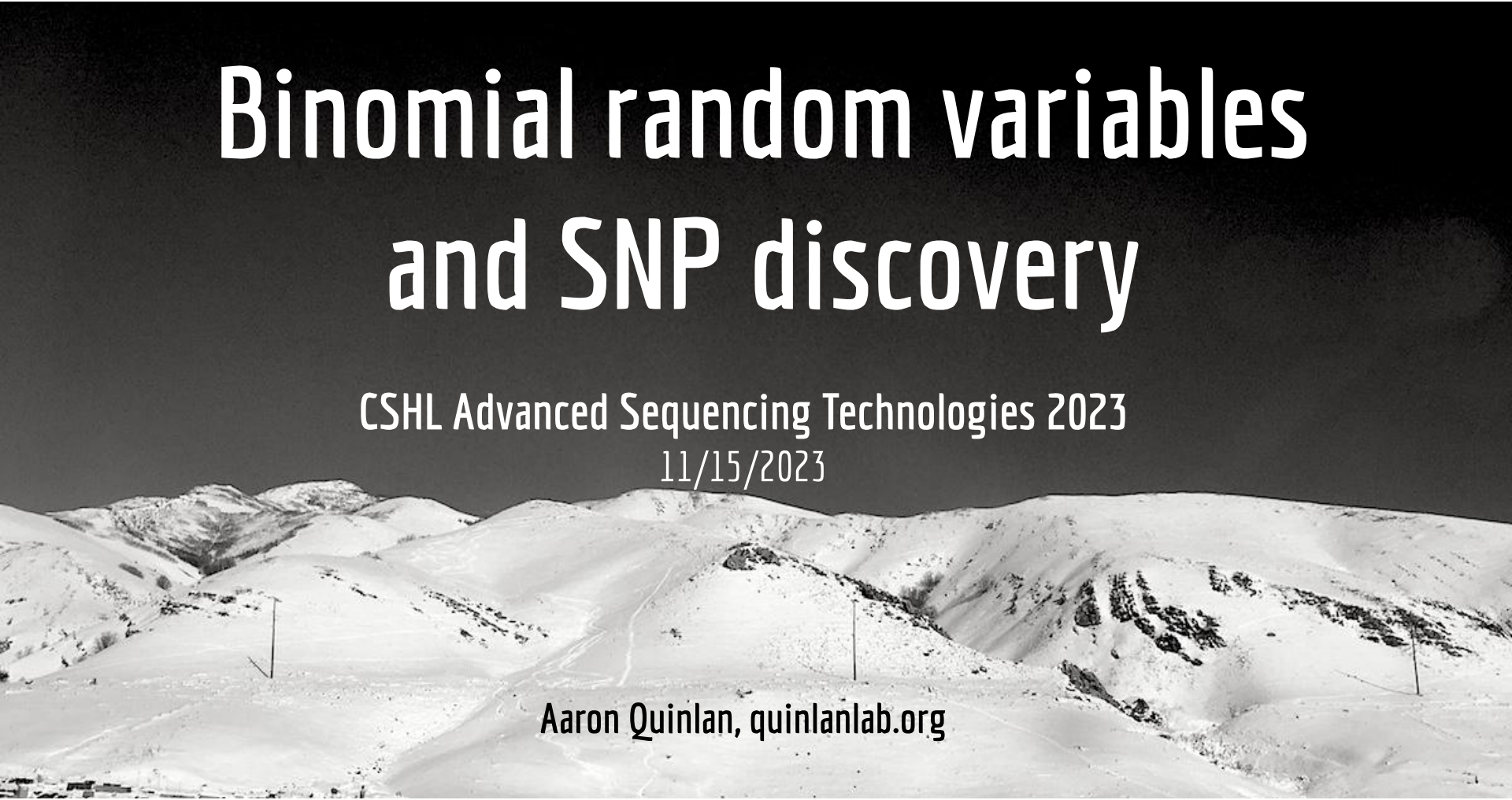
Binomial random variables and SNP discovery

CSHL Advanced Sequencing Technologies 2023

11/15/2023



Aaron Quinlan, quinlanlab.org

Thought experiment

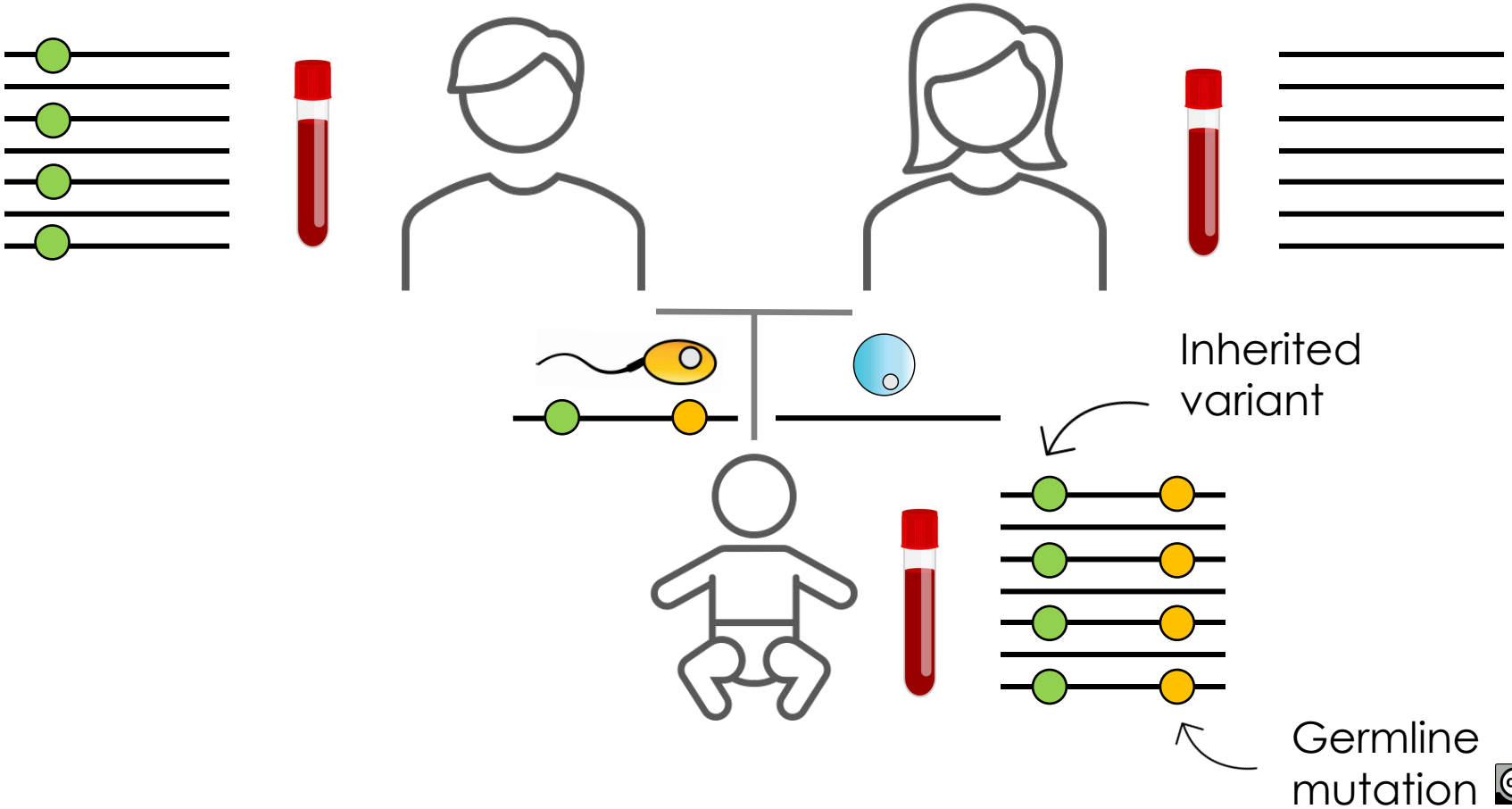
Mutations arise on sperm and egg prior to fertilization.

Such mutations are observed in the offspring's DNA, but not in the parent: "**germline mutations**"

How many should we expect in a typical child?

CC BY SA

Finding mutations with family genome sequencing



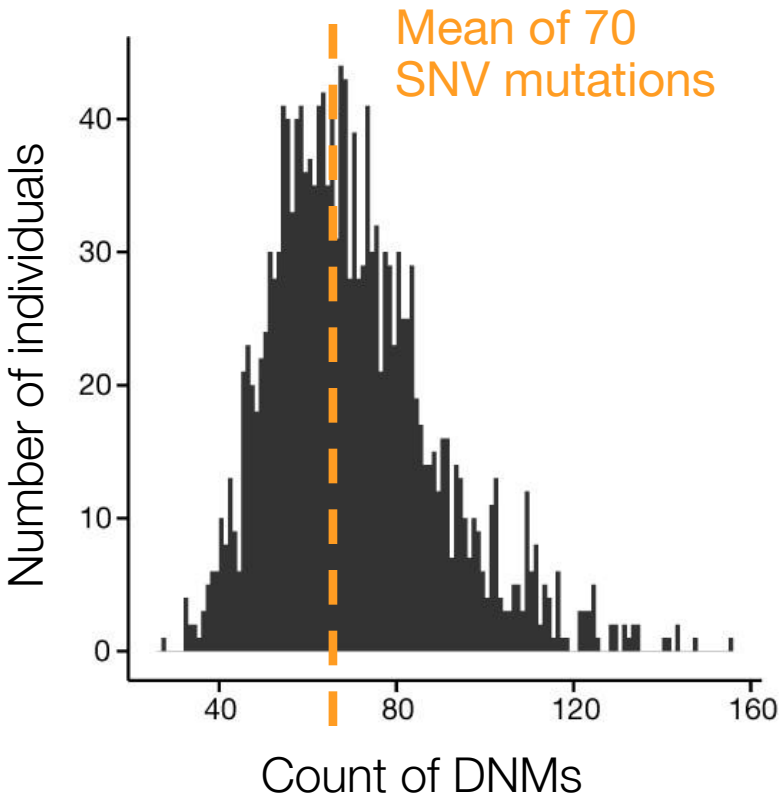
CC BY SA

Germline mutation rate

~1 DNM per 100 million per generation
(1 x 10-8 / bp / gen)

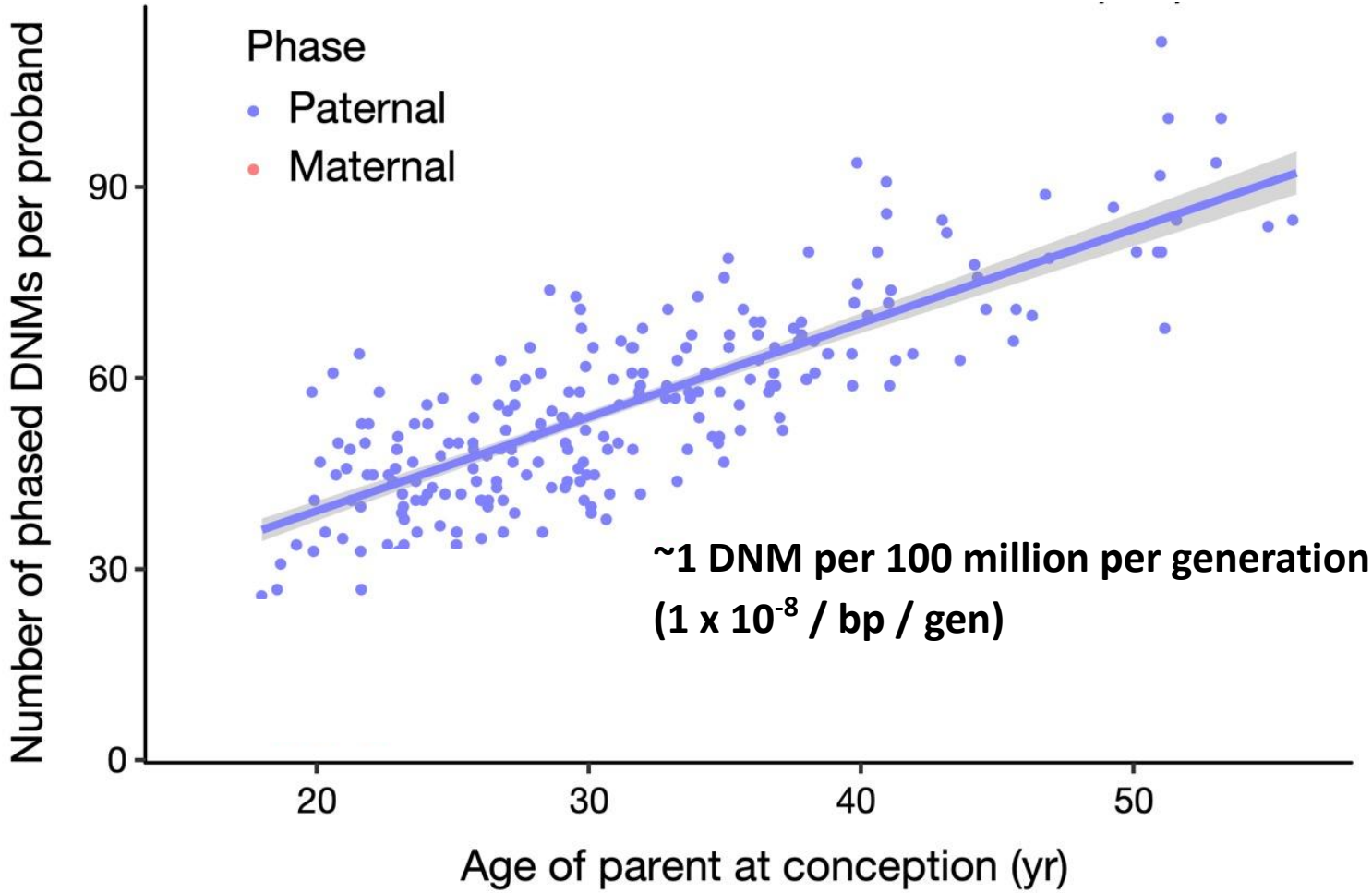
CC BY SA

A typical human has 70 de novo SNV mutations. However, the mutation burden is variable.



Jonnson et al, Nature, 2017

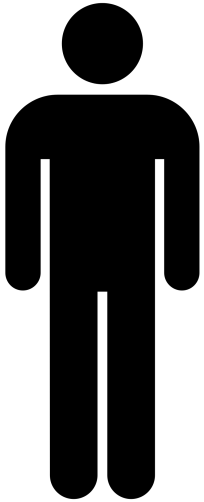
CC BY SA



Sasani et al, eLife, 2019

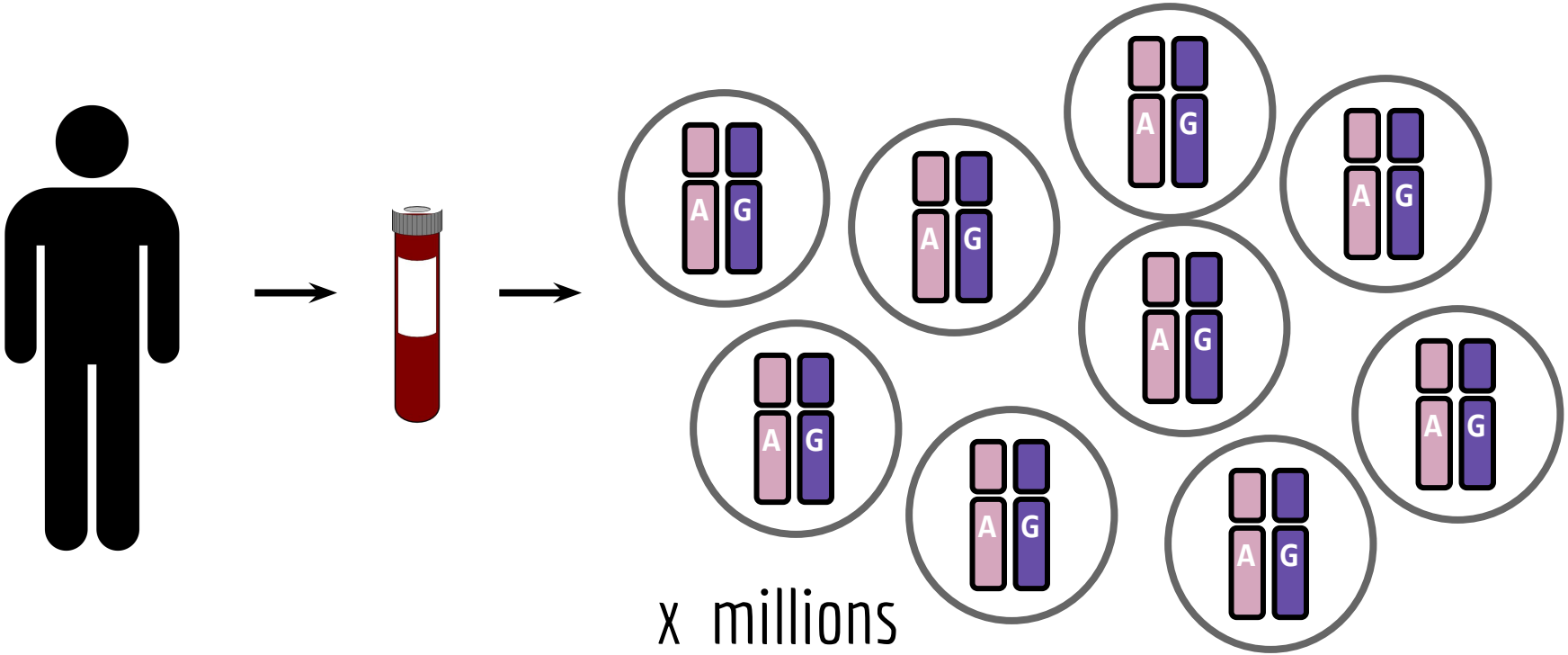
CC BY SA

Goal: find all germline mutations in an individual's diploid genome.



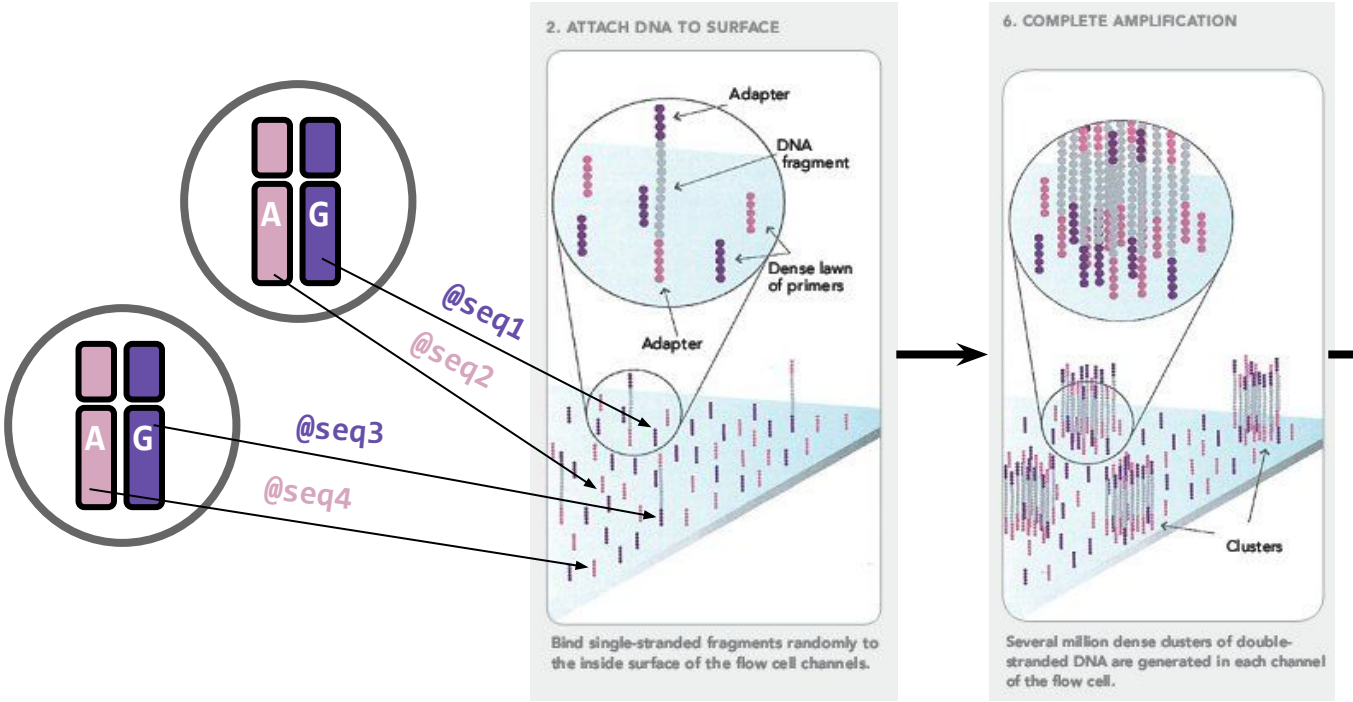
CC BY SA

Find all germline mutations by sequencing DNA from millions of cells



CC BY SA

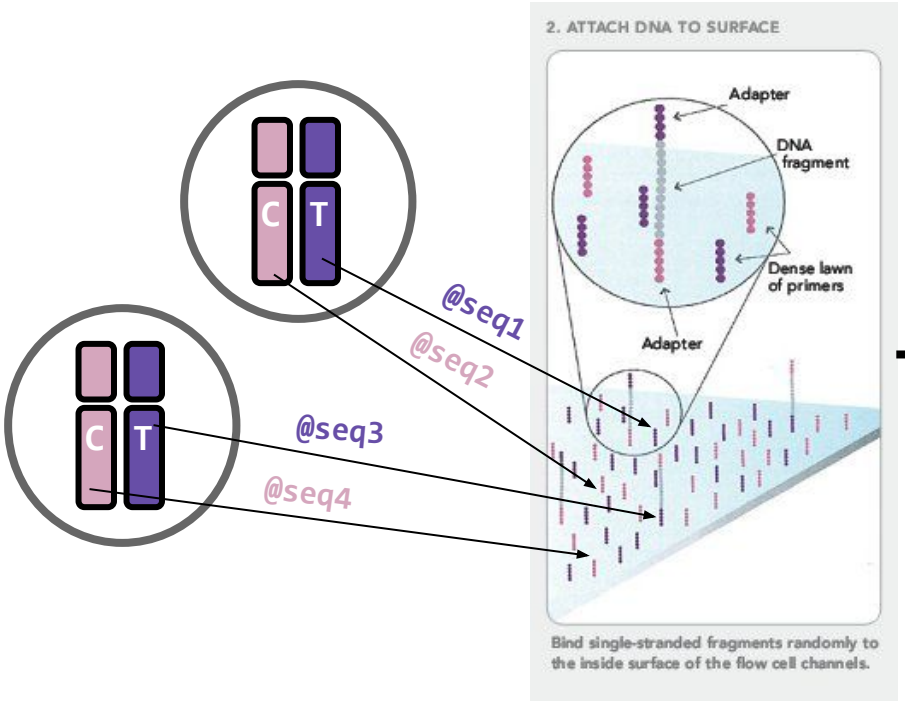
Each DNA cluster is amplified from a single strand from a single haploid chromosome from a single cell.




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

CC BY SA

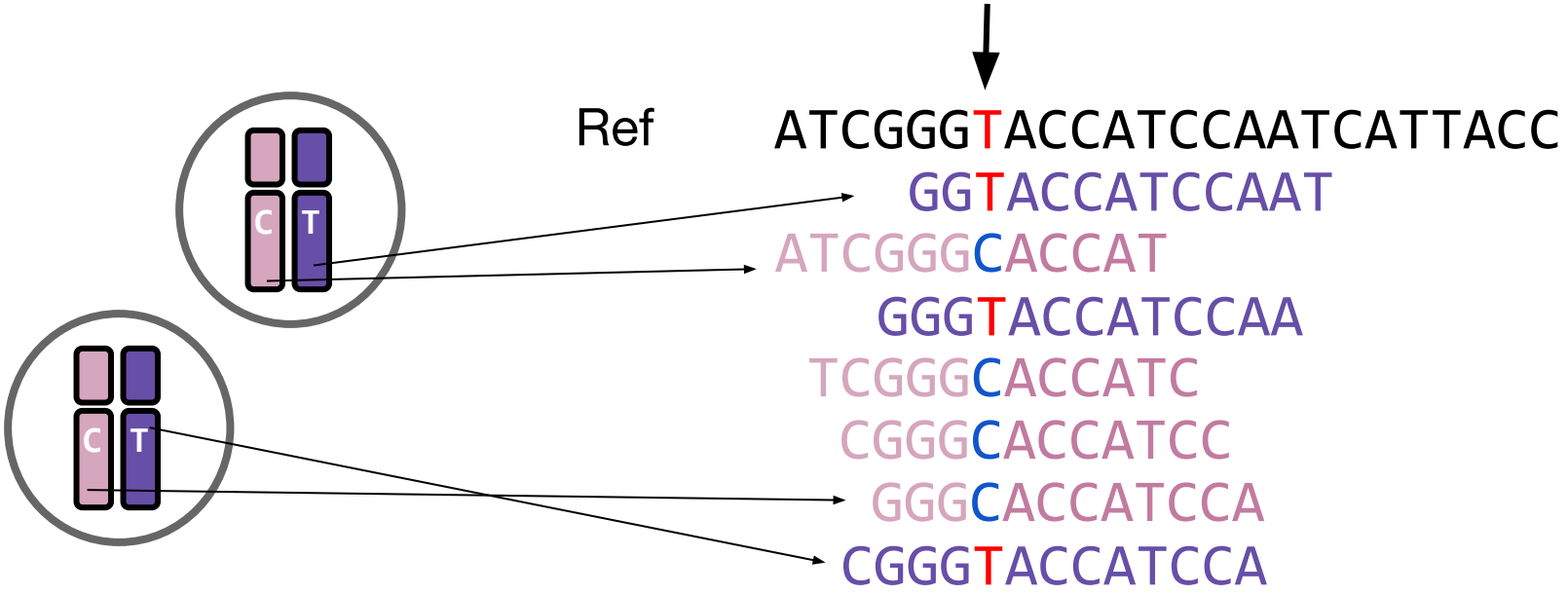
Scenario: An individual is heterozygous for an "alternate" allele.




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

CC BY SA

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



Why might finding heterozygous variants be harder?

CC BY SA

Binomial random variables: adventures in coin flipping

P(heads) = 0.5

P(tails) = 0.5

CC BY SA

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**

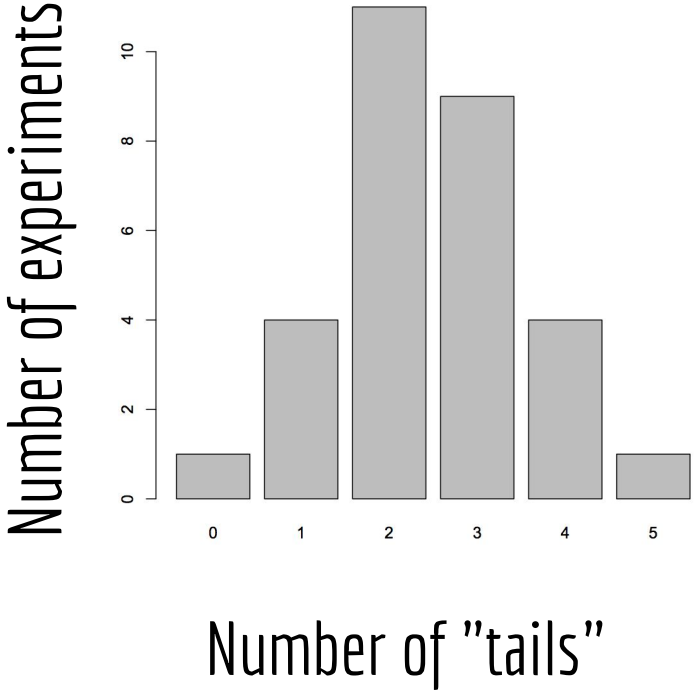
In R, the function would be: `dbinom(1, size=3, prob=0.5)`

CC BY SA

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

CC BY SA

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



R code:

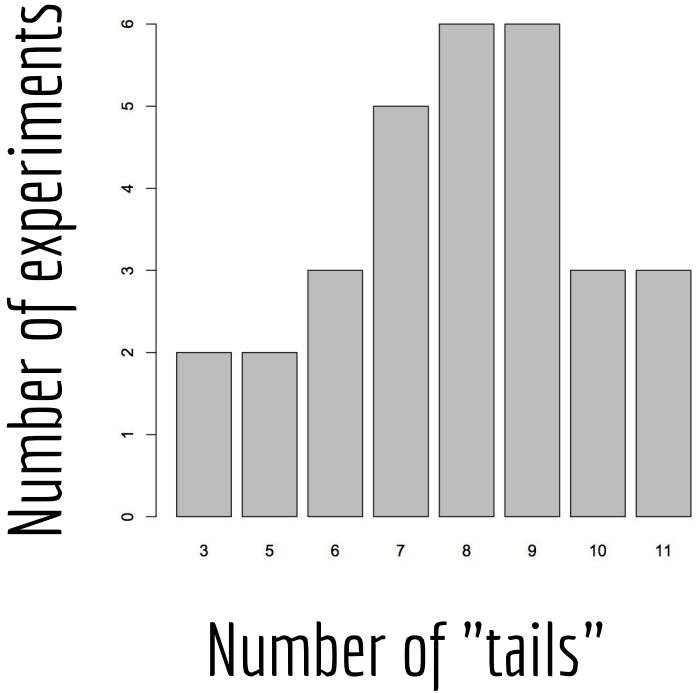
```
barplot(table(rbinom(30, 5, 0.5)))
```

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)?

CC BY SA

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

CC BY SA

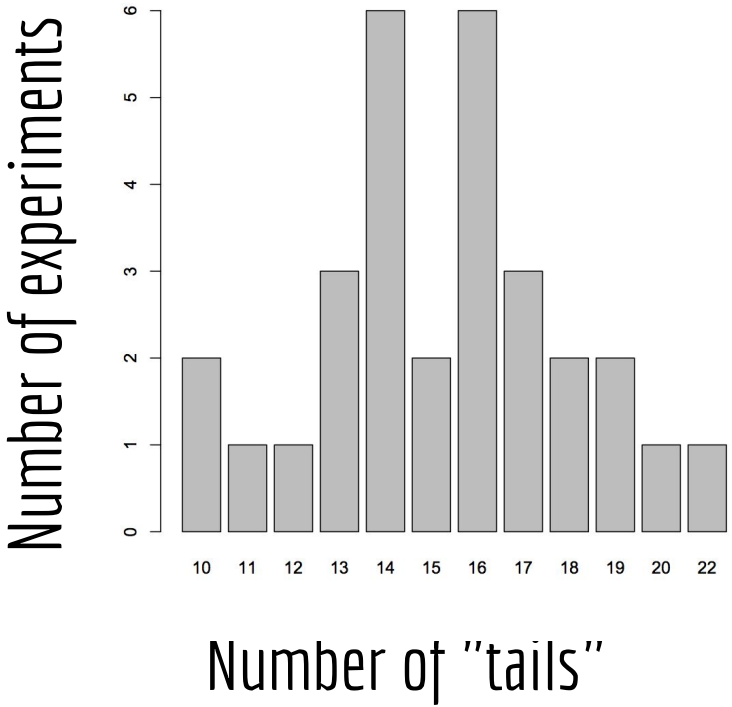
What is the distribution of tails (alternate alleles) do we expect to see after 30 tosses (sequence reads)?

Record your result in the following spreadsheet:

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

CC BY SA

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



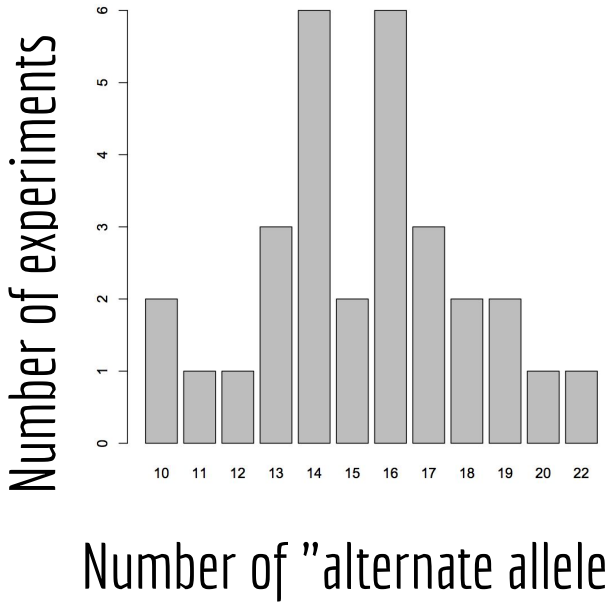
R code:

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

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

CC BY SA

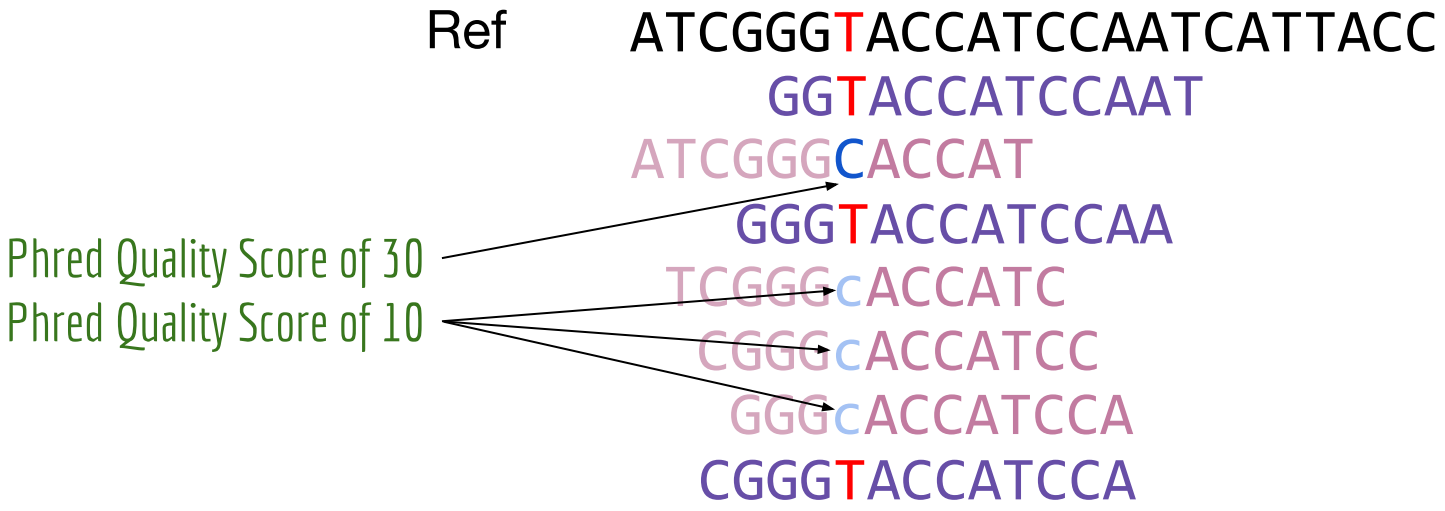
So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



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

CC BY SA

Depth tackles the allele sampling issue and lower quality scores

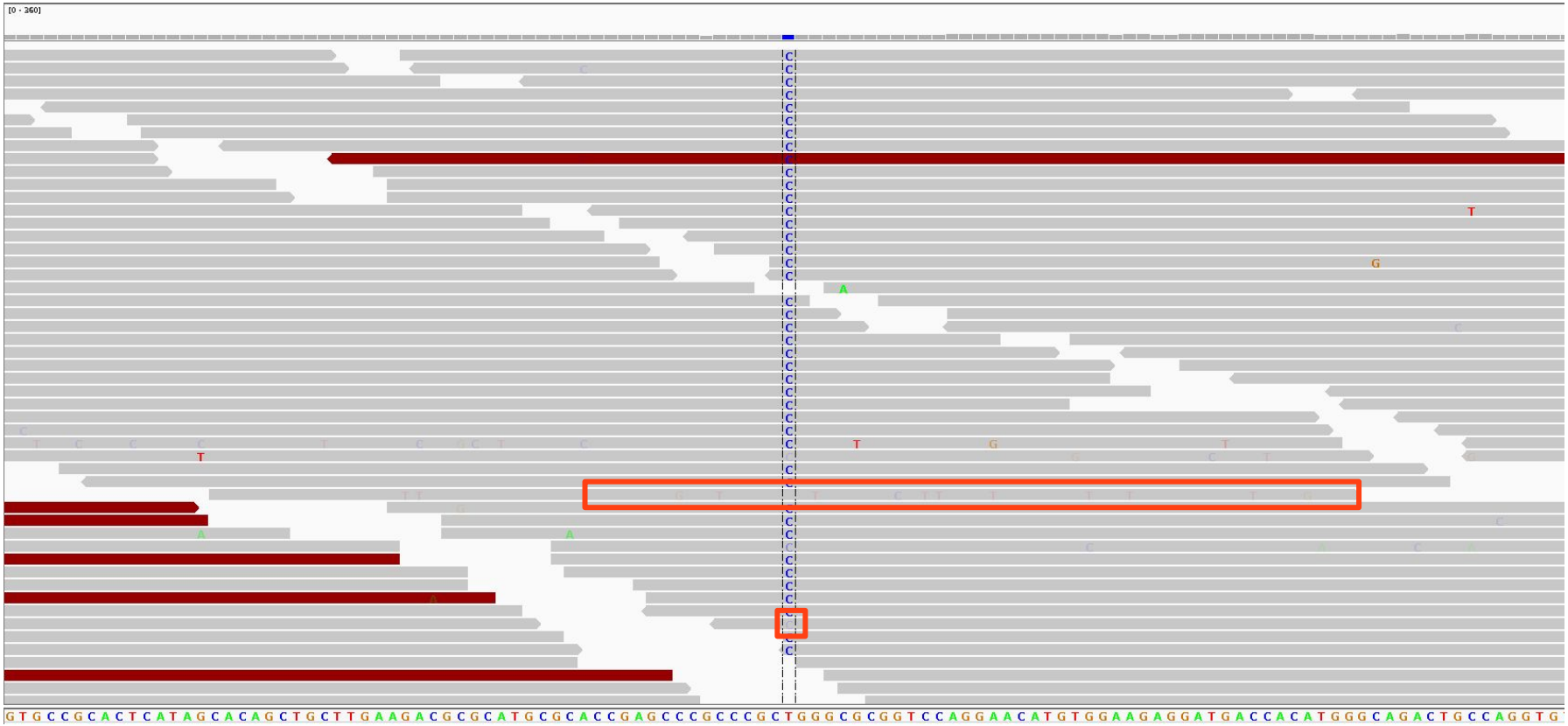


CC BY SA

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

CC BY SA

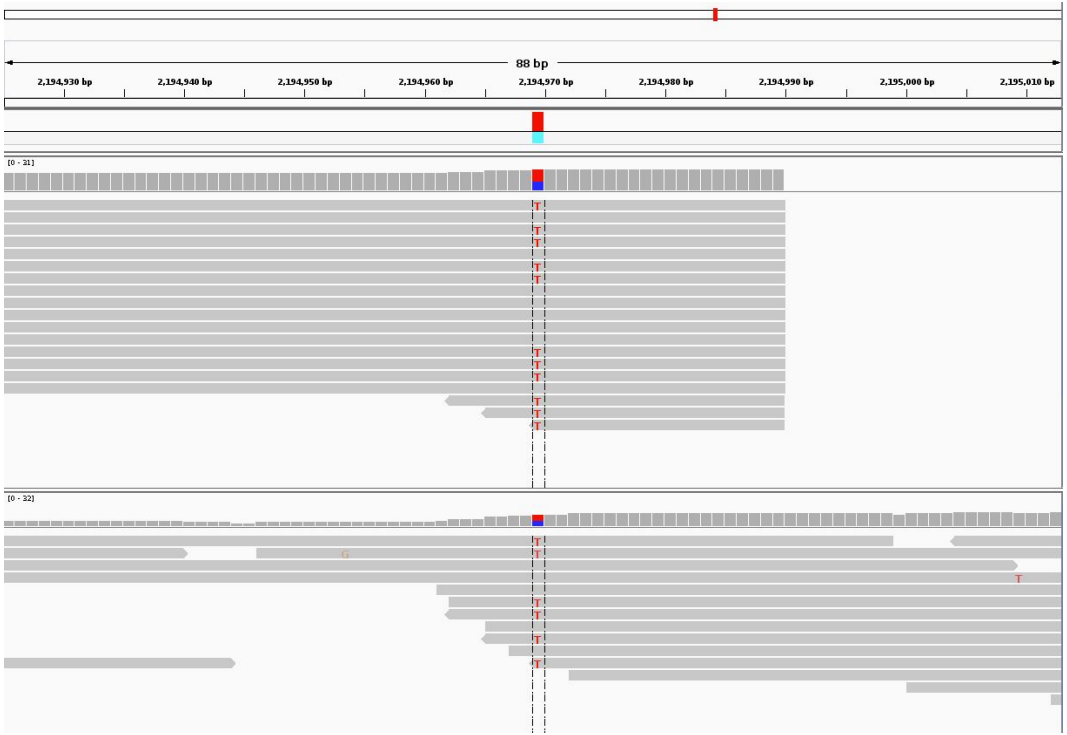
Homozygous for the "C" allele



Heterozygous for the alternate allele

Individual 1

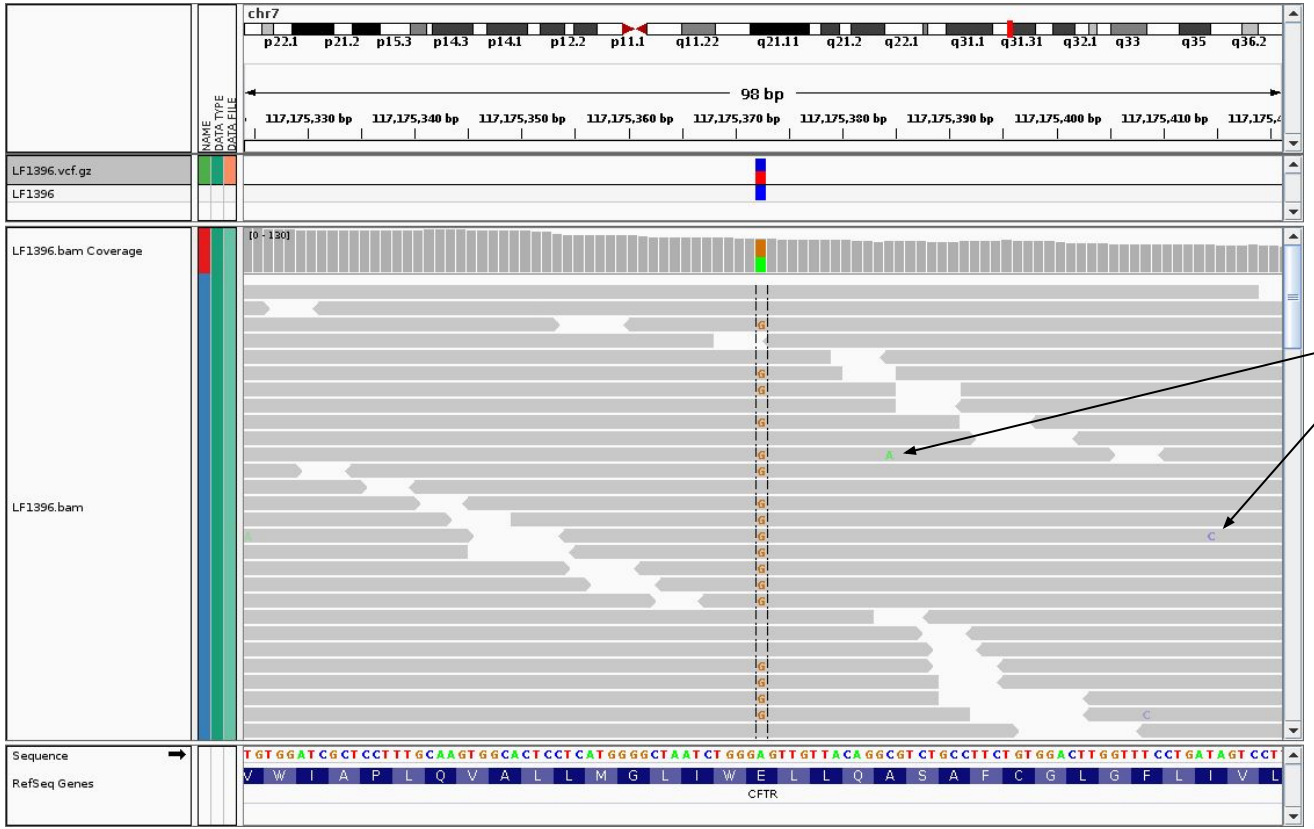
Individual 2



Which genotype prediction would you have more confidence in?

CC BY SA

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



Sequencing errors

CC BY SA

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