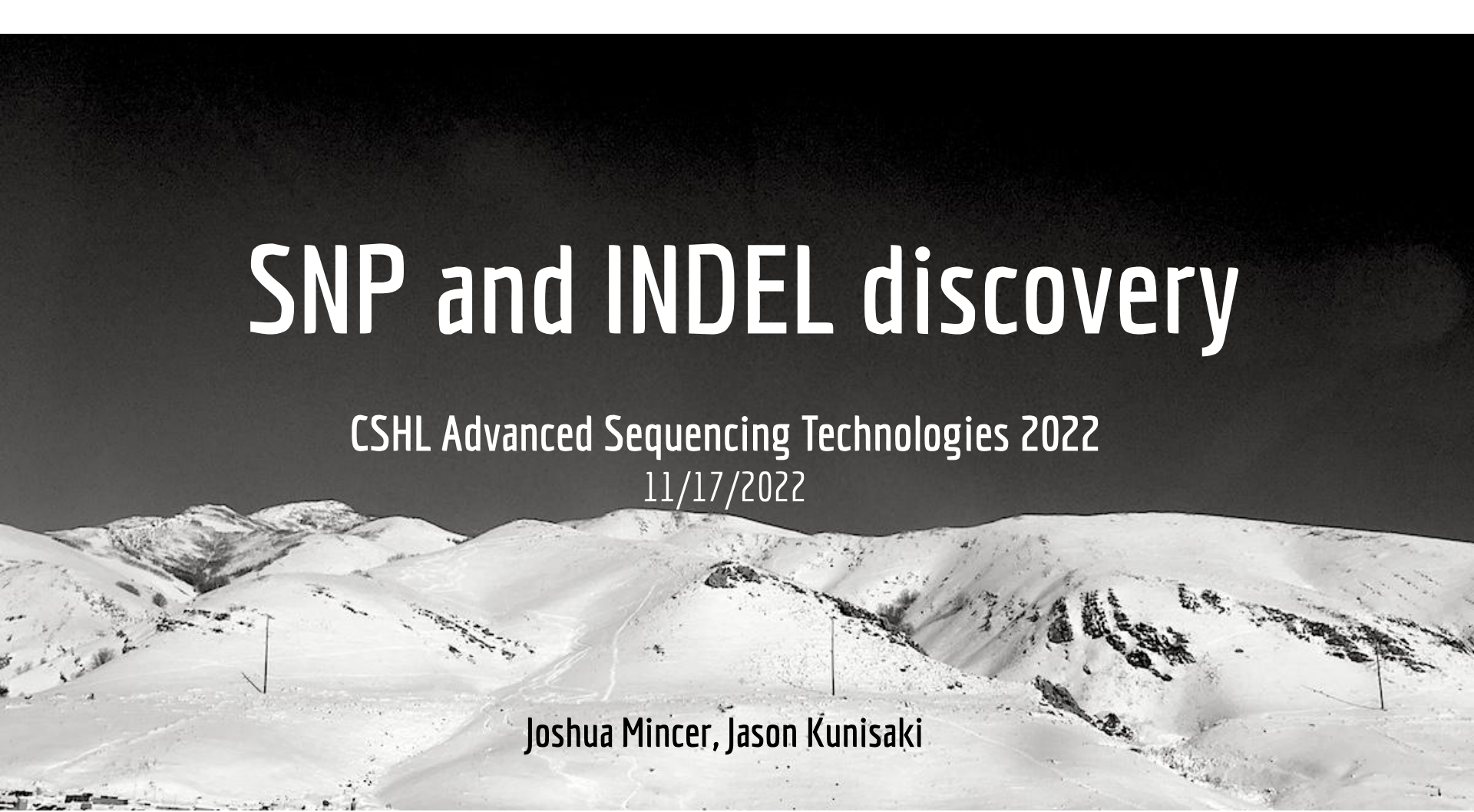
SNP and INDEL discovery

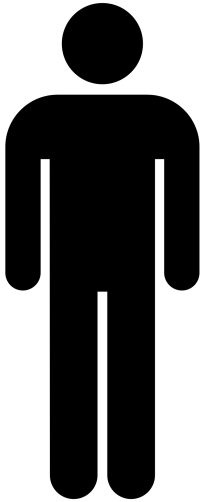
CSHL Advanced Sequencing Technologies 2022

11/17/2022

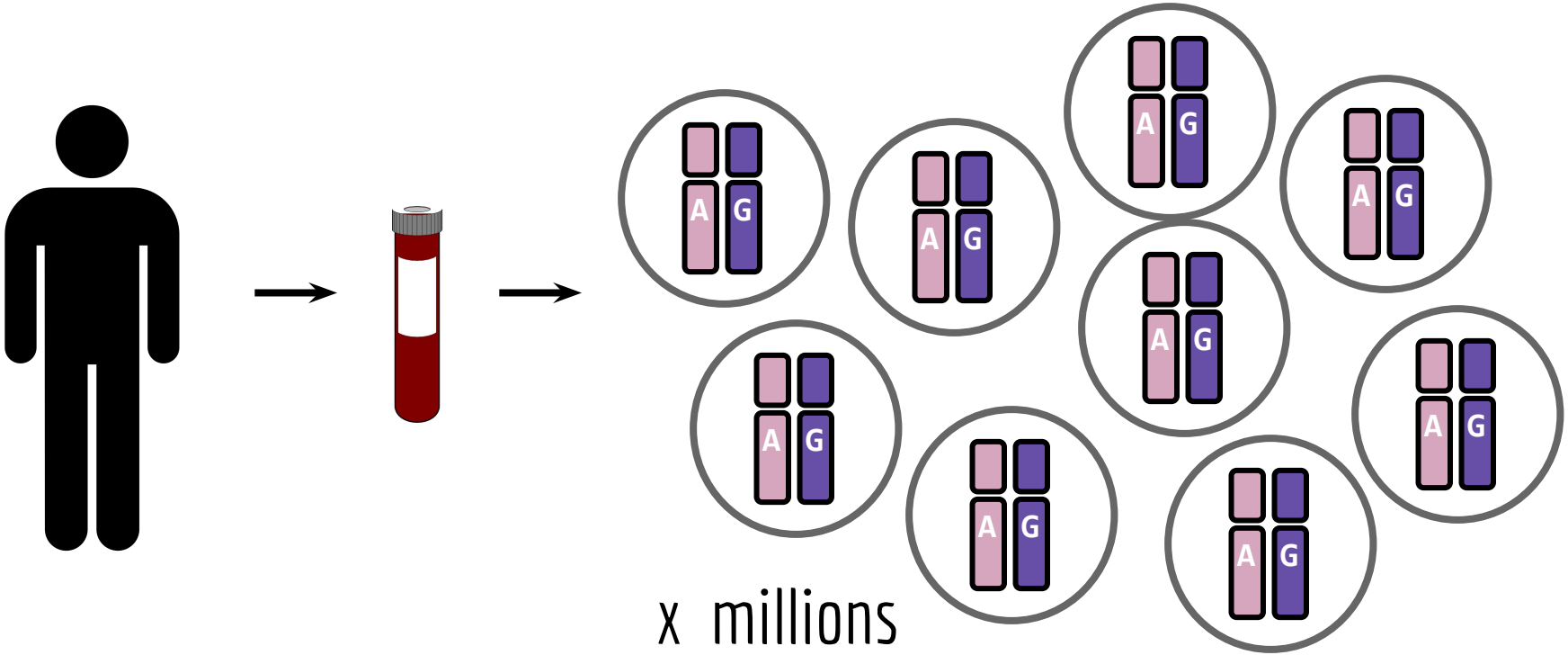


Joshua Mincer, Jason Kunisaki

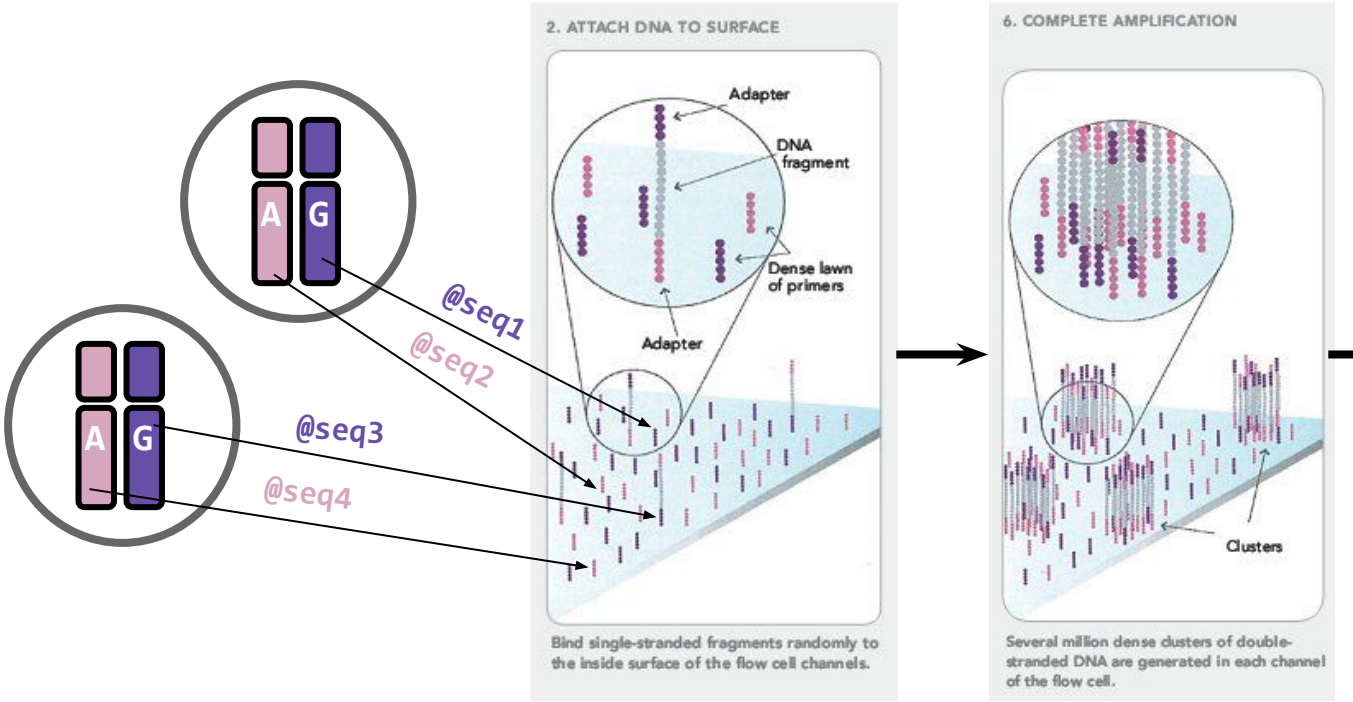
Goal: find all inherited variants in an individual's diploid genome.



Find inherited genetic variation by sequencing DNA from millions of cells

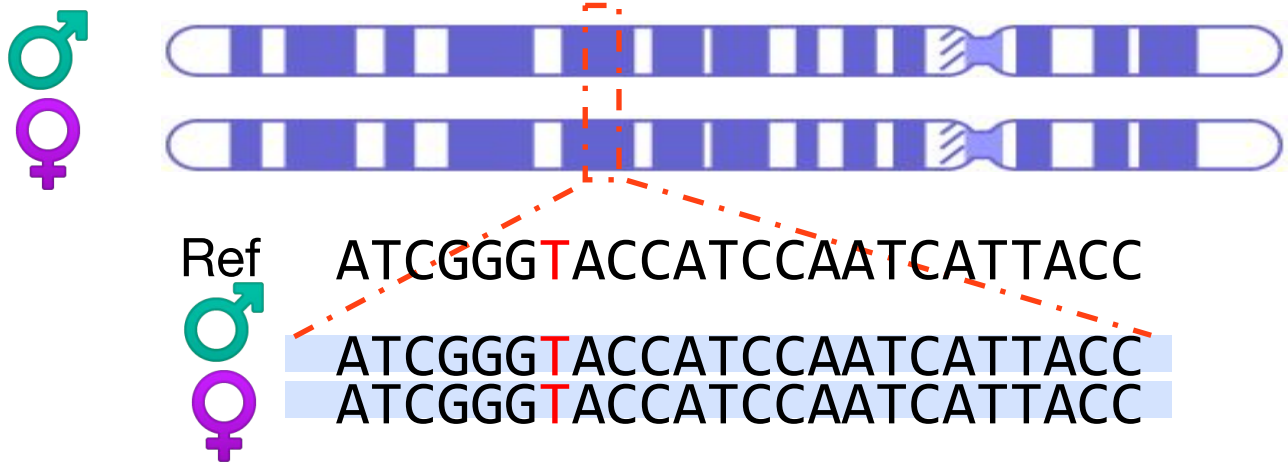


Each DNA cluster is amplified from a single strand from a single haploid chromosome from a single cell.

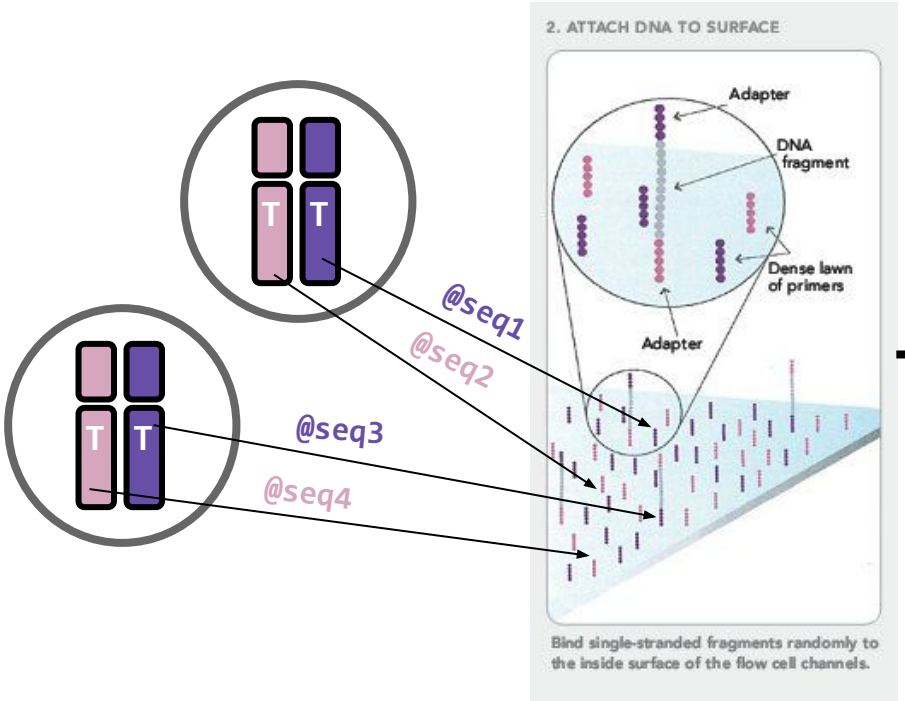



```
@seq1
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
@seq3
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq4
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
```

Scenario 1: An individual is homozygous for the "reference" allele.

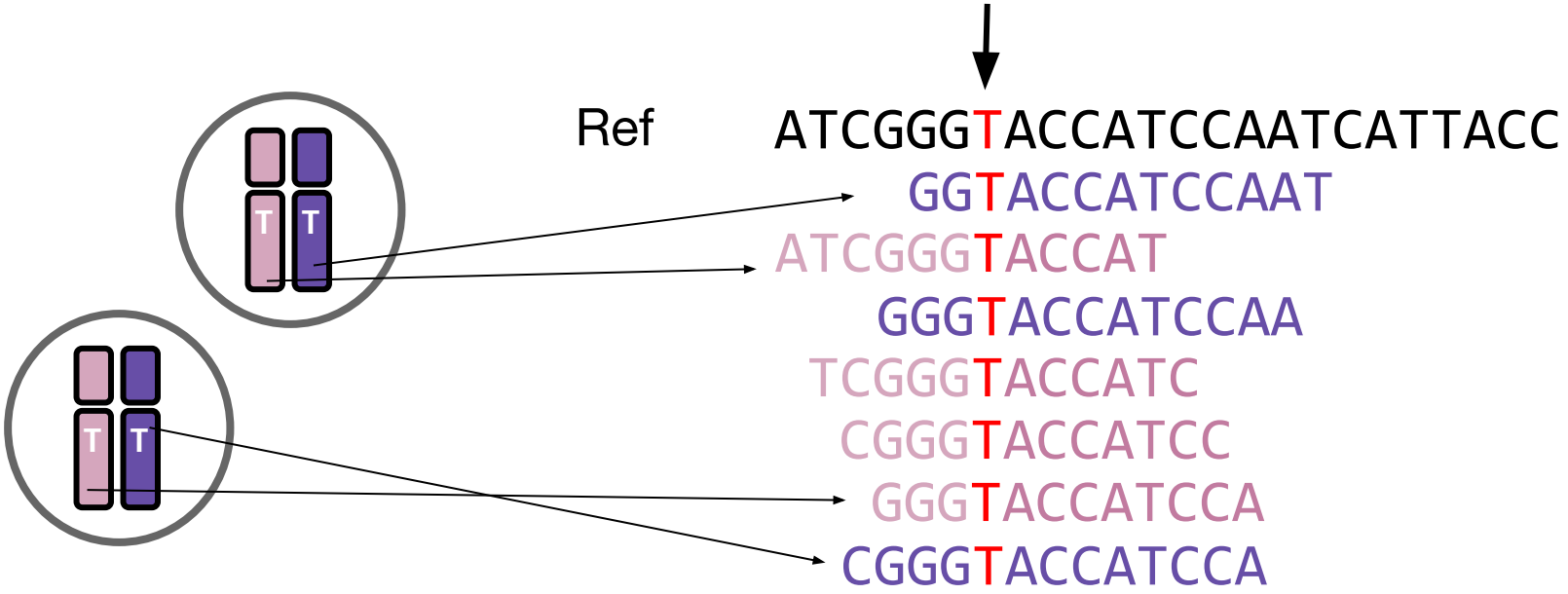


Scenario 1: An individual is homozygous for the "reference" allele.

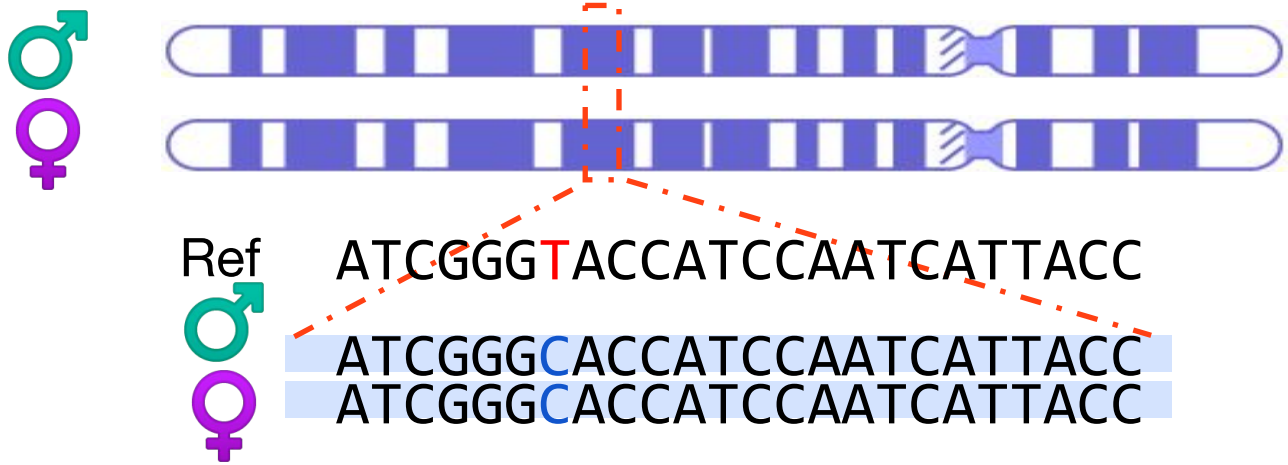



```
@seq1
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
@seq3
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq4
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
```

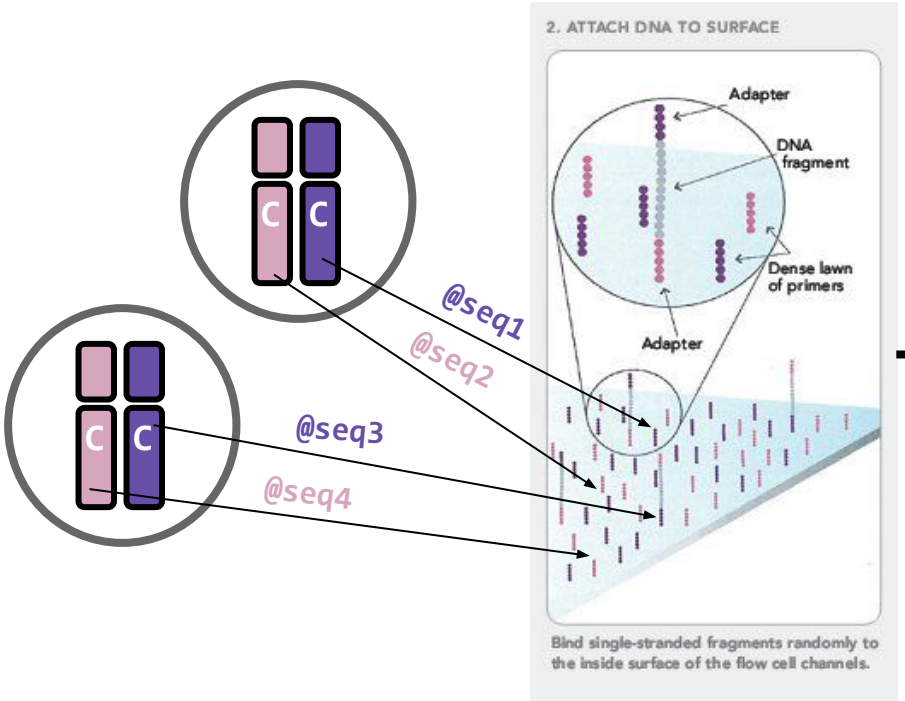
Scenario 1: An individual is homozygous for the "reference" allele.

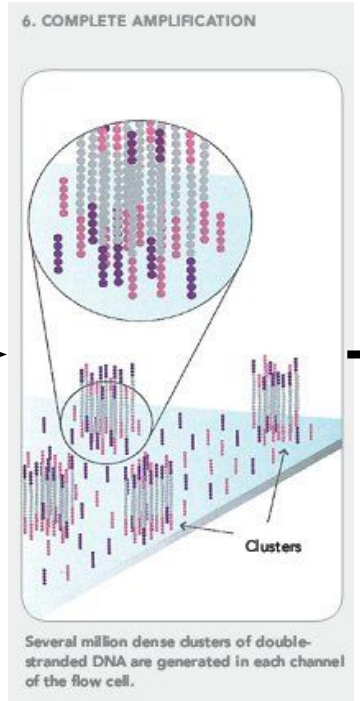


Scenario 2: An individual is homozygous for an "alternate" allele.



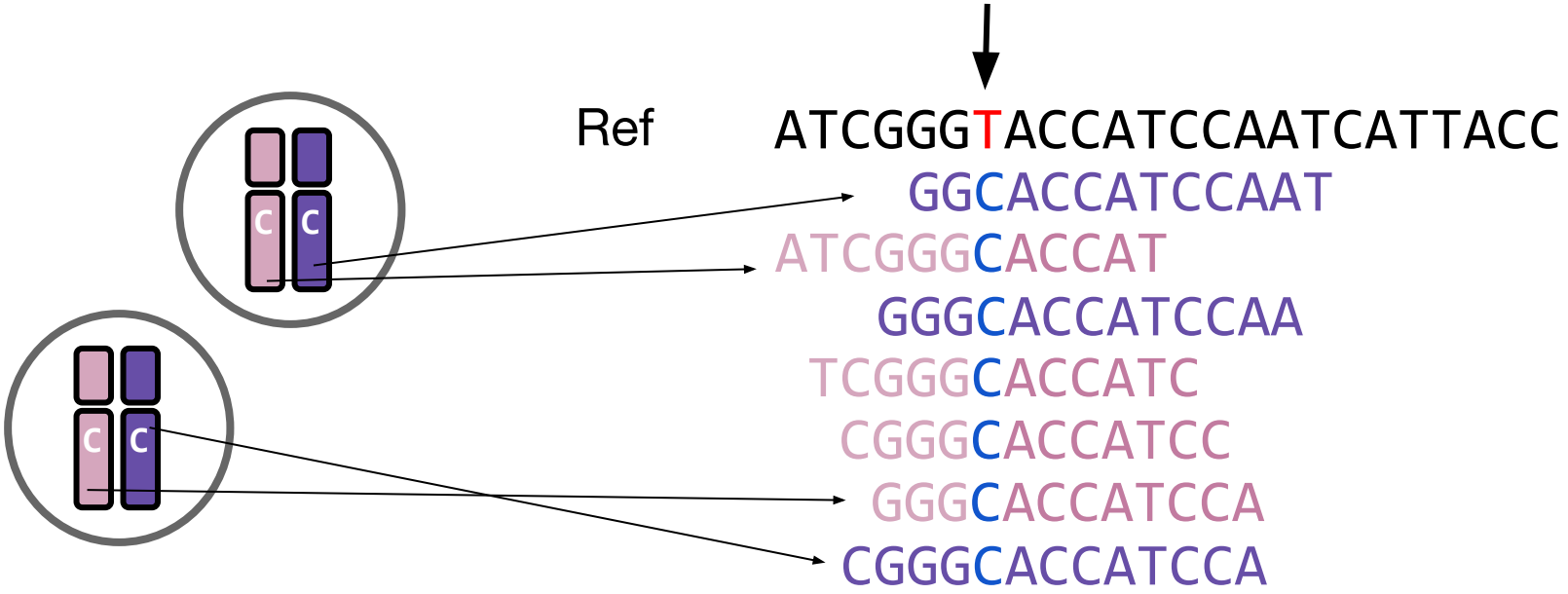
Scenario 2: An individual is homozygous for an "alternate" allele.



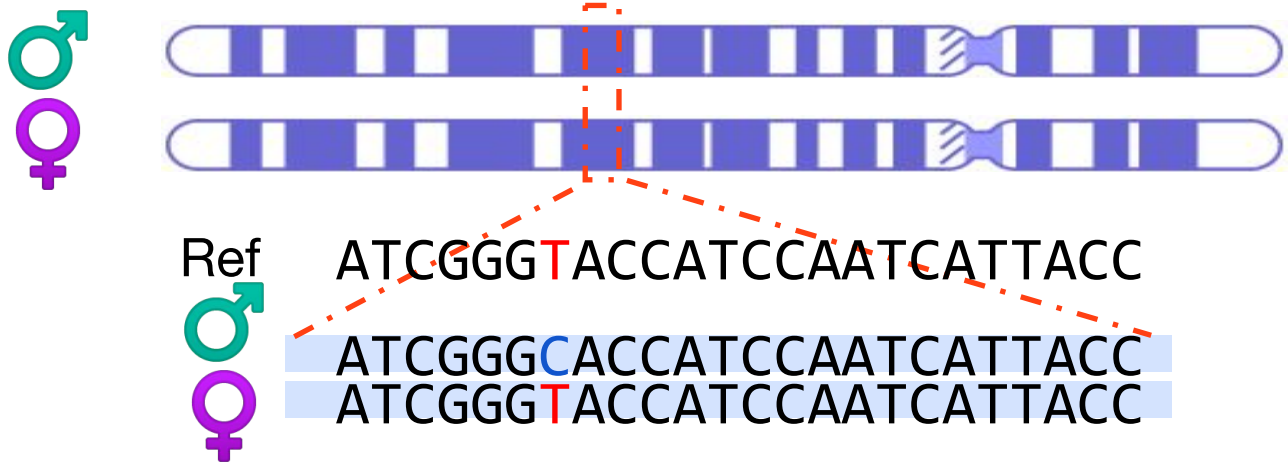



```
@seq1
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
@seq3
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq4
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
```

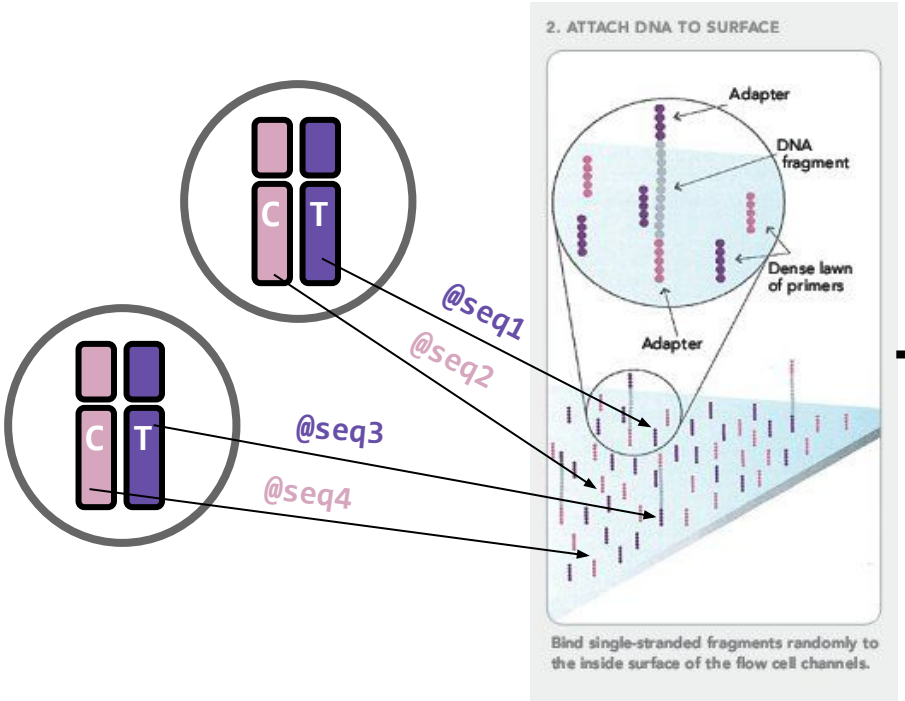
Scenario 2: An individual is homozygous for an "alternate" allele.

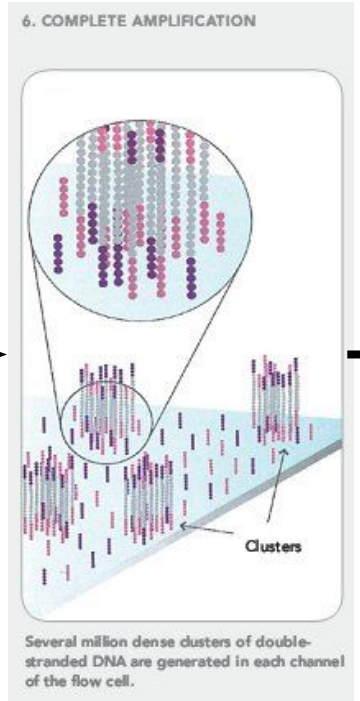


Scenario 3: An individual is heterozygous for an "alternate" allele.



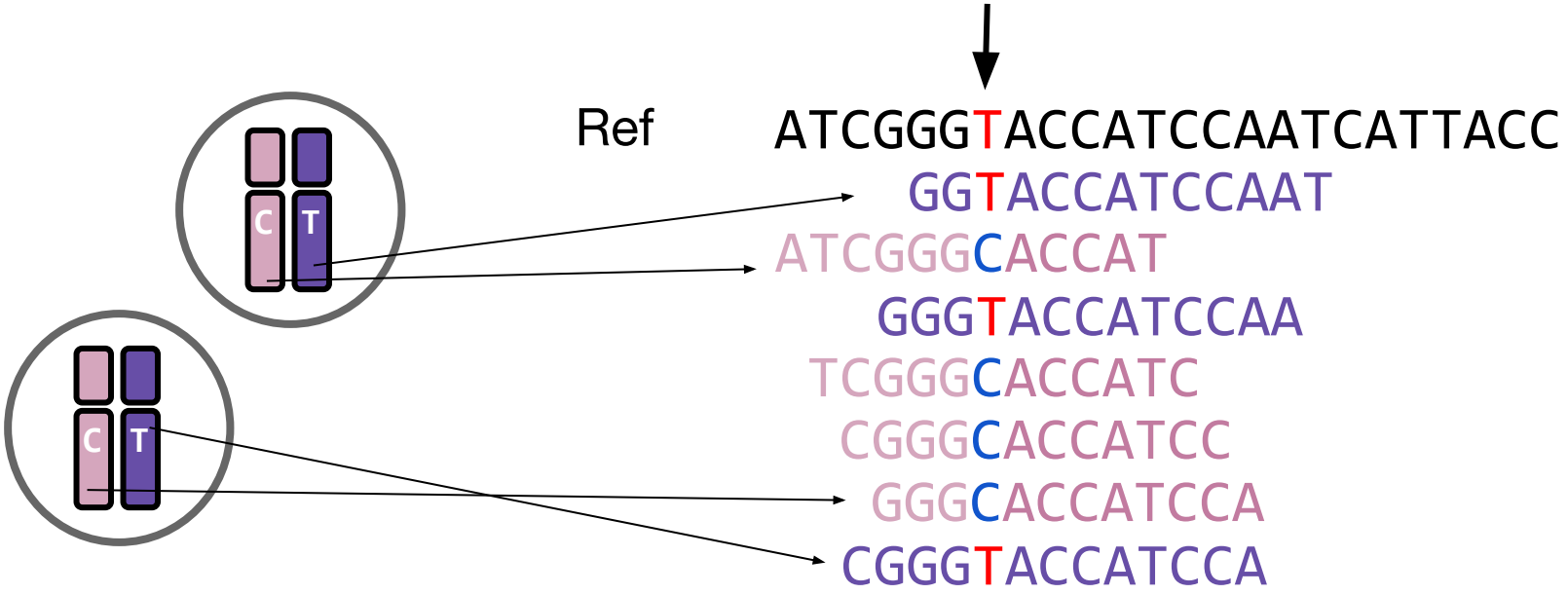
Scenario 3: An individual is heterozygous for an "alternate" allele.






```
@seq1
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
@seq3
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq4
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
```

Scenario 3: An individual is heterozygous for an "alternate" allele.



Why might finding heterozygous variants be harder?

The binomial distribution: adventures in coin flipping

P(heads) = 0.5

P(tails) = 0.5

Thinking about allele sampling with the binomial distribution

The **binomial distribution** with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of *n* independent yes (e.g., "heads" or "reference allele") or no (e.g., "tails", or "alternate allele") experiments, each of which yields success with probability *p*.

The probability of getting exactly k successes in n trials is given by the probability mass function:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

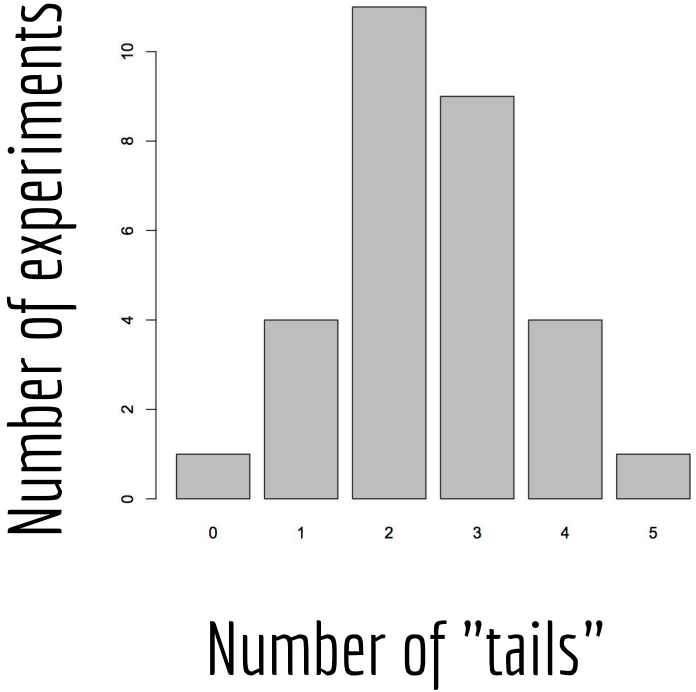
What is the probability of seeing k=1 tails in n=3 flips of a fair coin with the probability of a tail (p) = 0.5?

3 choose 1 = 3; 0.5^1 = 0.5; $(1-0.5)^{(3-1)}$ = 0.25. So.... 3*0.5*0.25 = **0.375**

In R, the function would be: `dbinom(1, size=3, prob=0.5)`

What is the distribution of tails (alternate alleles) do we expect to see after 5 tosses (sequence reads)?

What is the distribution of tails (alternate alleles) do we expect to see after 5 tosses (sequence reads)?



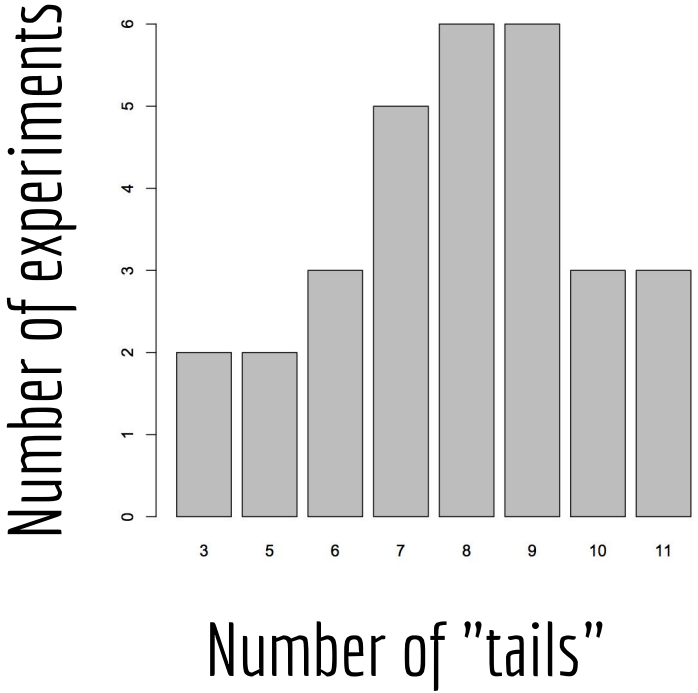
R code:

```
barplot(table(rbinom(30, 5, 0.5)))
```

30 experiments (students tossing coins)
5 tosses each
Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 15 tosses (sequence reads)?

What is the distribution of tails (alternate alleles) do we expect to see after 15 tosses (sequence reads)?



R code:

```
barplot(table(rbinom(30, 15, 0.5)))
```

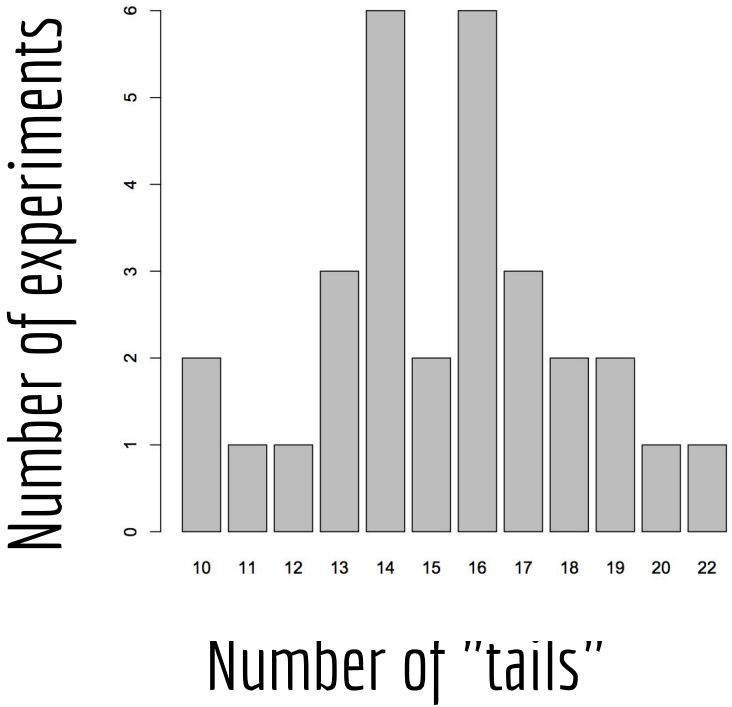
30 experiments (students tossing coins)
15 tosses each
Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 30 tosses (sequence reads)?

Record your result in the following spreadsheet:

https://docs.google.com/spreadsheets/d/1i8sA1KMeYc9UhWTnCg0tLFjCy8x5LIsBITcXrz5La94/edit?usp=sharing

What is the distribution of tails (alternate alleles) do we expect to see after 30 tosses (sequence reads)?

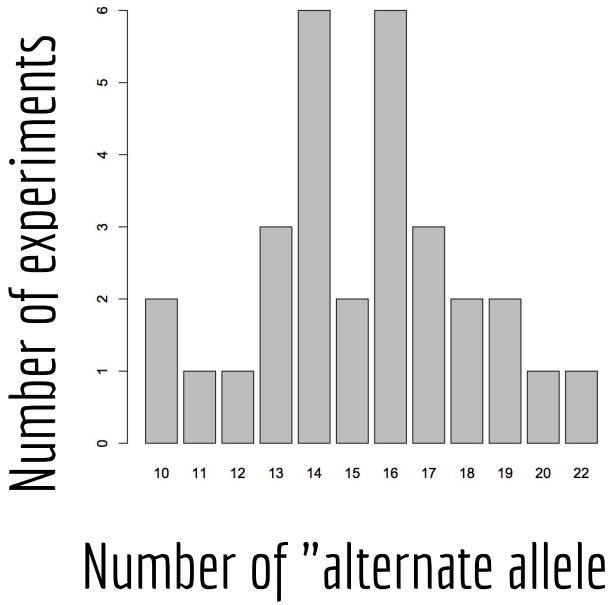


R code:

```
barplot(table(rbinom(30, 30, 0.5)))
```

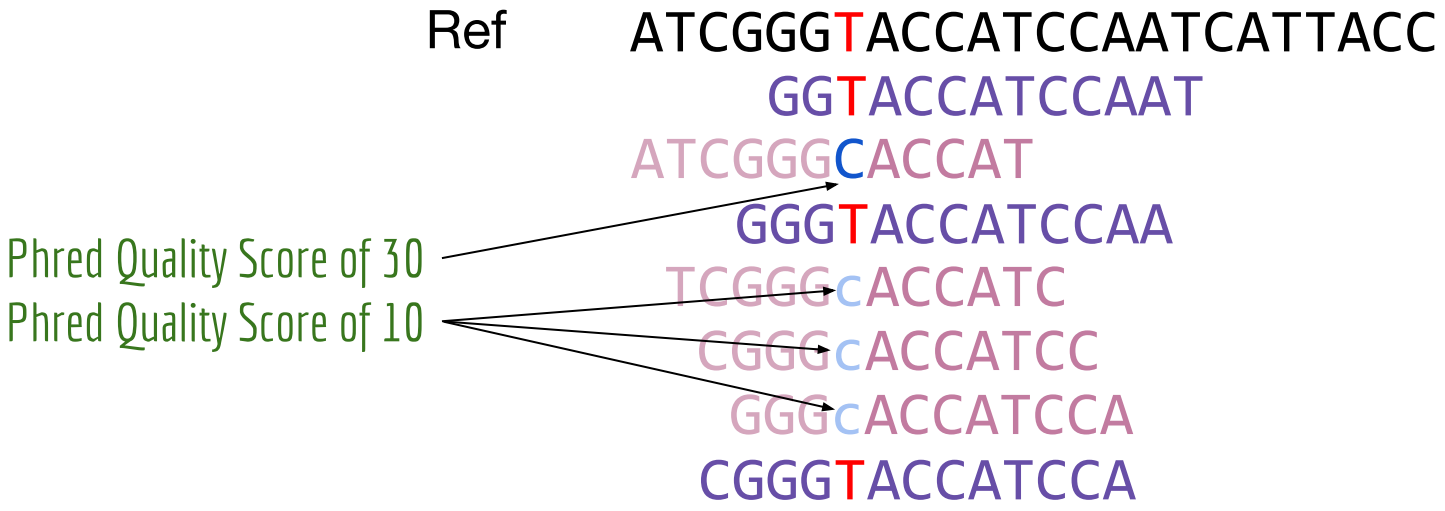
30 experiments (students tossing coins)
30 tosses each
Probability of Tails

So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



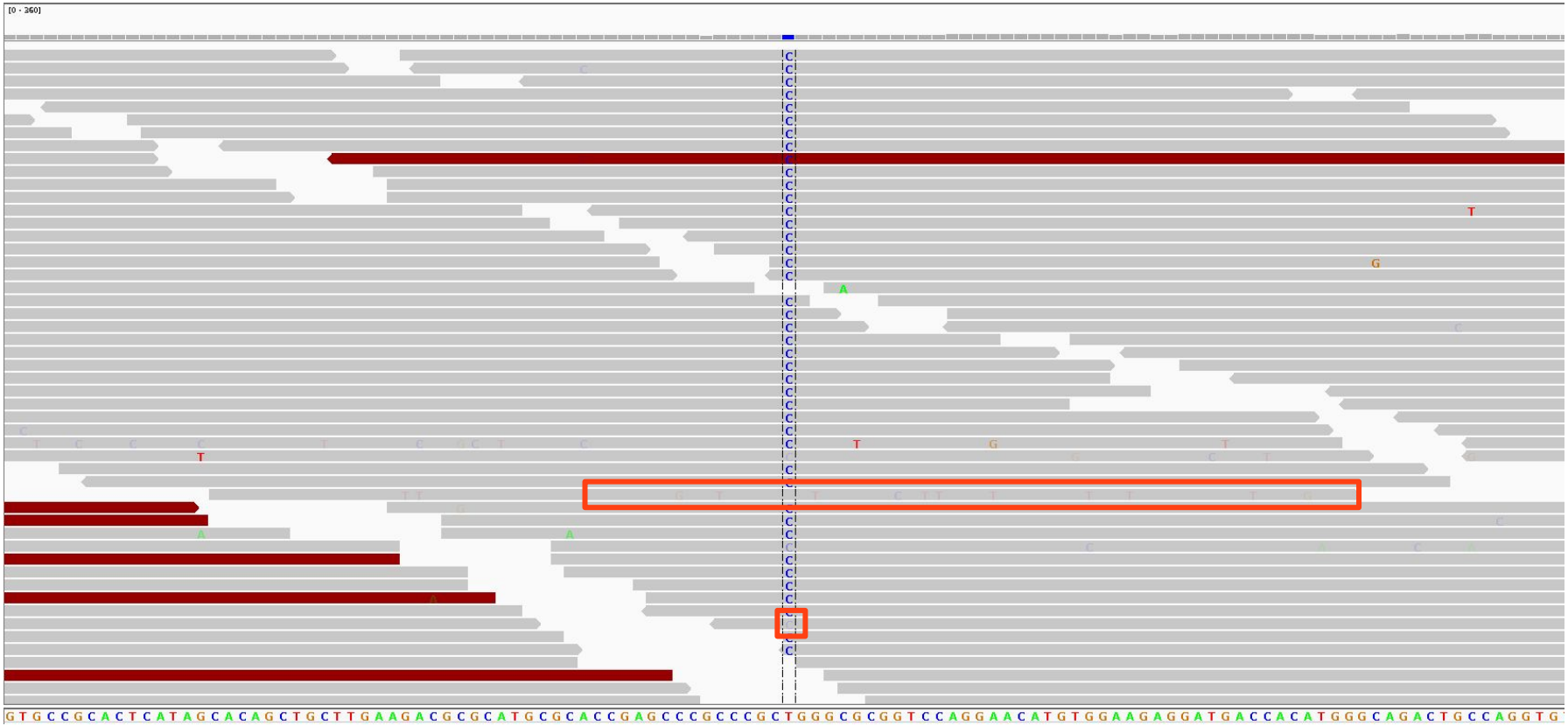
This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to find the majority of heterozygous alleles

Depth tackles the allele sampling issue and lower quality scores



Some real examples of SNPs in IGV: validating variants via manual review

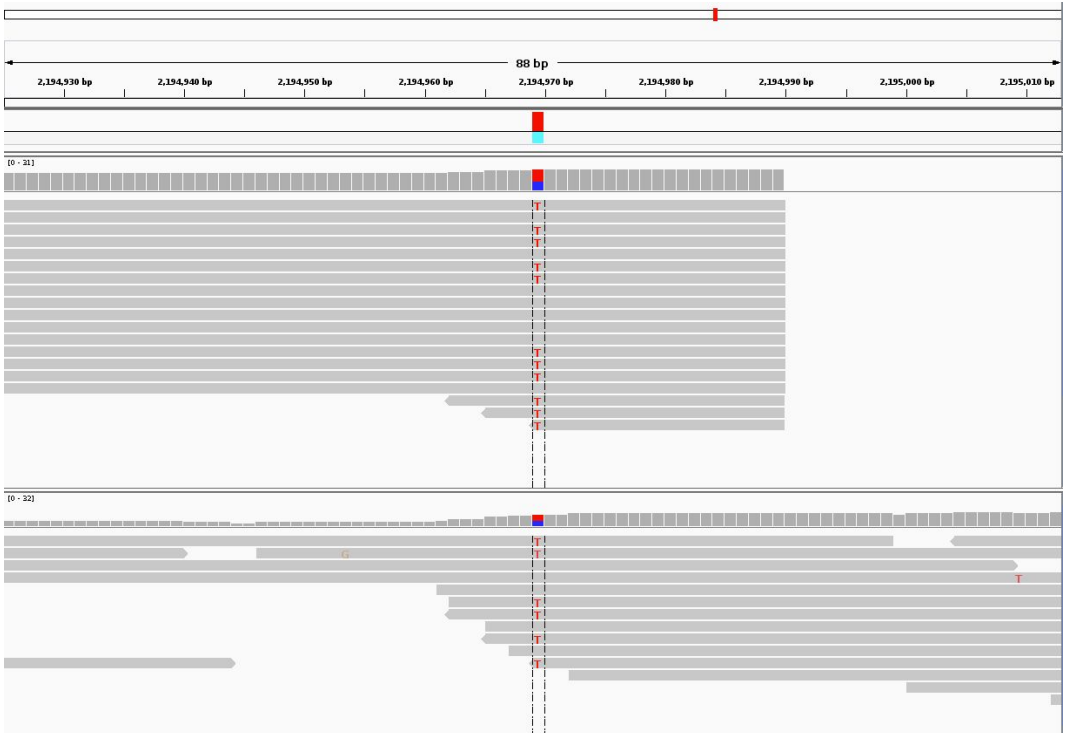
Homozygous for the "C" allele



Heterozygous for the alternate allele

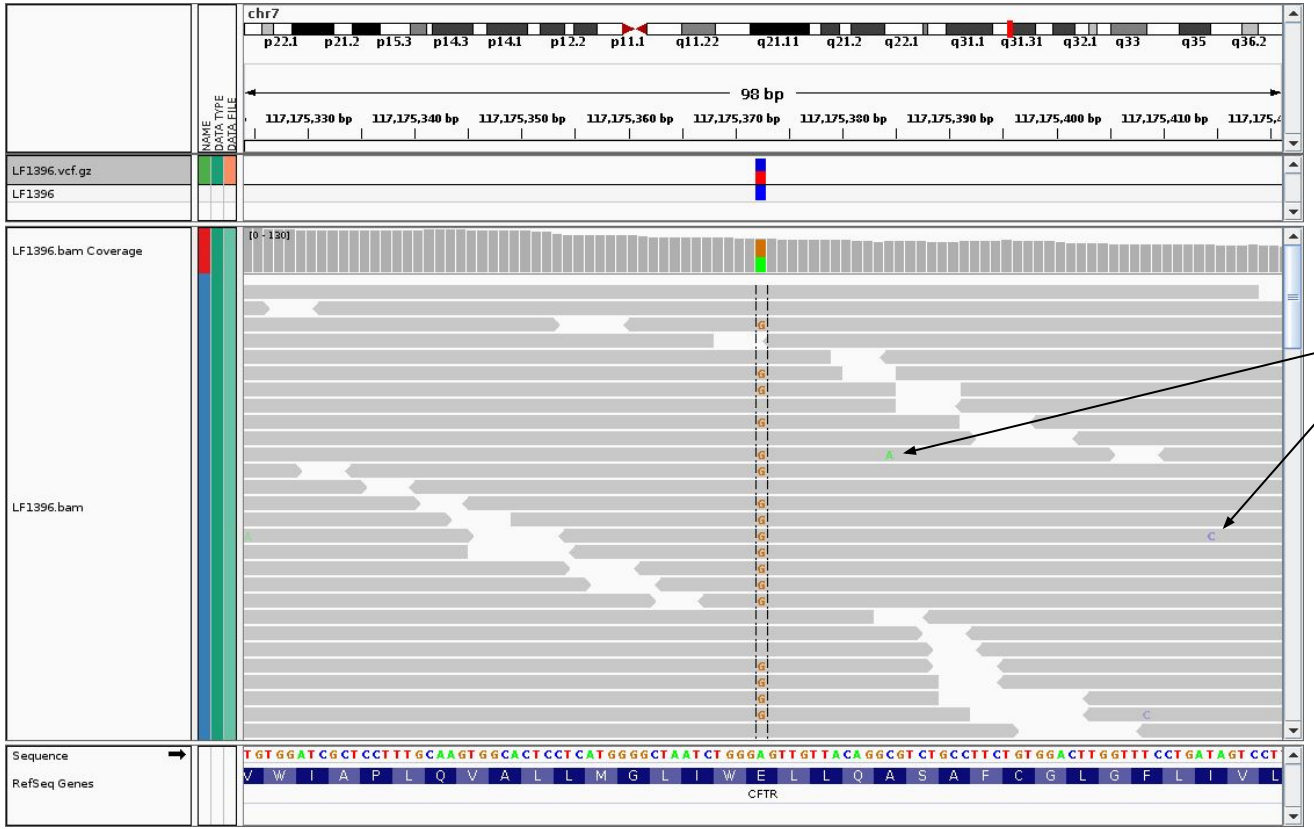
Individual 1

Individual 2



Which genotype prediction would you have more confidence in?

Sequencing errors fall out as noise (most of the time)

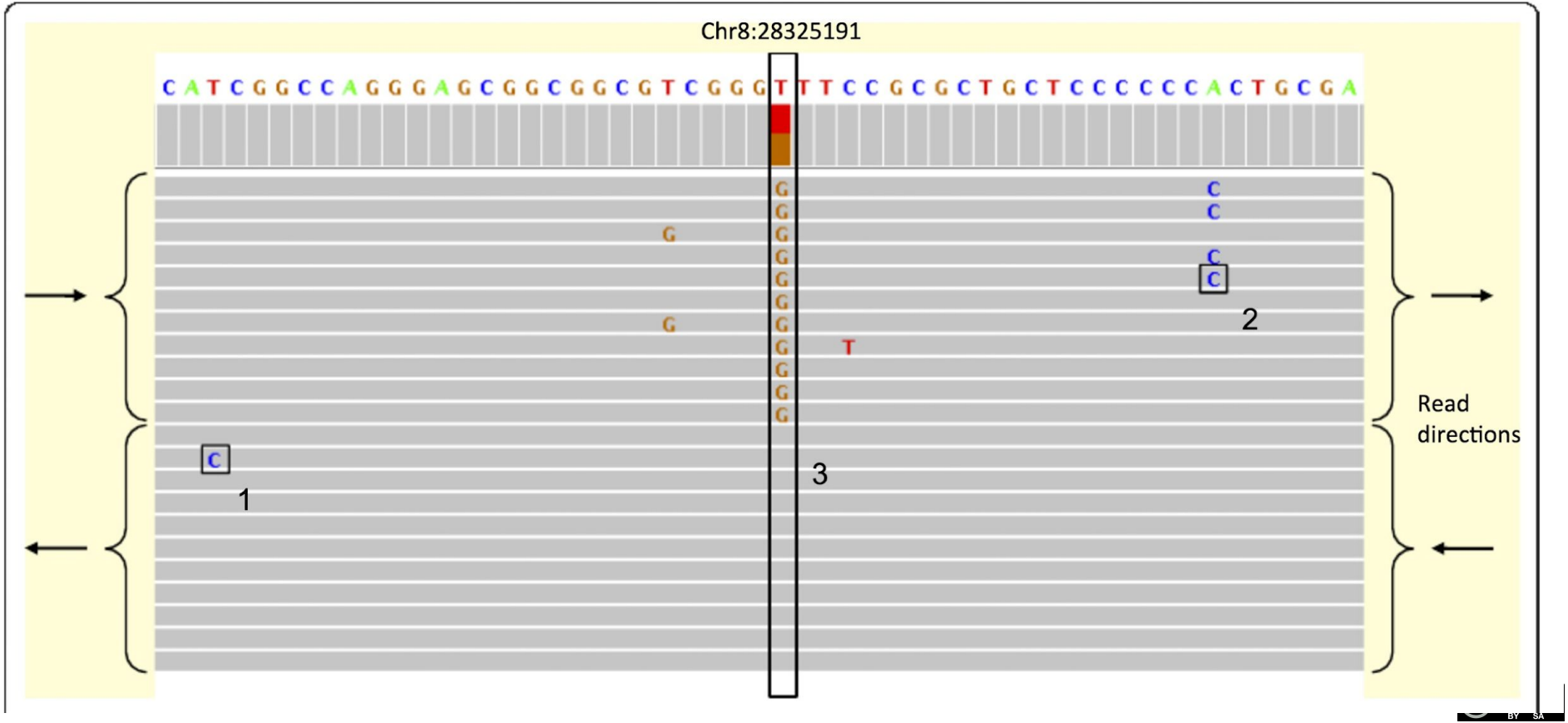


Sequencing errors

https://jchoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

It is not always so easy

Random versus systematic error



Random versus systematic error

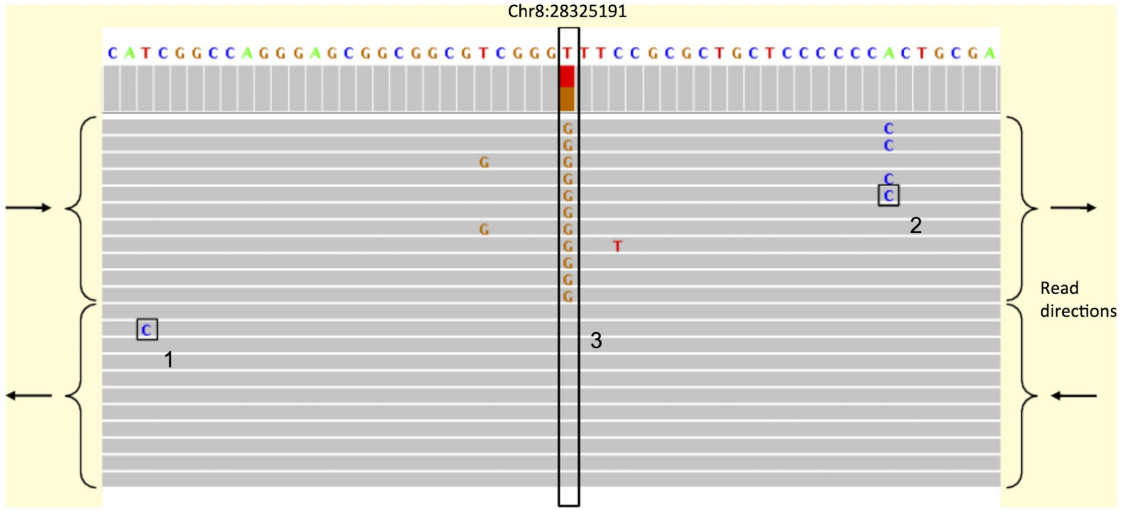
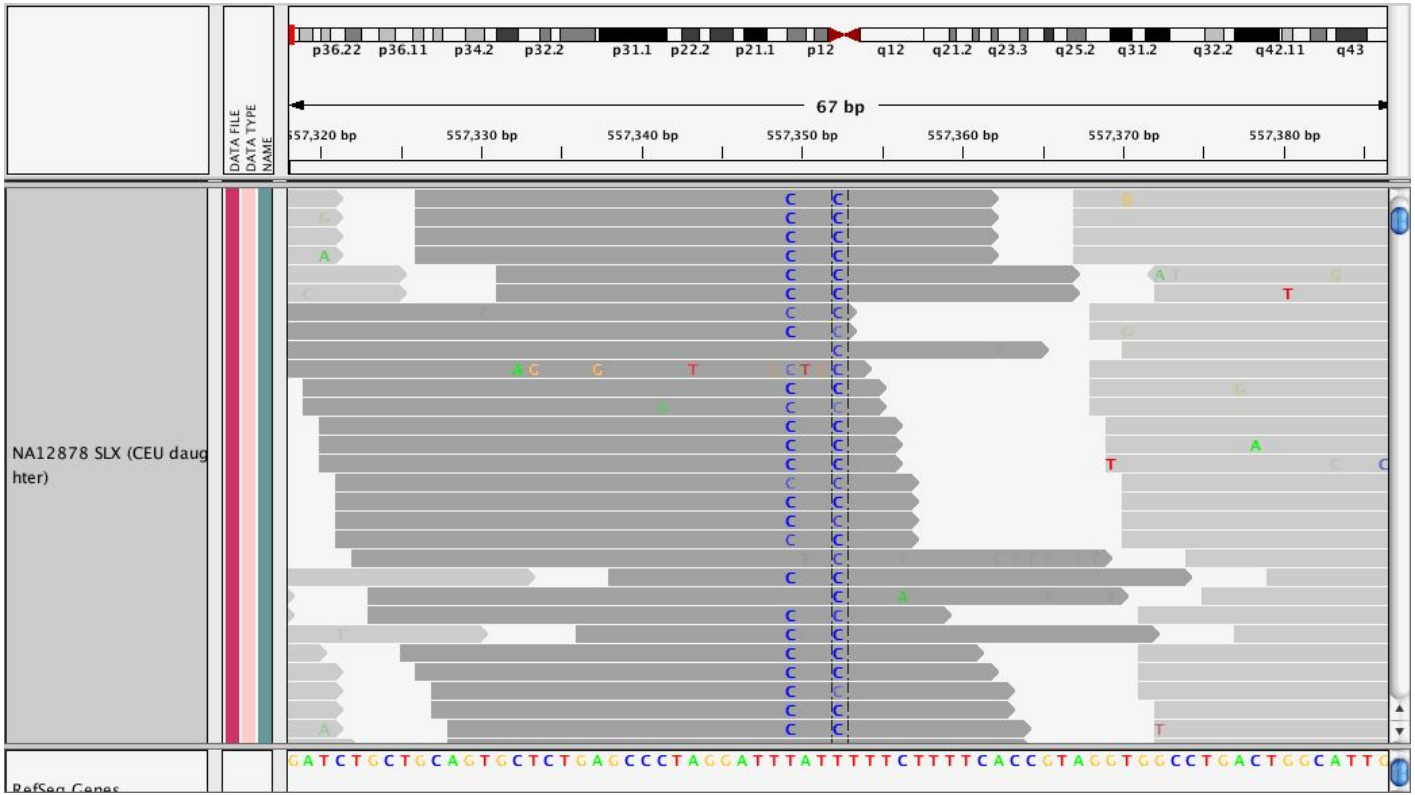
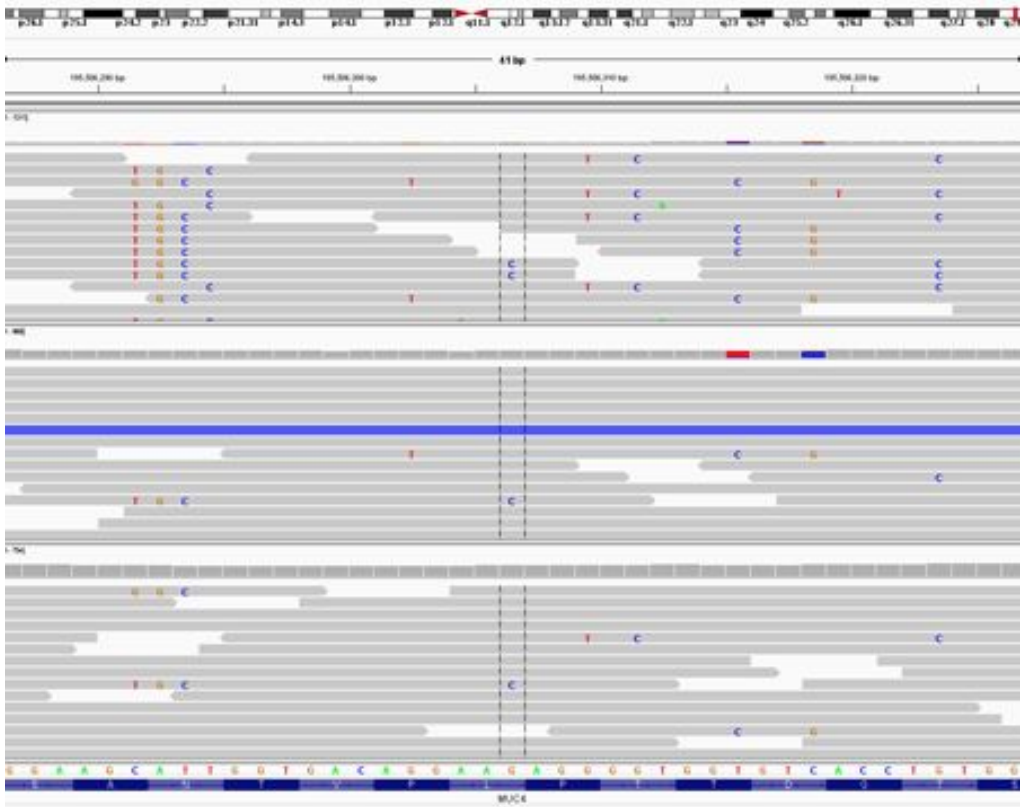


Figure 1 Types of errors. A screenshot from the IGV browser [21] showing three types of error in reads from an Illumina sequencing experiment: (1) A random error likely due to the fact that the *position* is close to the end of the read. (2) Random error likely due to *sequence specific* error- in this case a sequence of Cs are probably inducing errors at the end of the low complexity repeat. (3) *Systematic error*: although it is likely that the GGT sequence motif and the GGC motifs before it created phasing problems leading to the errors, the extent of error is not explained by a random error model. In this case, all the base calls in one direction are wrong as revealed by the 11 overlapping mate-pairs. In particular, all differences from the reference genome are base-call errors, verified by the mate-pair reads, which do not differ from the reference. Given the background error rate, the probability of observing 11 *error-pairs* at a single location, given that 11 mate-pair reads overlap the location, is 1.5×10^{-26}. Moreover, given the presence of such errors at a single location, the probability that all of the errors occur on the same strand (i.e., on the forward mate pair) is $\frac{1}{1024} = 0.00098$. Note that the IGV browser made an incorrect SNP call at the systematic error site (colored bar in top panel).

Strand bias from PCR



Pileups of many differences from paralogy



RESEARCH ARTICLE | OPEN ACCESS

FLAGS, frequently mutated genes in public exomes

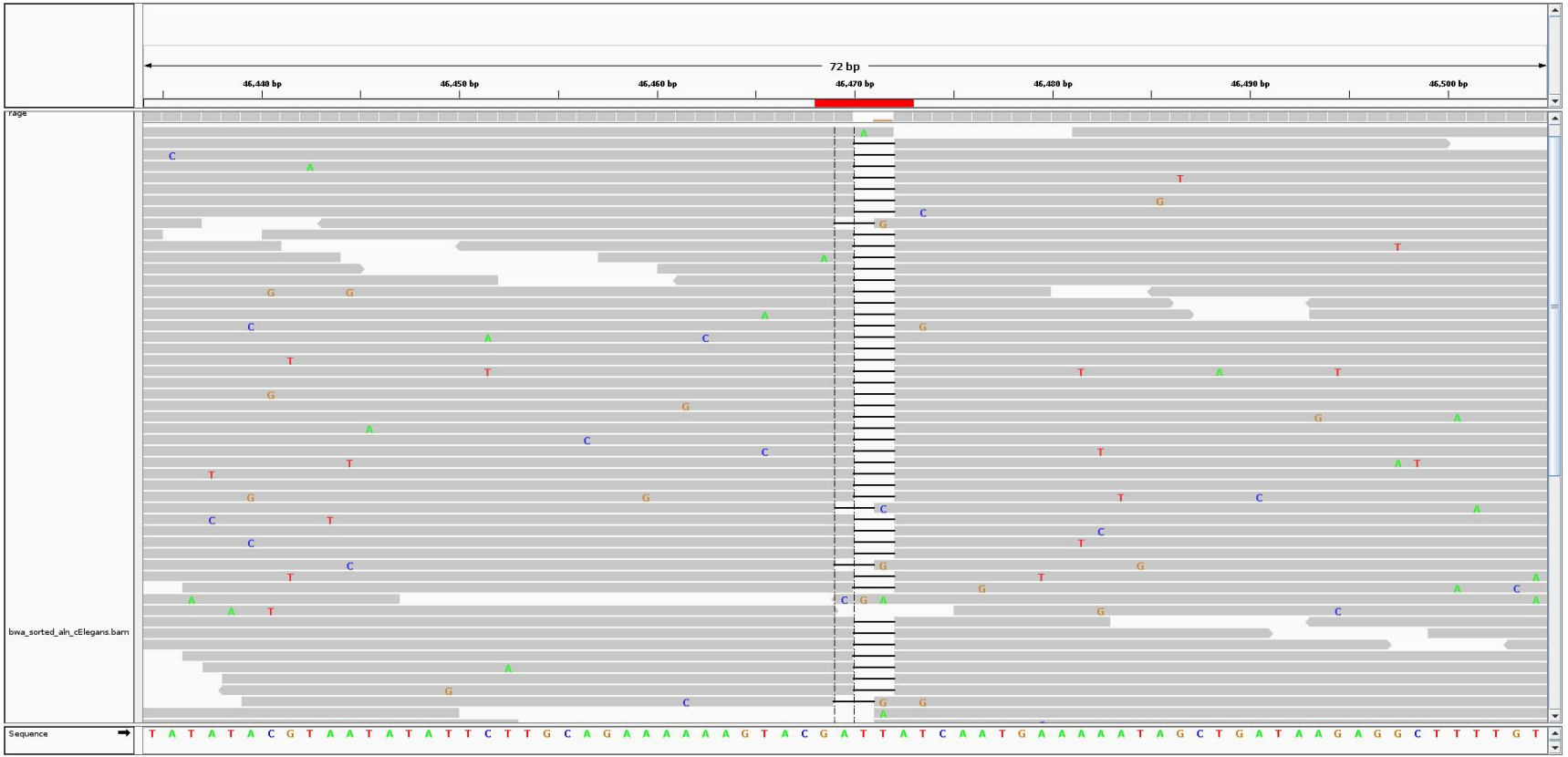
Casper Shyr, Maja Tarailo-Graovac, Michael Gottlieb, Jessica JY Lee, Clara van Karnebeek and Wyeth W Wasserman

BMC Medical Genomics 2014 7:64 | DOI: 10.1186/s12920-014-0064-y

Received: 16 June 2014 | Accepted: 24 October 2014 | Published: 3 December 2014

Open Peer Review reports

Calling INDELs is _much_ harder than SNPs



INDEL "realignment"

Some excellent resources to learn about manual review

Griffith Lab guides to manual review in IGV:

- https://rnabio.org/module-02-alignment/0002/04/01/IGV/
- Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples