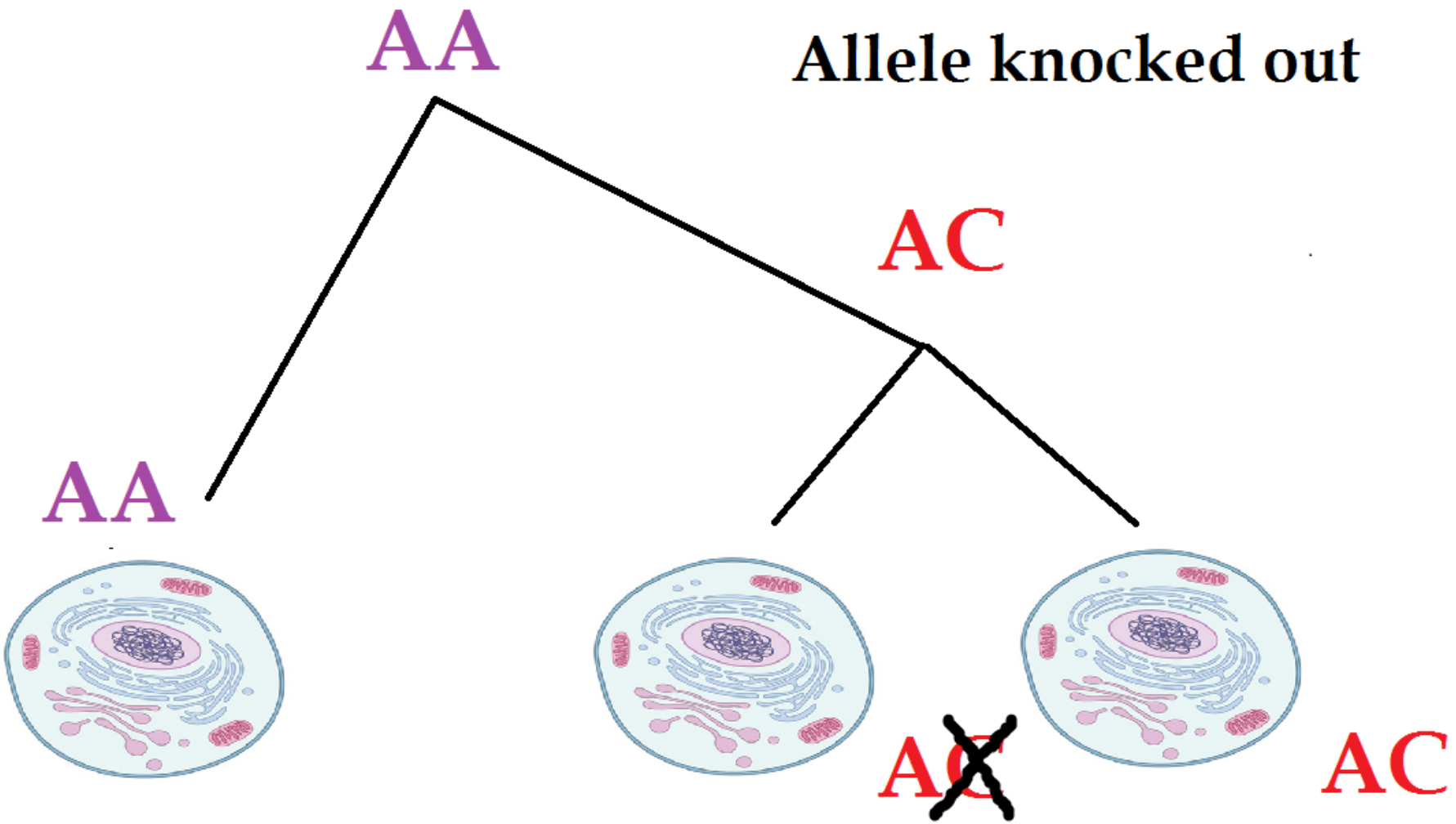
A–G–G–G–T–T–A–C

Allele dropout has destroyed the mutation...this happens like 40% of time in SC data

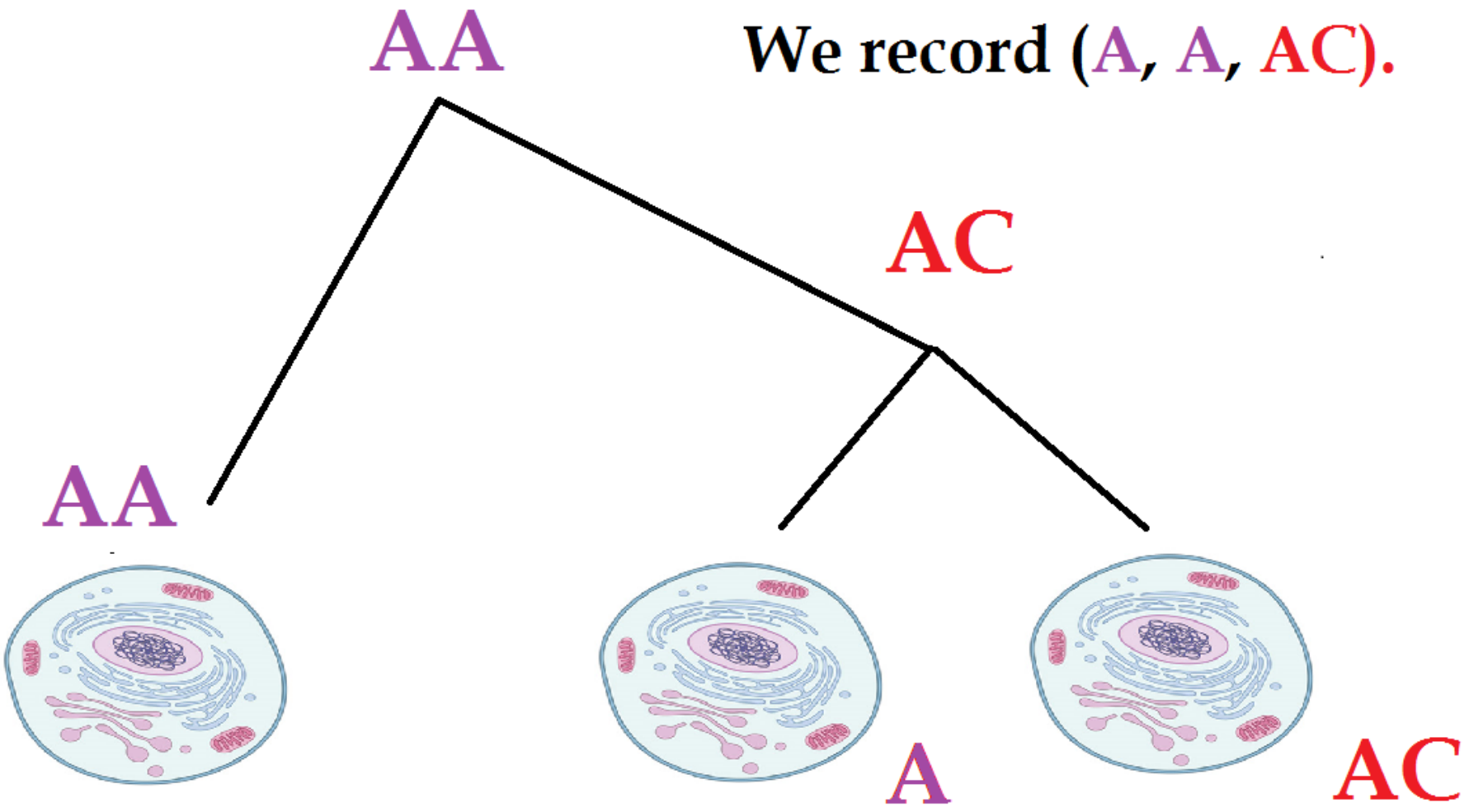
Why this is a problem



Allele dropout: example

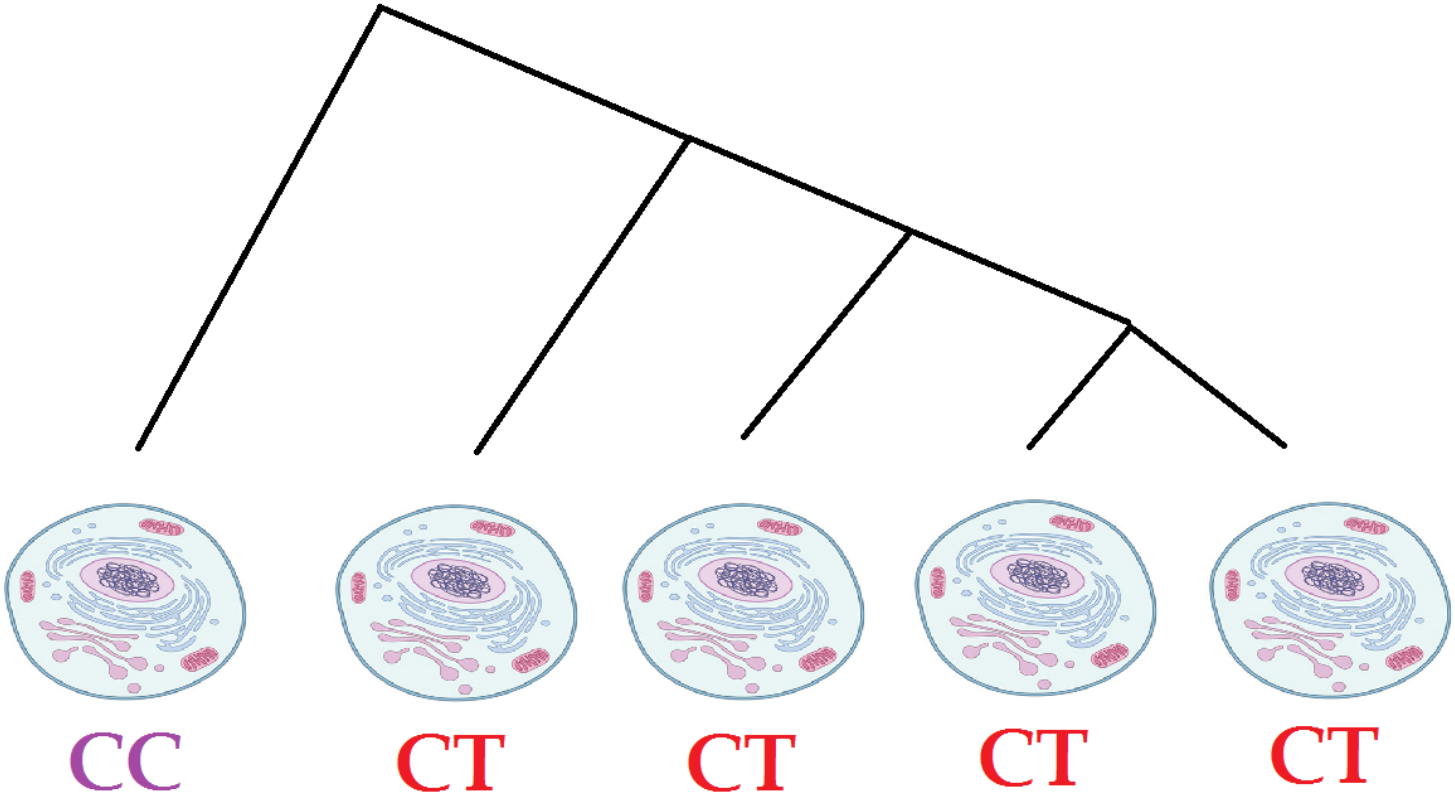


Mutation lost



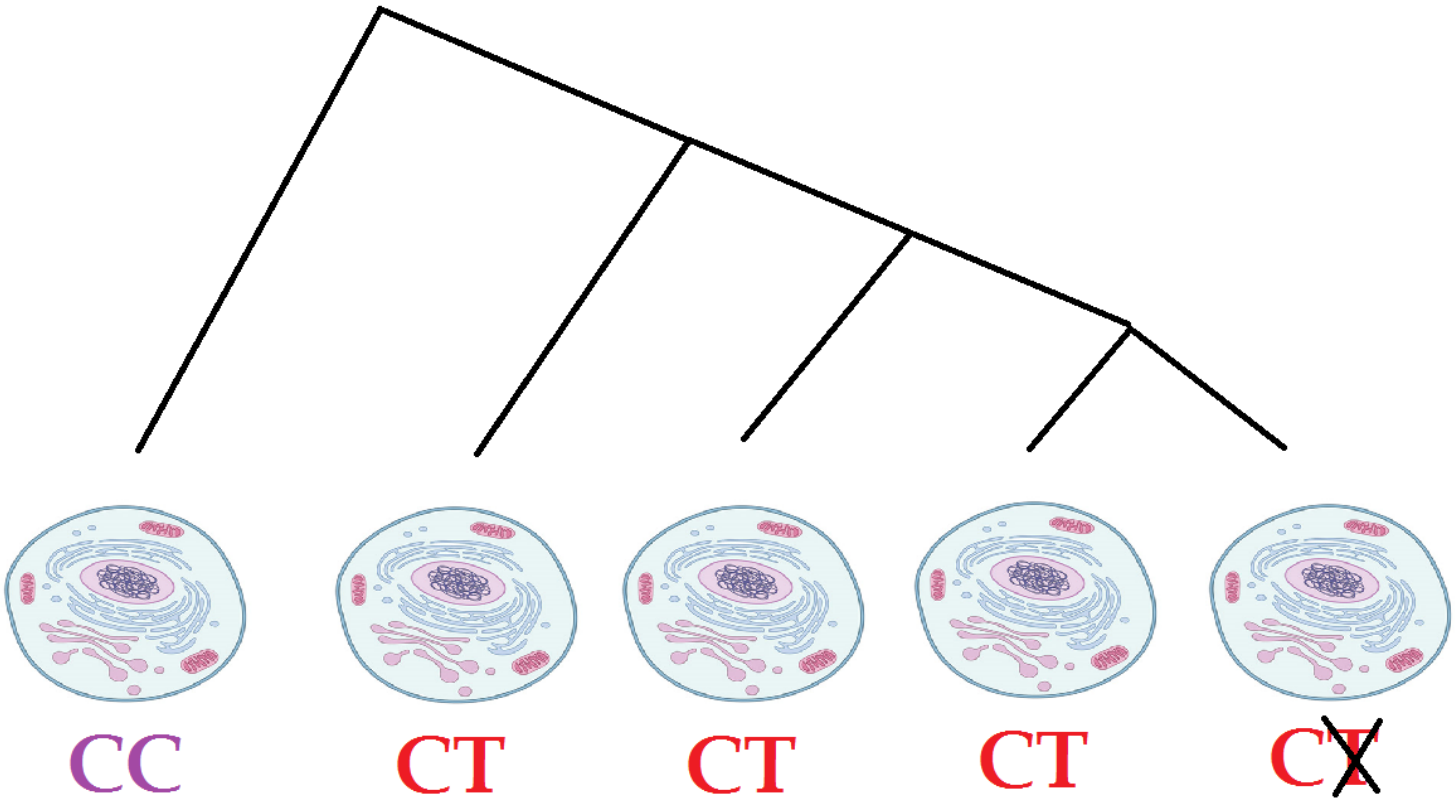
Bad

This is kind of our worst nightmare – can lead to false inference about closeness of relationships



Worst possible inference

In this example, it suggests that the most distantly related cells are actually closest



The bottom line

So the bottom line is, allele dropout makes our data totally self-contradictory.

The bottom line

So the bottom line is, allele dropout makes our data totally self-contradictory.

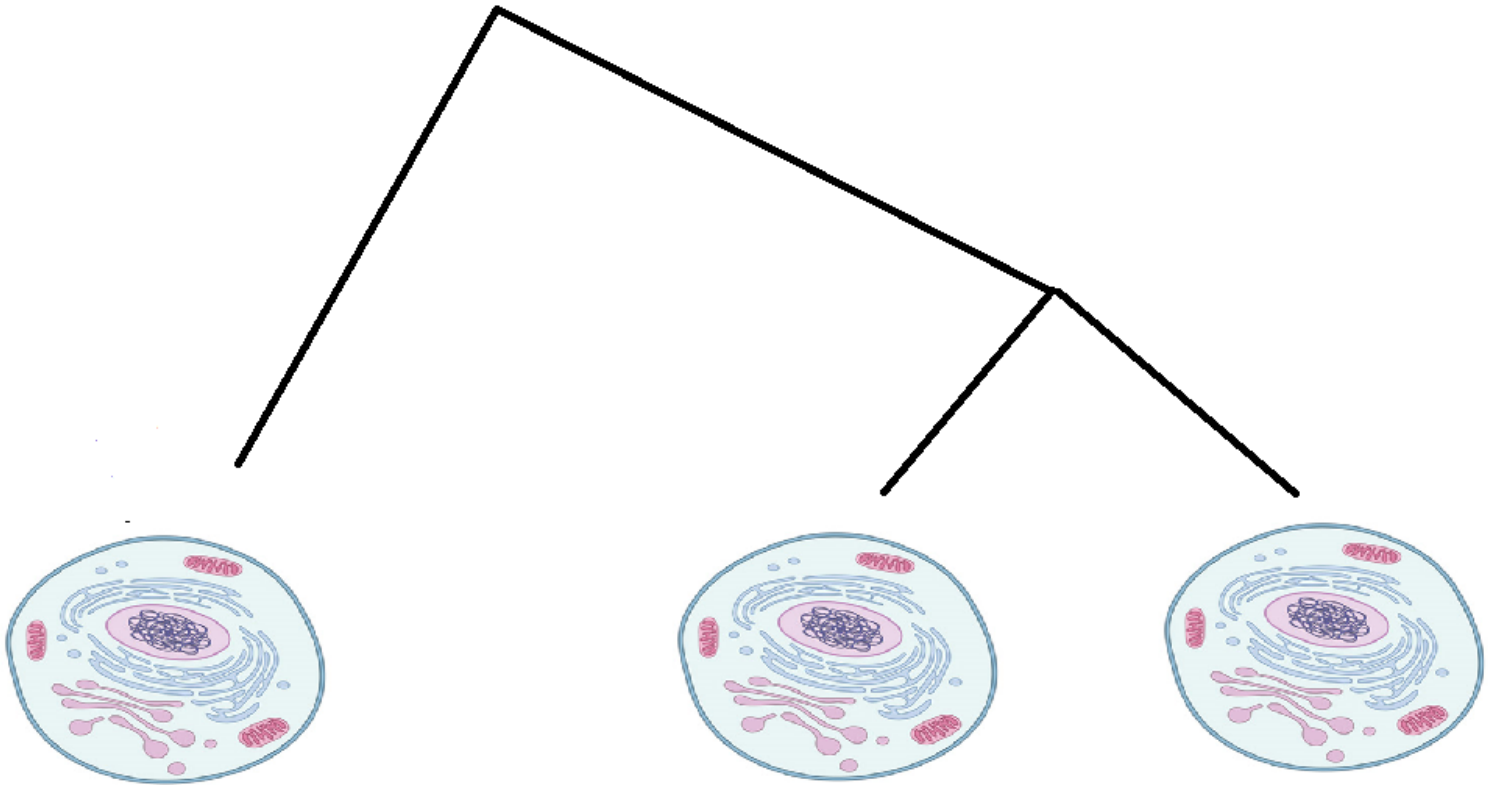
So how to deal with this problem? Try to isolate allele dropout one triplet of cells at a time.

How to address

Assume true tree is as below, and let

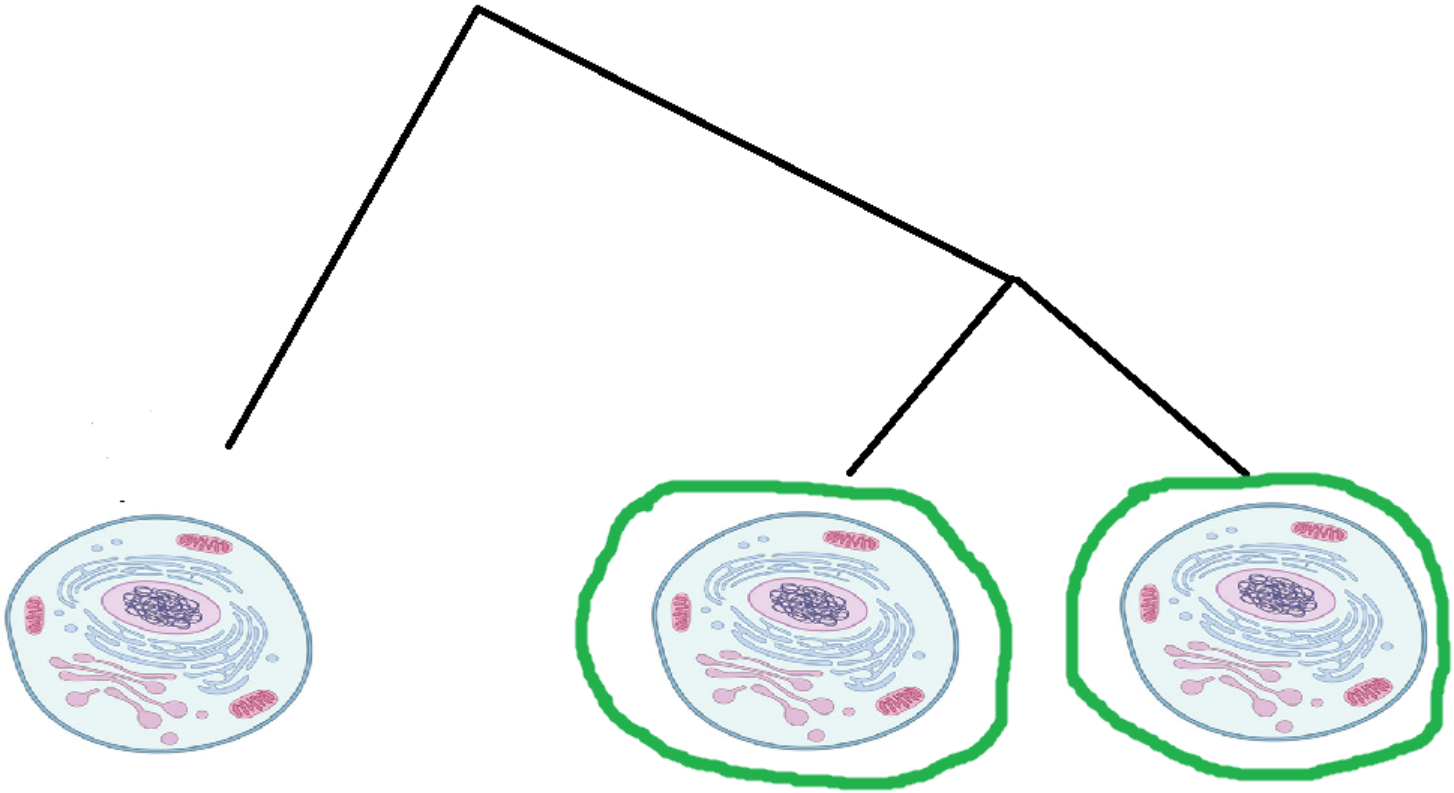
$$(M_1, M_2, M_3) \in \{0, 1\}^3$$

be pattern at given site for cells 1,2,3 (out of 55 total)



How to address

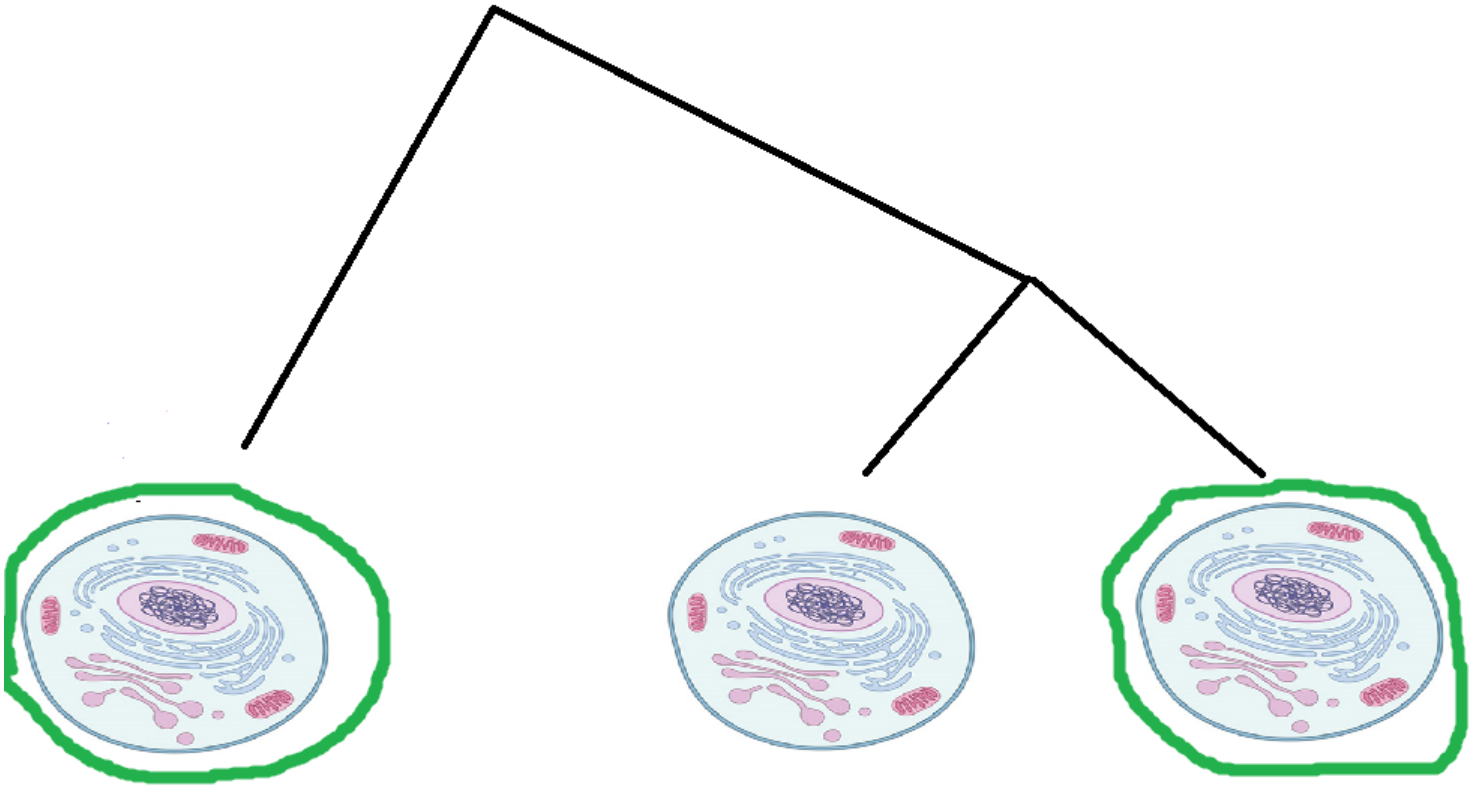
Then even with allele dropout, we expect to see $(0, 1, 1)$



How to address

Then even with allele dropout, we expect to see $(0, 1, 1)$

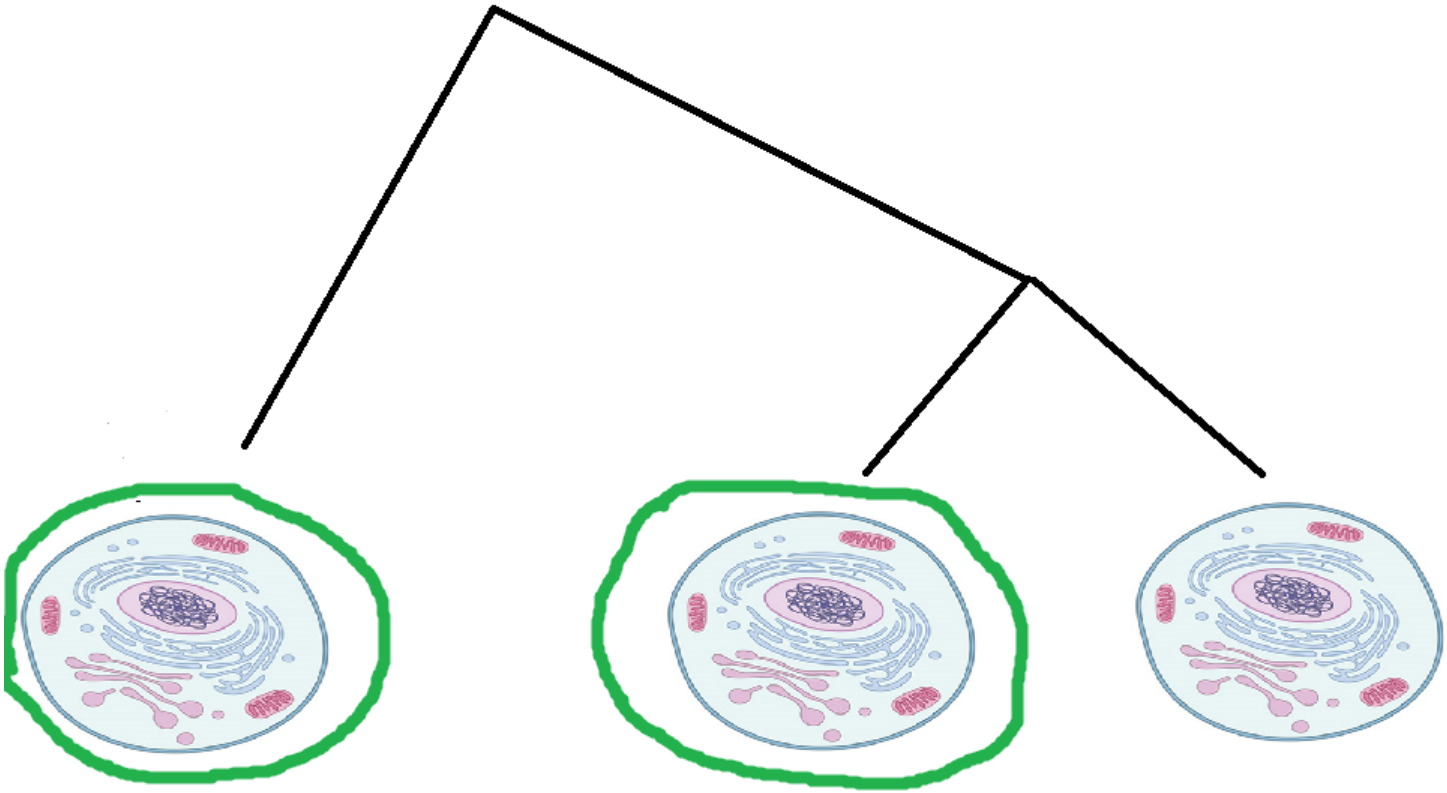
far more frequently than $(1, 0, 1)$



How to address

Then even with allele dropout, we expect to see $(0, 1, 1)$

far more frequently than $(1, 0, 1)$ or $(1, 1, 0)$

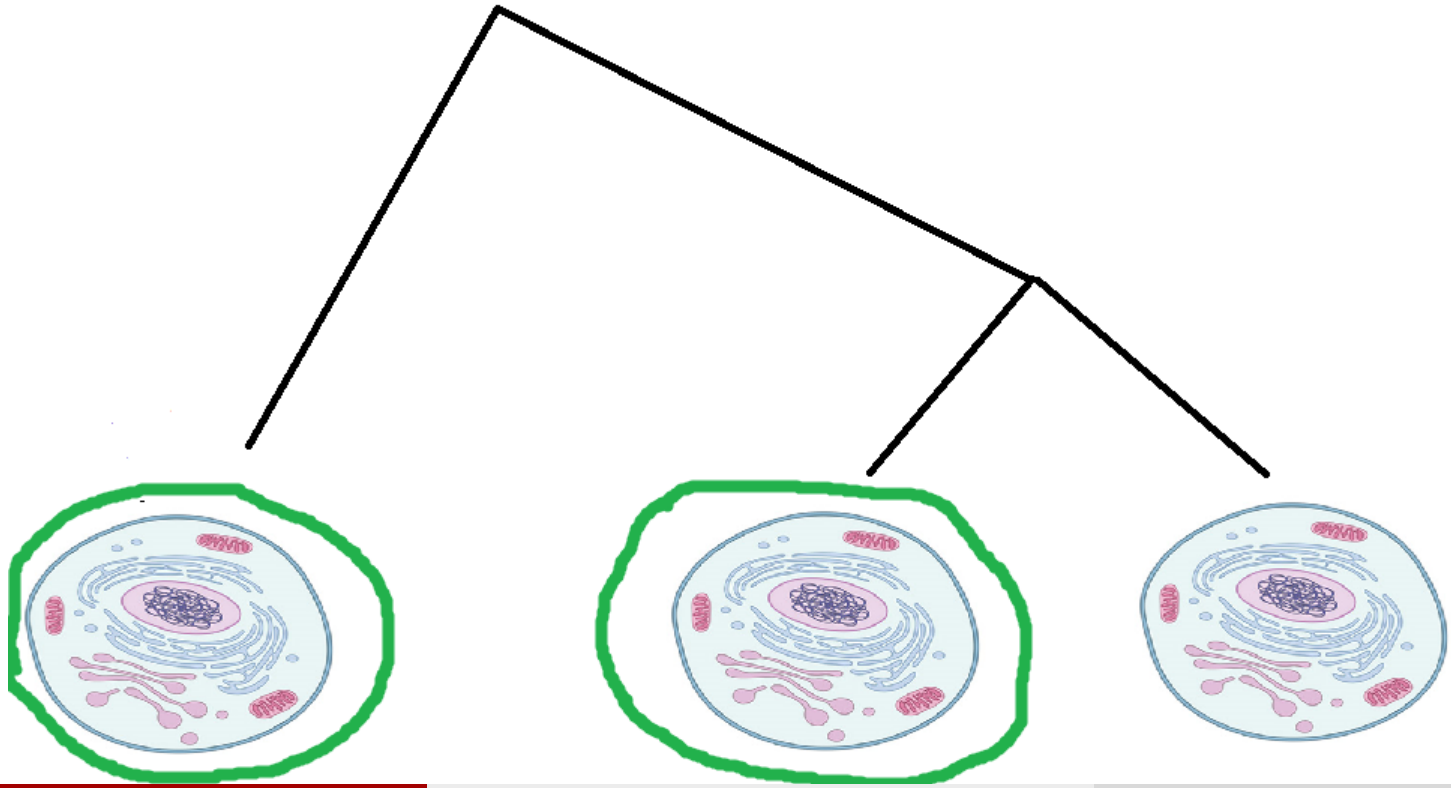


How to address

Then even with allele dropout, we expect to see $(0, 1, 1)$

far more frequently than $(1, 0, 1)$ or $(1, 1, 0)$

And we expect to see $(1, 0, 1)$ or $(1, 1, 0)$ in roughly equal proportions



Liklihood ratio test

So, out of 55 cells, for each triplet of cells $i < j < k$ we count the possible mutation-states in which exactly one cell is left unmutated,

$$n_1 = \#\text{ sites with } (M_i, M_j, M_k) = (0, 1, 1)$$
$$n_2 = \#\text{ sites with } (M_i, M_j, M_k) = (1, 0, 1)$$
$$n_3 = \#\text{ sites with } (M_i, M_j, M_k) = (1, 1, 0)$$

We assume $n_1, n_2, n_3 \sim \text{Multinom}(n_1 + n_2 + n_3,\ p_1,\ p_2,\ p_3)$.

Then we evaluate the hypotheses

$$H_1 \ : \ p_2 = p_3 < p_1$$
$$H_2 \ : \ p_1 = p_3 < p_2$$
$$H_3 \ : \ p_1 = p_2 < p_3$$

Bayesian Multinomial

Multinomial probability of n_1, n_2, n_3 is

$$\mathbf{P}(n_1, n_2, n_3) = \frac{(n_1 + n_2 + n_3)!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}.$$

Let's be *Bayesians*:

$$\mathbf{P}(n_1, n_2, n_3 \mid p_1, p_2, p_3) = \frac{(n_1 + n_2 + n_3)!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}.$$

Likelihood Ratio Test

Somewhat amazing fact: if the distribution really satisfies $p_1 = p_2$, then the quantity

$$\Lambda(n_1, n_2, n_3) = \frac{\max_{p_1,p_2,p_3 \mid p_1=p_2} \mathbf{P}(n_1, n_2, n_3 \mid p_1, p_2, p_3)}{\max_{p_1,p_2,p_3} \mathbf{P}(n_1, n_2, n_3 \mid p_1, p_2, p_3)},$$

called the *likelihood ratio*, exhibits the distributional convergence

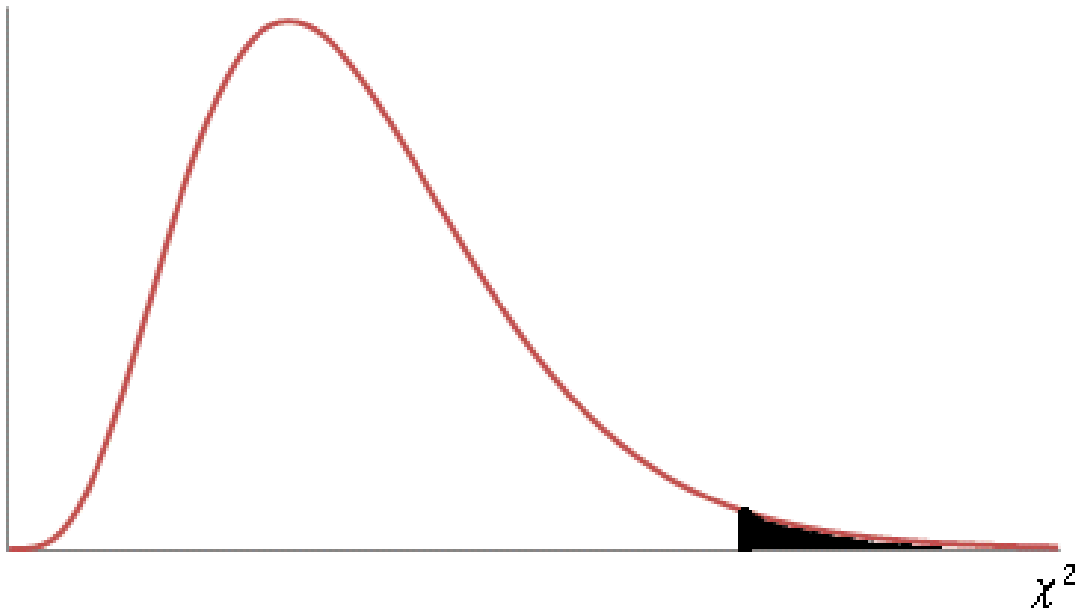
$$-2\log(\Lambda(n_1, n_2, n_3)) \xrightarrow{d} \chi^2(1).$$

Hypothesis Test

So if e.g.,

$$\mathbf{P}(\chi^2 > -2\log(\Lambda(n_1, n_2, n_3))) < .05,$$

then we can reject the hypothesis that $p_1 = p_2$ at a 5% level of significance.



Plan of attack

For each triplet of cells $i < j < k$, test hypotheses

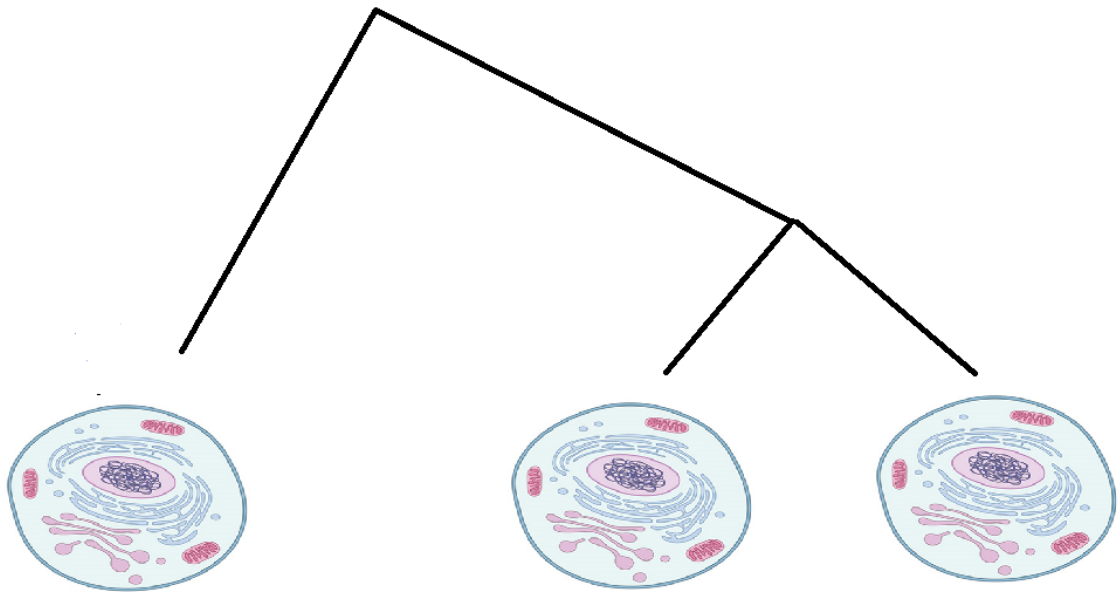
$$\begin{aligned} H_1 &: p_j = p_k < p_i \\ H_2 &: p_i = p_k < p_j \\ H_3 &: p_i = p_f < p_k \end{aligned}$$

at significance-evels $\alpha = .05,\ .01$.

If we can reject two hypotheses at $\alpha = .01$ and fail to reject the third at $\alpha = .05$, then we conclude that the third is true, and any data contradicting it is due to allele dropout.

Changing the data

If e.g. we accept H_1, we then change all data-values of $(1, 0, 1)$ and $(1, 1, 0)$ to $(1, 1, 1)$.



Because we've established that the above pic holds, and $(1, 0, 1)$ and $(1, 1, 0)$ are pretty unlikely given this topology.

Does this help?

Does this help? Yes!

We ultimately build our tree using SVDquartets (Kubatko and Chifman, 2014) in combination with PAUP (Swafford, 2002).

Does this help?

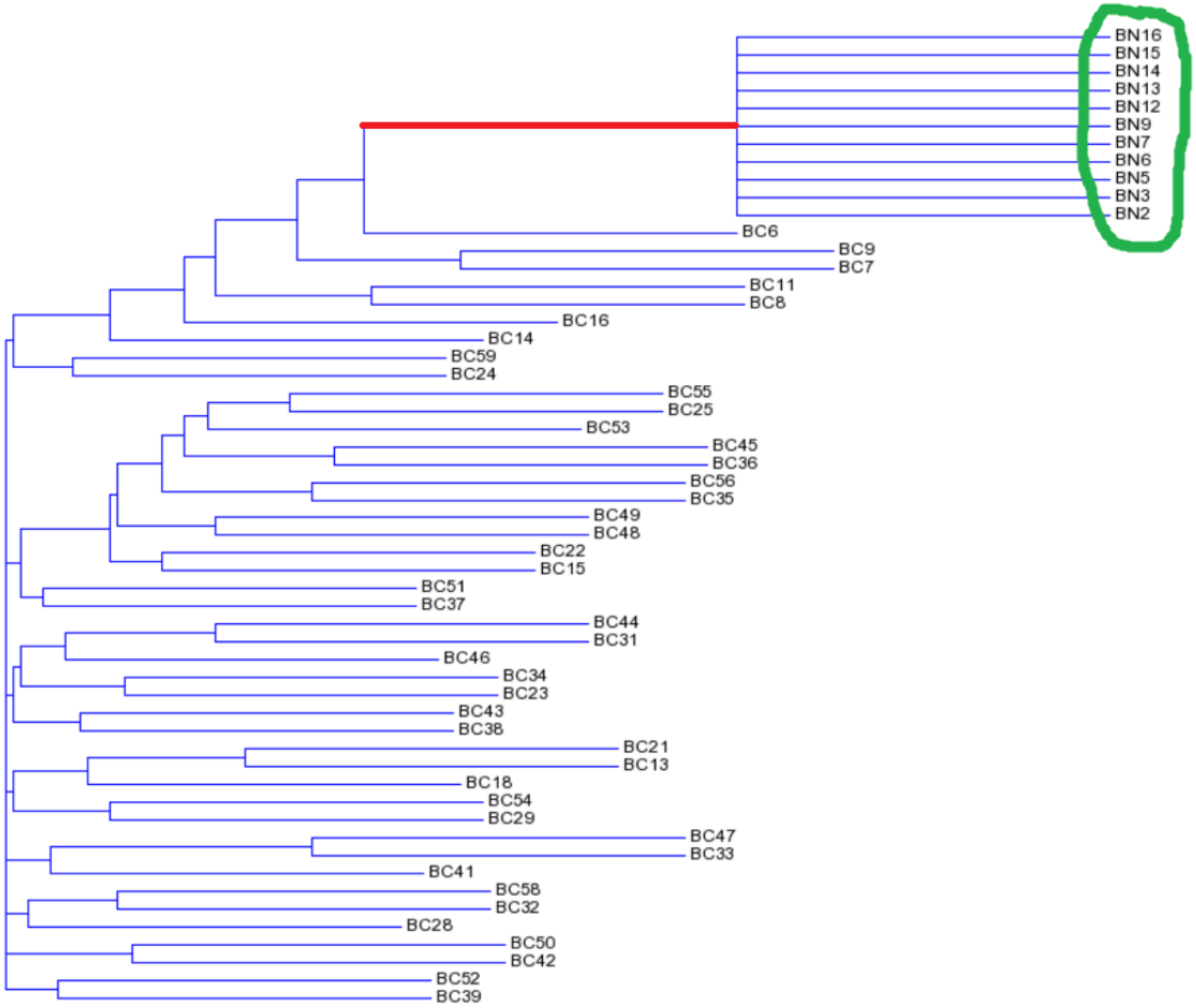
Does this help? Yes!

We ultimately build our tree using SVDquartets (Kubatko and Chifman, 2014) in combination with PAUP (Swafford, 2002).

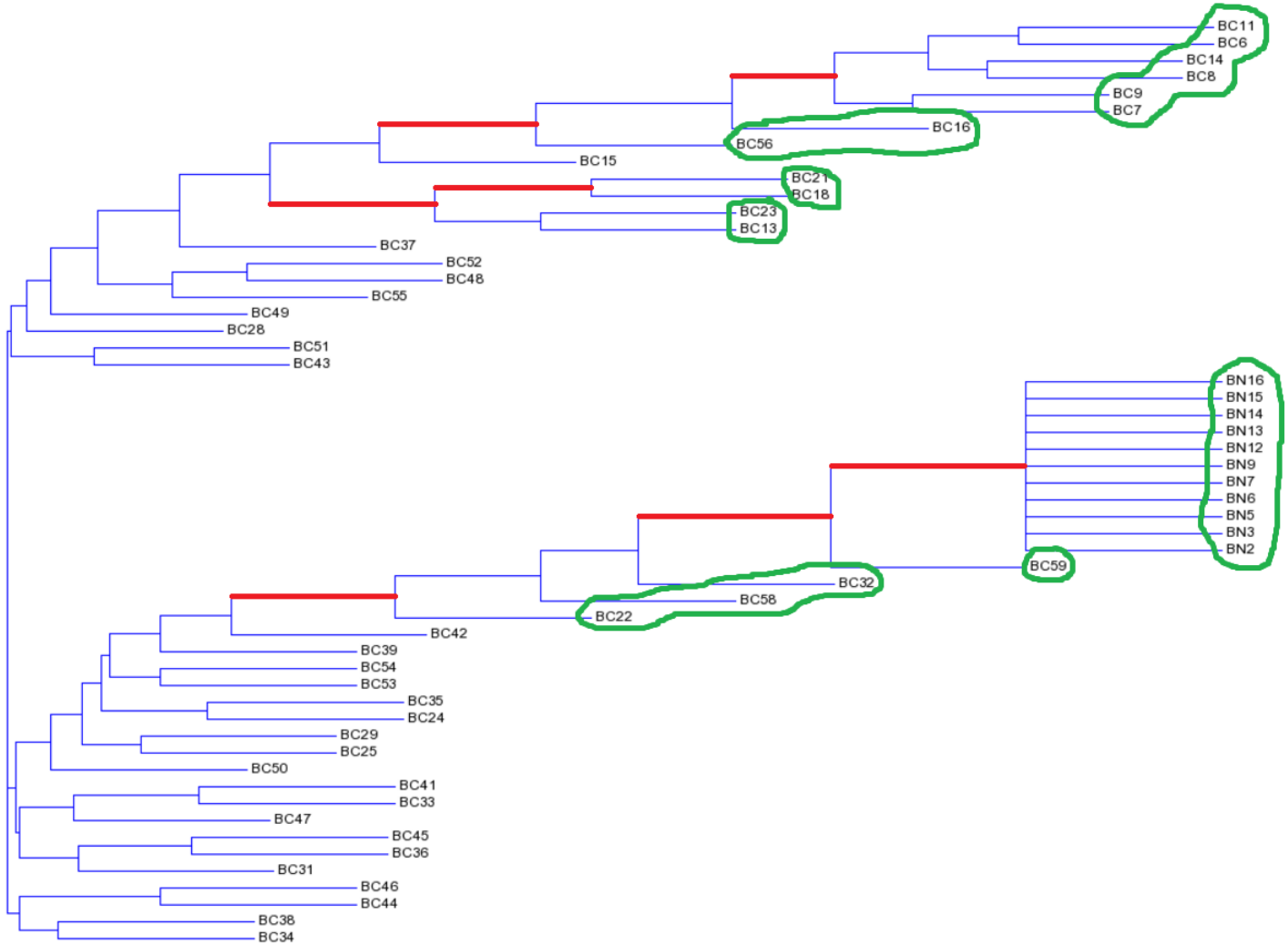
We judge its accuracy via *bootstrapping* (sampling dataset with replacement).

A branch which appears in 80% of bootstrapped samples is considered well-supported.

Without triplet procedure



With triplet procedure



Future directions

- Try using the triplets to actually reconstruct the tree
- Look for canonical mutations in well-supported branches
- Iterate . . . ?

Acknowledgments

Thanks to collaborators

Laura Kubatko

Julia Chifman

Kate Hartmann

Acknowledgments

Thank you for your attention!