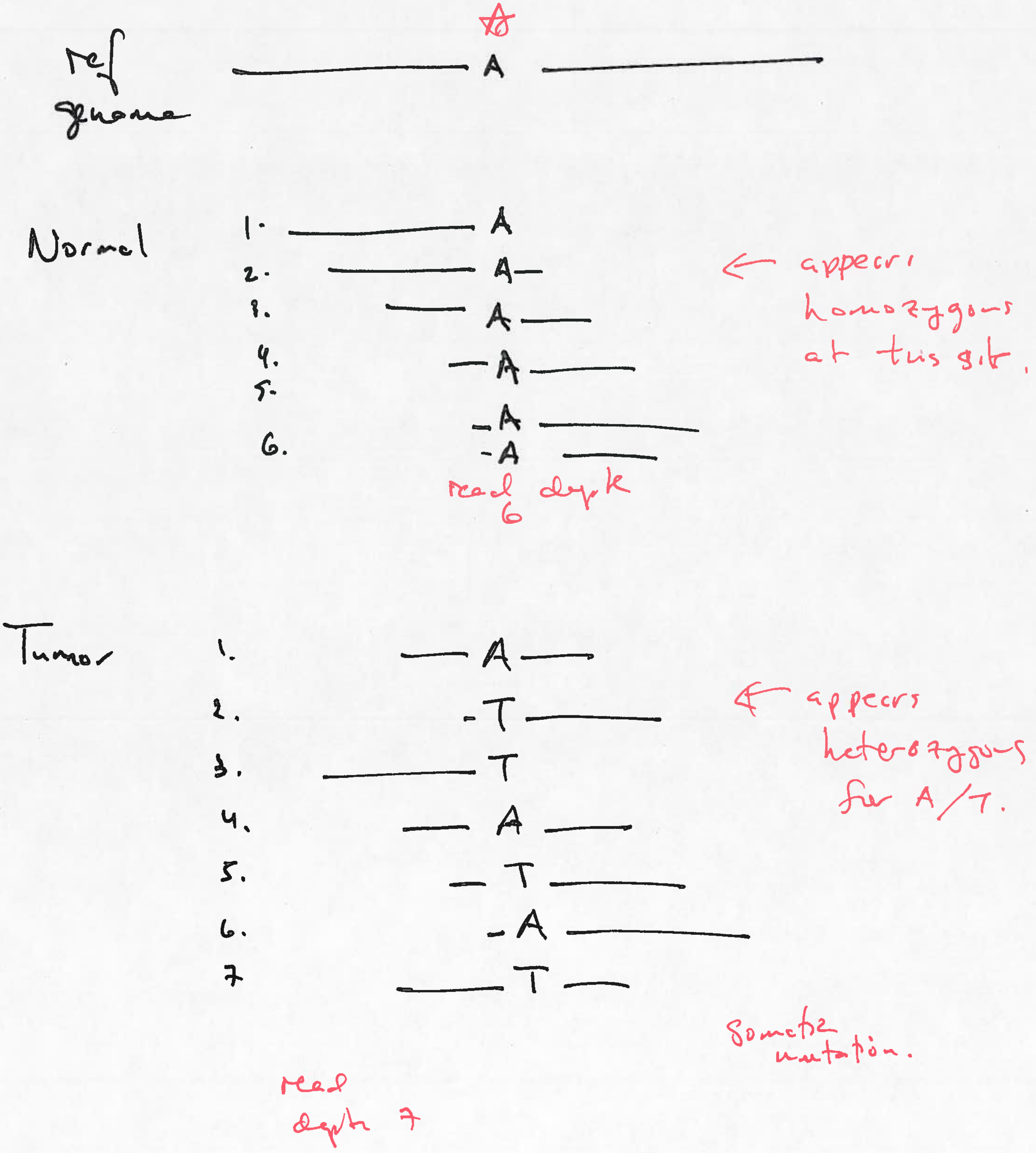
Identifying Somatic Mutations via MuTect (from the Broad Institute)



read depth 6

← appears homozygous at this site.

← appears heterozygous for A/T.

Somatic mutation.

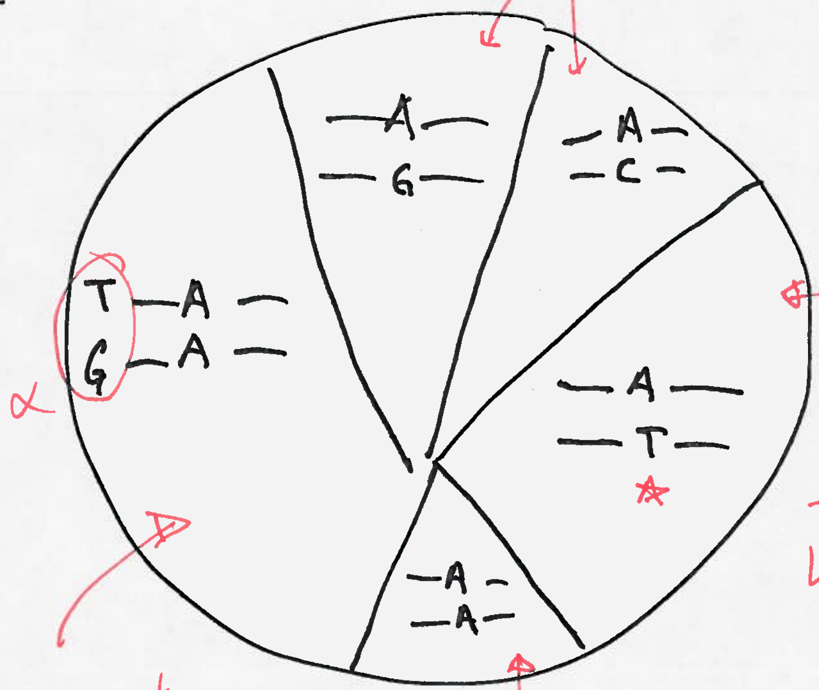
read depth 7

Tumors have clonal structure.

Humans have 2 copies of each genomic site.

They have a different mutation. we assume this doesn't occur

Think of this as a pie chart.



This is the fraction of cells in the tumor sample that have te mutation

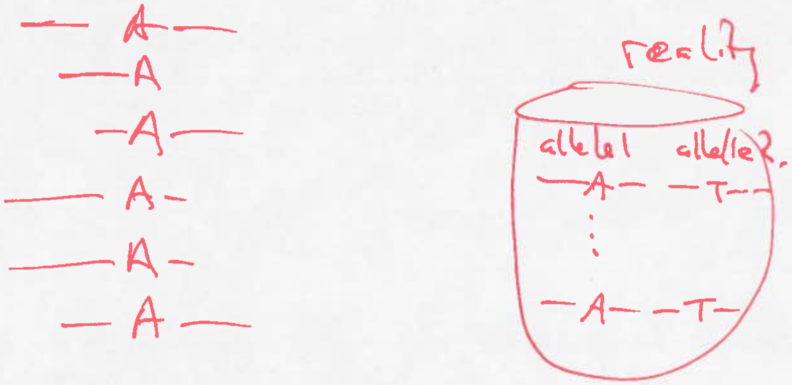
These cells are tumor cells but they don't have the A→T mutation at site ☆. They might have other somatic mutations though. (eg. T-G upstream α)

this fraction ~~is~~ corresponds to normal cells in the tumor sample. So "contamination". But some normal cells will always be harvested with tumor cells. (What affect does this have on our analysis?)

To call somatic mutations in an accurate manner (highly specific & sensitive), there are at least 4 parameters:

1. depth of sequence coverage in tumor & normal.

Imagine depth 6 (only) in tumor. What is the prob. that she is heterozygous at a site where all 6 reads are A?



6 coin tosses; all 6 are heads.
(the sequencer reaches in the bag and pulls out 1 of 2 copies of the locus at random).

2. Error rate of sequencer.

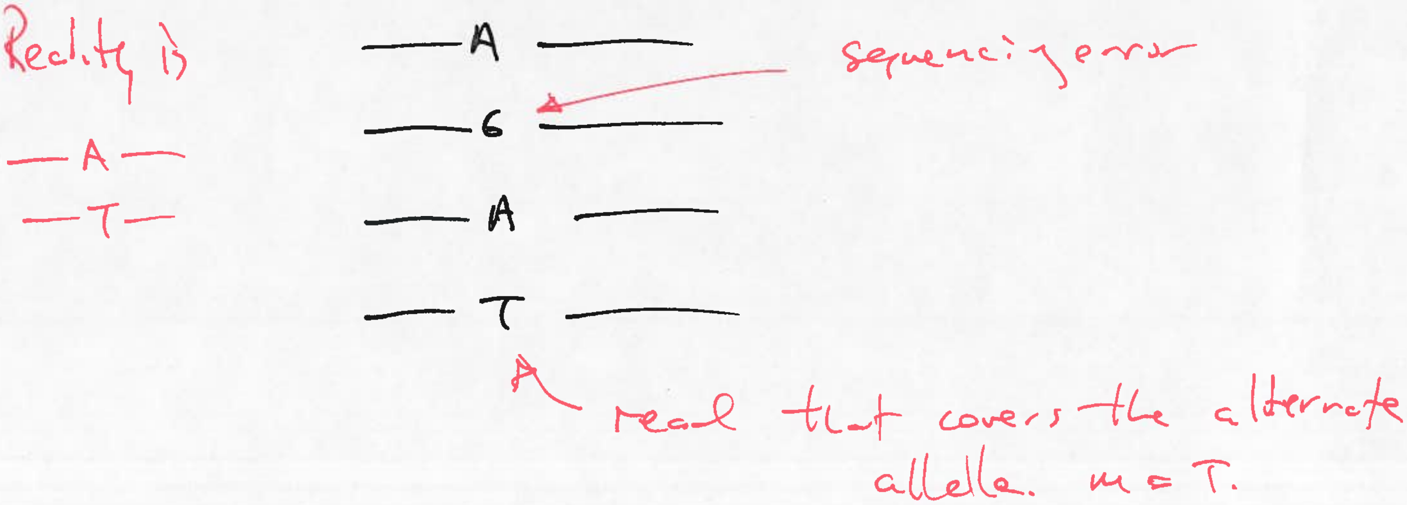
PHRED measures how reliable a signal is for a given genomic locus.

Value 0..1.

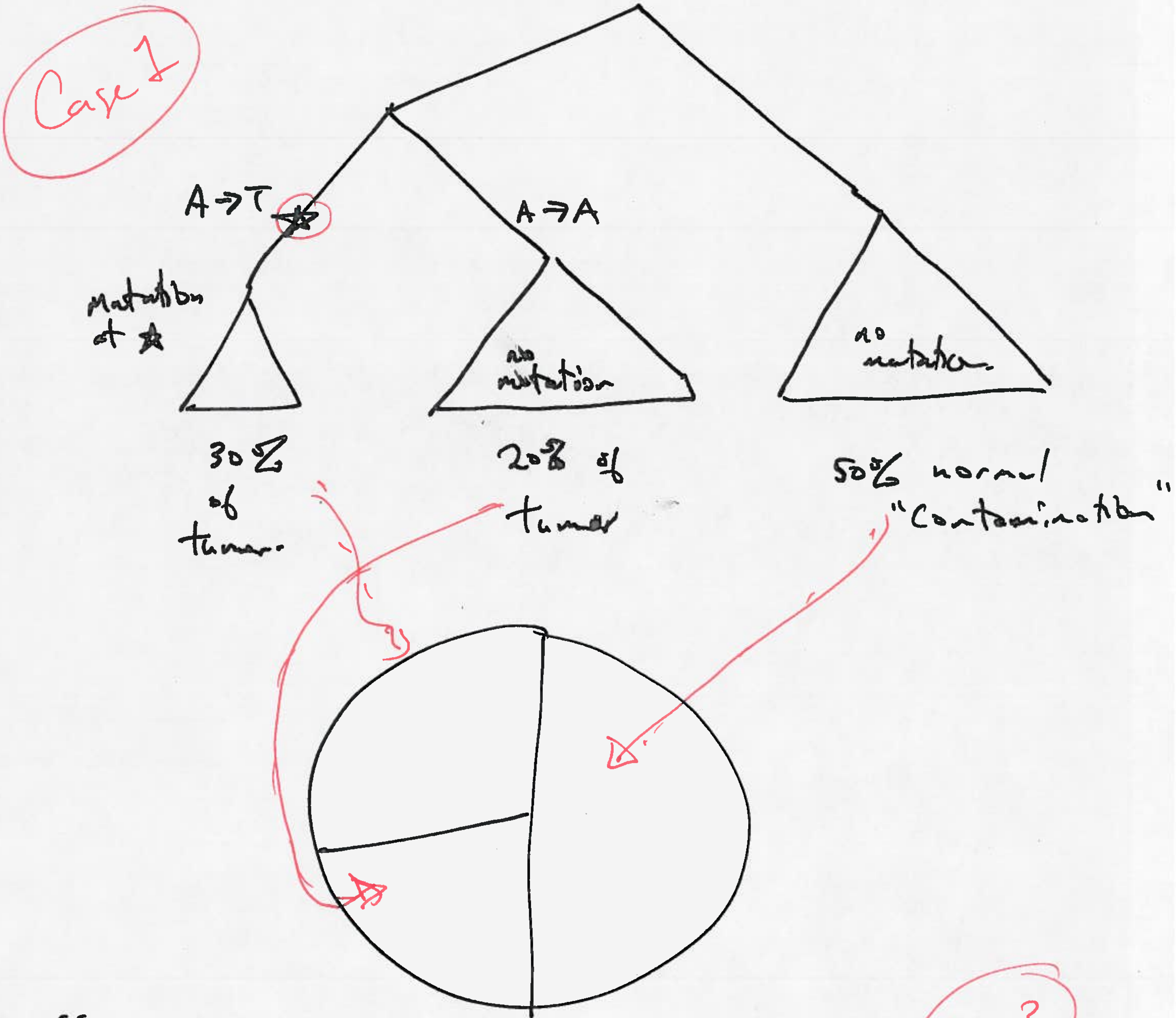
The idea is that we can measure the probability the machine makes a mistake

e.g. $\frac{1}{10,000}$ b.p.

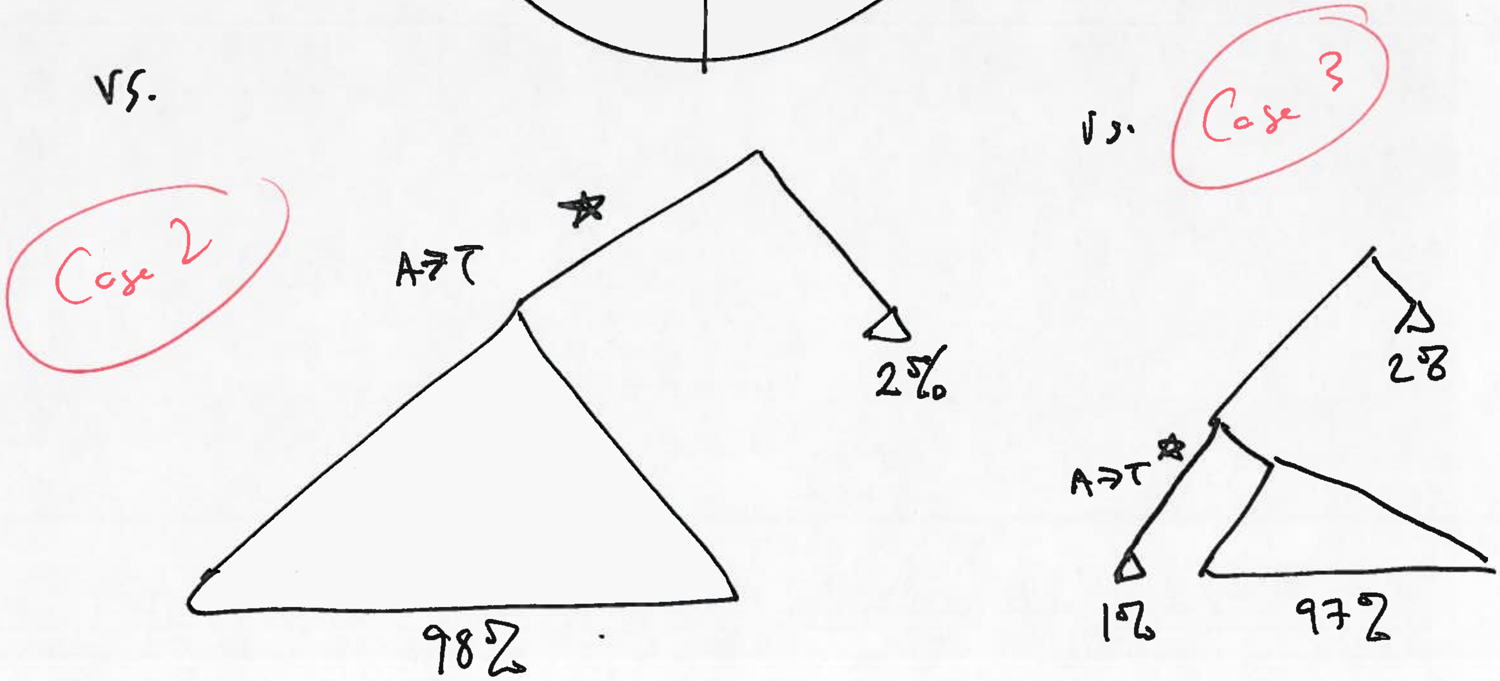
Clearly it is difficult to distinguish between somatic mutations & sequencing errors @ low depth



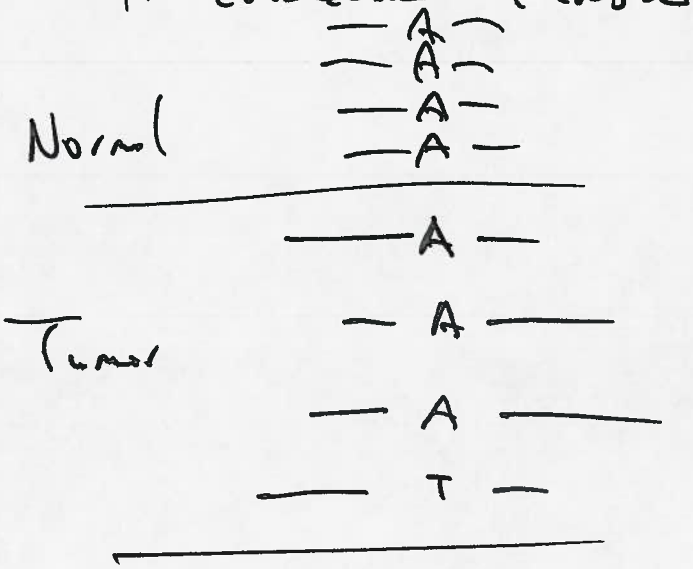
3. Allelic fraction of the mutation



vs.



4. evidence thresholds.



With low read depth, difficult to decide if 1 T in 4 reads is enough.

Heterozygous or sequencing mistake?

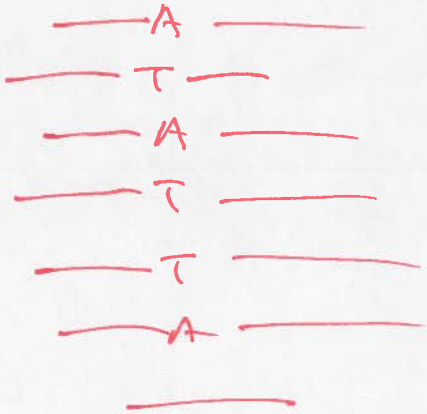
How certain do you want to be?

Can you tolerate a few mistakes?

Model M_0 : no variant at thee site.
all non-ref bases are random
sequencing errors.

Alt. model M_f^m : the site contains a true somatic
mutation m with allelic freq f.

We could estimate this
by counting the fraction
of reads with the
variant m versus the
total # of reads



50/50
A/T.

$f = .50$

log. odds

$$\log_{10}\left(\frac{\mathcal{L}[M_f^m]\cdot \Pr[m,f]}{\mathcal{L}[M_0]\cdot(1-\Pr[m,f])}\right) \geq \log_{10} \delta$$

arbitrary base ... anything > 1 is ok, I guess.

It's easiest to think of δ as 0.

So if the numerator > denominator the data supports the alt. model more then the null model.

In practice δ = 6.3 is used.

Implying the alt. model must be $10^{6.3} : 1$ favored over the null model.

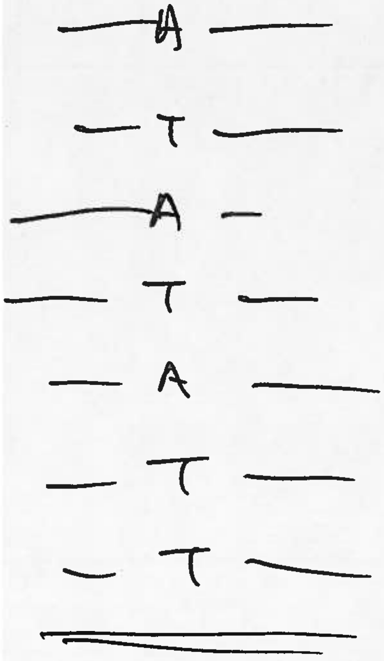
More conservative that δ = 0 ⇒ fewer sites are called somatic mutations.

(Don't worry about where 6.3 comes from)

For our site i, let $r \in \{A, C, G, T\}$ represent the reference allelle for the woman.

Here we ~~[illegible]~~ are going to assume that she is homozygous for either A, C, G, or T at site i. (Analysis that allows heterozygosity is a bit more cumbersome but essentially the same.)

Suppose that we have d reads that cover site i.



	b_i	e_i
1	A	0.1
2	T	0.01
3	A	0.3
4	T	$\boxed{1}$
5	A	0.9 -bad
6	T	0.0001
7	T	0.000001 -good.

e_4 is a probability that the site is a sequencing error $0 \le e_i \le 1$.

It is derived by a program called PHRED.

~~That is basically a program to [illegible]~~

Now observe that

$$M_0 = M^m_f \quad \text{where} \quad f=0$$

(so a variant m exists at the site i but it has frequency 0... just a mathematical refor~~mod~~ulation to save us a bit of time).

$$\mathcal{L}(M^m_f) = \Pr[\{b_i\} \mid \{e_i\}, r, m, f]$$

~~over~~ over all the different reads.

(the prob. ~~of all~~ al considering all read calls bi given what PHRED thinks ({ei}), the reference allele of the site, and assuming the site has somatic mutation m with frequency f).

$$= \prod_{i=1}^{d} \Pr[b_i \mid e_i, r, m, f]$$

Let's assume all substitutions (mutations) away from the reference r occur with equal probability $e_i/3$.

$\boxed{(e_i/3 \text{ because there are 3 alternatives vs. the observed reference})}$

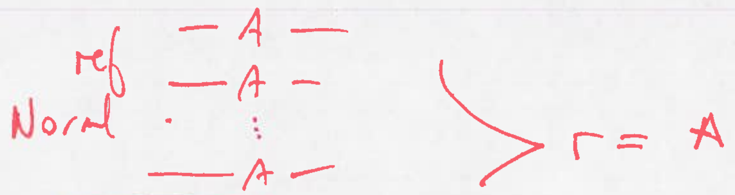
$$\Pr[b_i \mid e_i, r, m, f] = \begin{cases} f \cdot e_i/3 + (1-f)(1-e_i) & \text{if } b_i = r \\ f(1-e_i) + (1-f)(e_i/3) & \text{if } b_i = \\ e_i/3 & \text{otherwise} \end{cases}$$

$e_i/3$ → $\overset{\text{prob. of}}{\text{sequencing error}}$

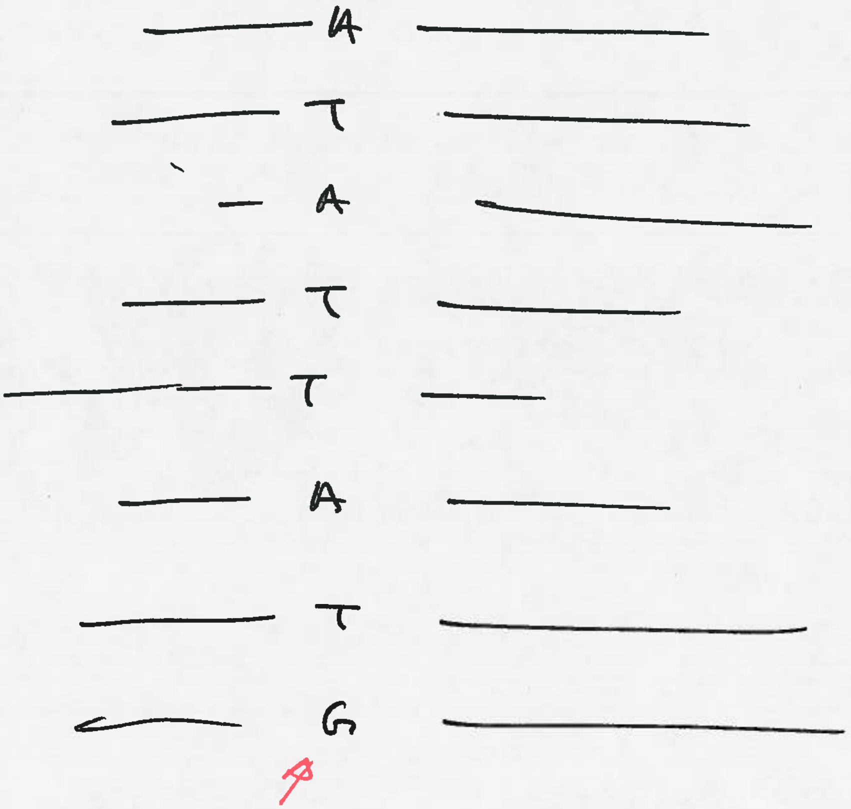
$1-e_i/3$ → $\overset{\text{prob. of}}{\text{no sequencing error}}$

f → frequency of the alt. allele m

$1-f$ → ~~frequency~~ the site has the ref r.



$\boxed{\text{ref} = A}$



$b_8 = G, \quad e_8 = 0.5$

Consider

$\Pr[b_8 = G \mid r = A, m = T, f = 50\%]$ (case 3)
$e_8 = 0.5$

$\Pr[b_8 = G \mid r = A, m = C, f = 1\%]$ (case 3)

$\Pr[b_8 = G \mid r = A, m = G, f = 20\%]$ (case 3)

Which do you think has the highest probability?

Consider

$Pr[b_2 = T \mid e_2 = 0.0001, r = A, m = T, f = 50\%]$

vs.

$Pr[b_2 = T \mid e_2 = 0.0001, r = A, m = T, f = 90\%]$

vs

$Pr[b_2 = T \mid e_2 = 0.9, r = A, m = T, f = 90\%]$

Consider

$Pr[b_3 = A \mid e_2 = 0.5, r = A, m = T, f = 90\%]$

vs.

$Pr[b_3 = A \mid e_2 = 0.5, r = A, m = T, f = 50\%$

$Pr[b_3 = A \mid e_2 = 0.00001, r = A, m = T, f = 60\%]$

Are we sure the sequencer is correct?