Program of Medical Informatics
Faculty of Computers and Information
Mansoura University

Detecting Genetic Variations -I

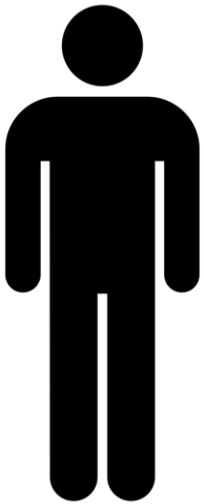
(Genomics)

Sara El-Metwally, Ph.D.
Faculty of Computers and Information,
Mansoura University, Egypt.

Email: sarah_almetwally4@mans.edu.eg
sara.elmetwally.2007@gmail.com

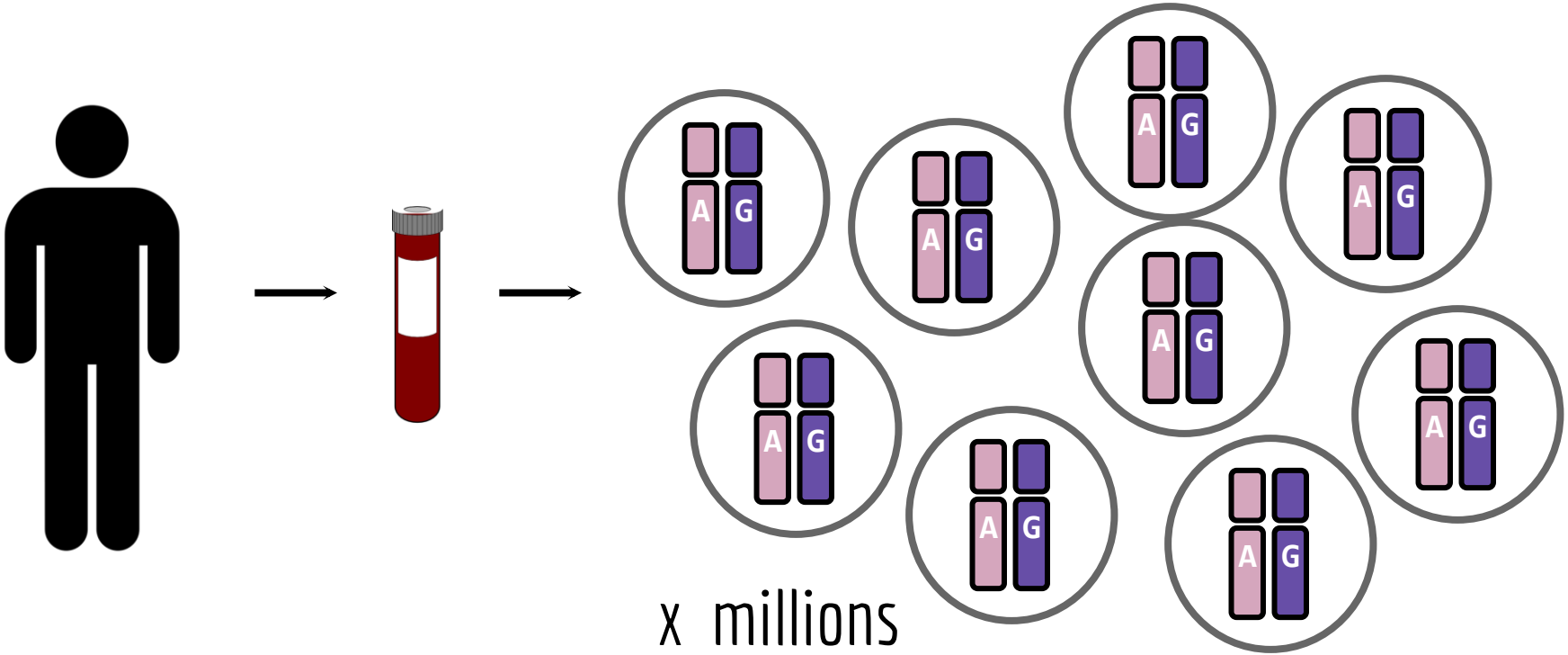
Office: Faculty of CIS, third floor

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



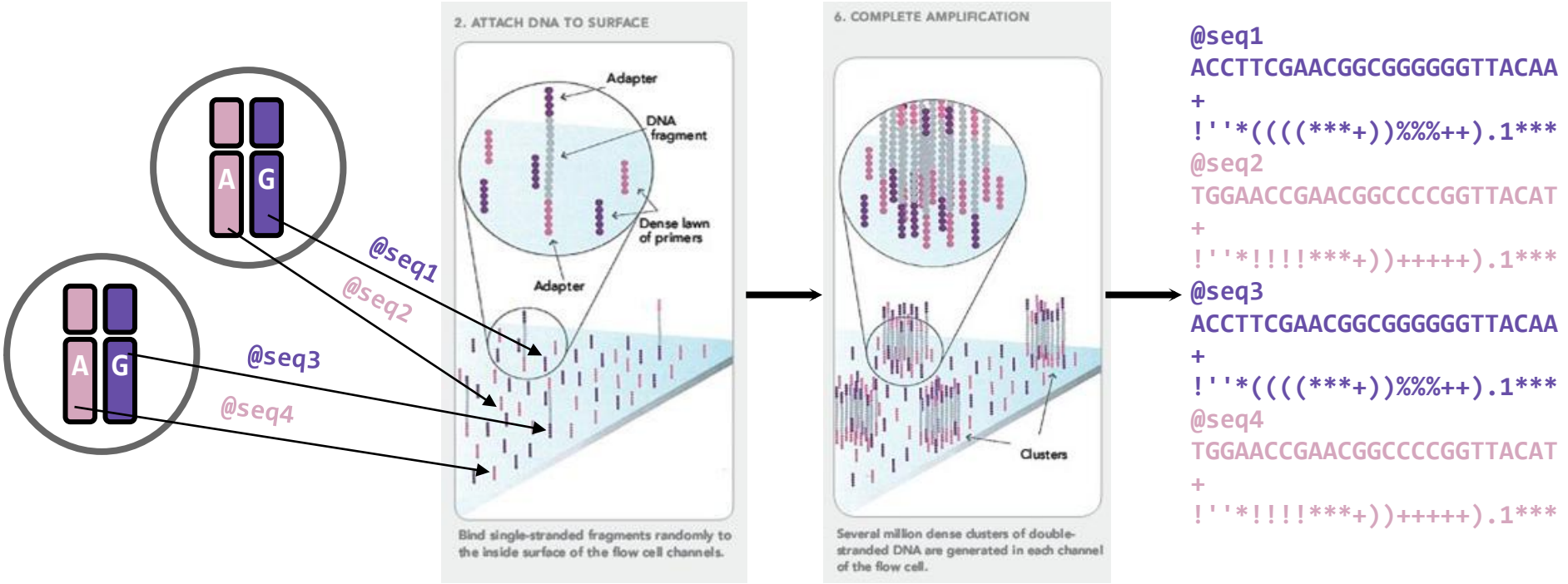
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Find inherited genetic variation by sequencing DNA from millions of cells



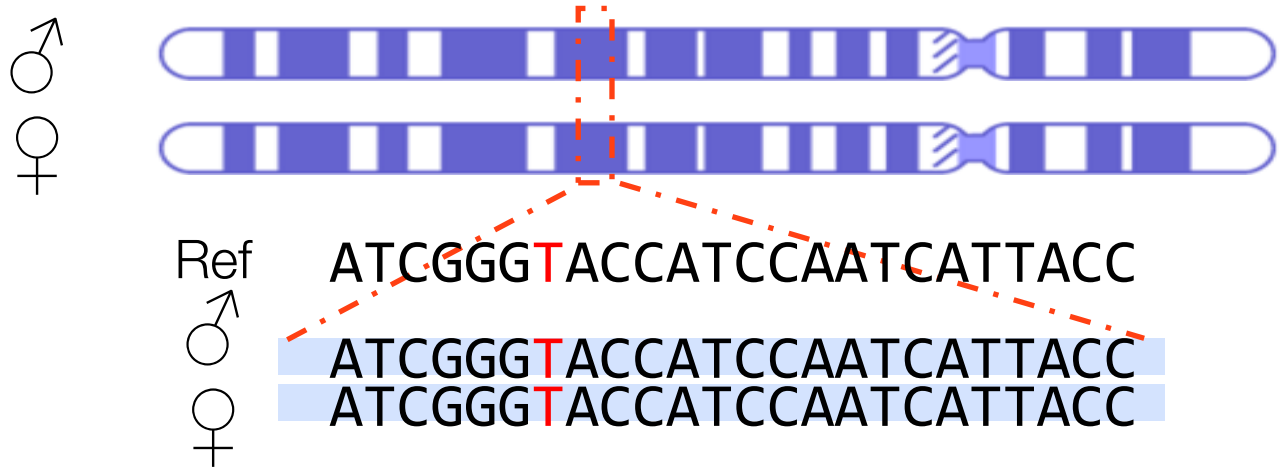
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



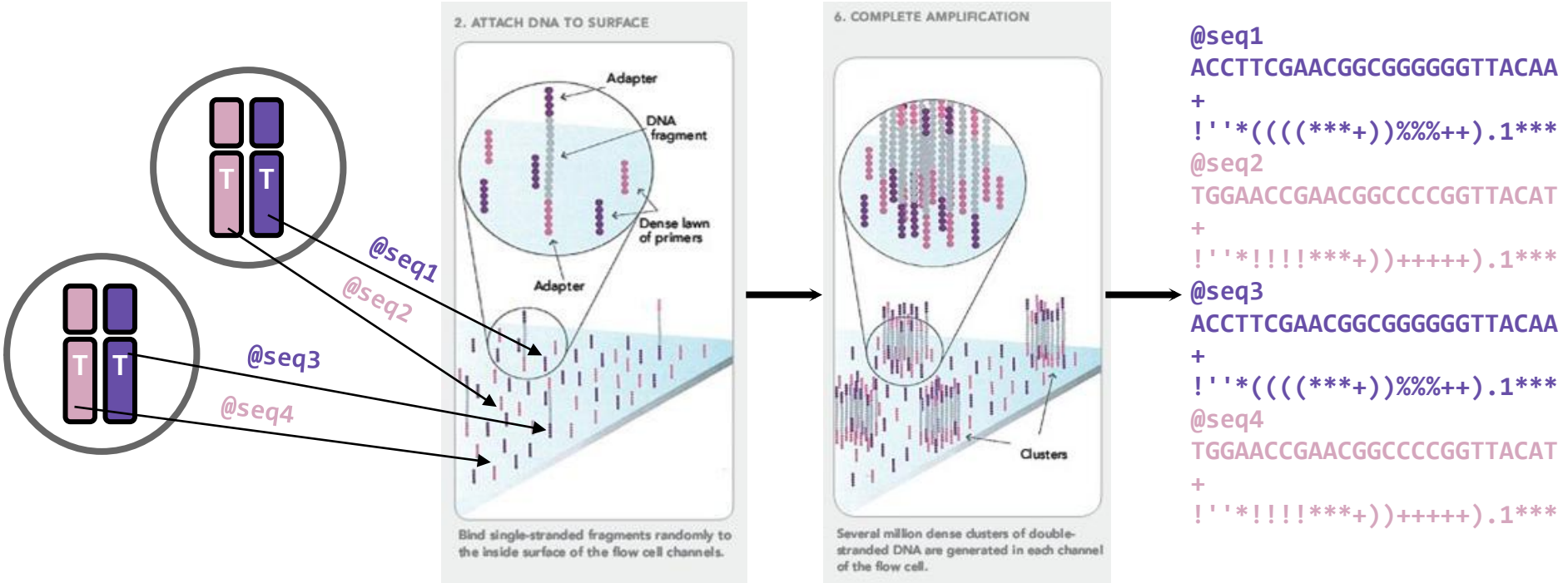
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



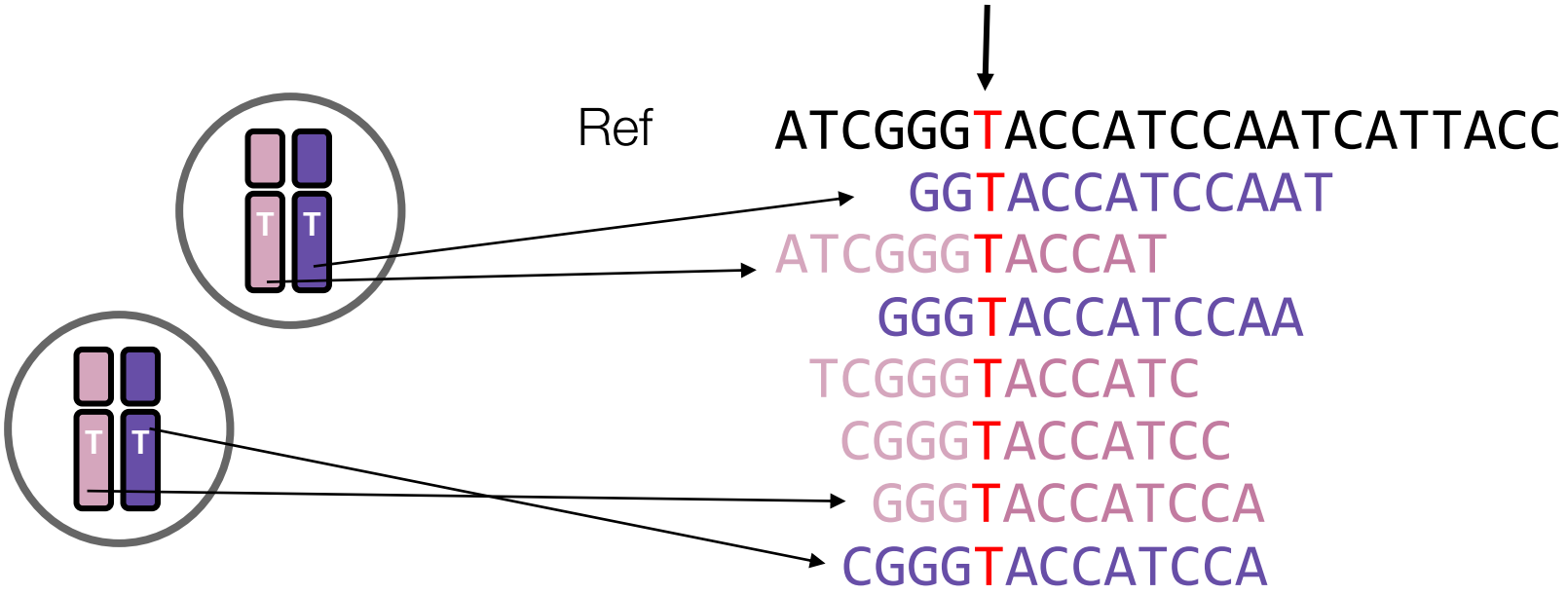
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



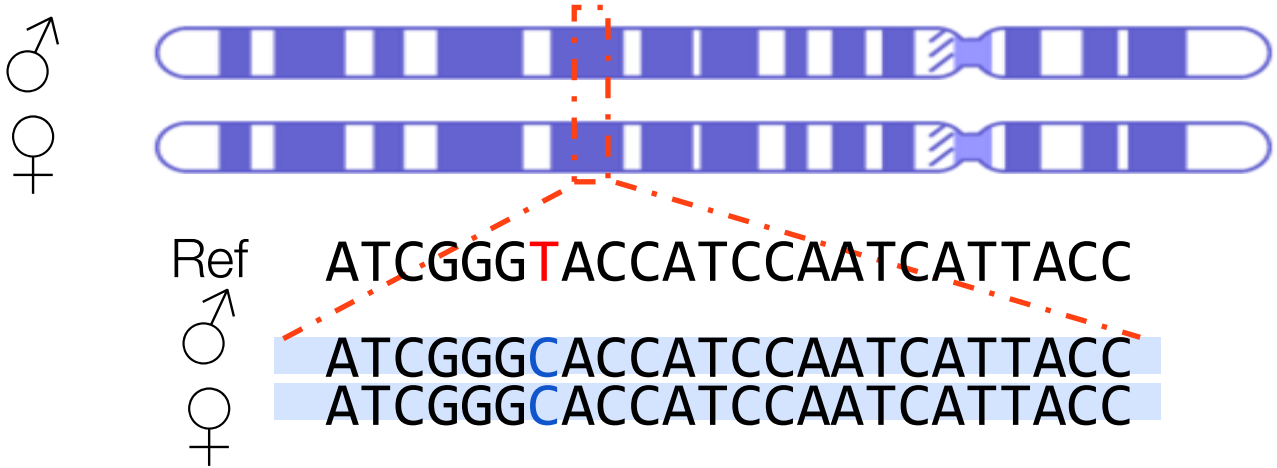
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



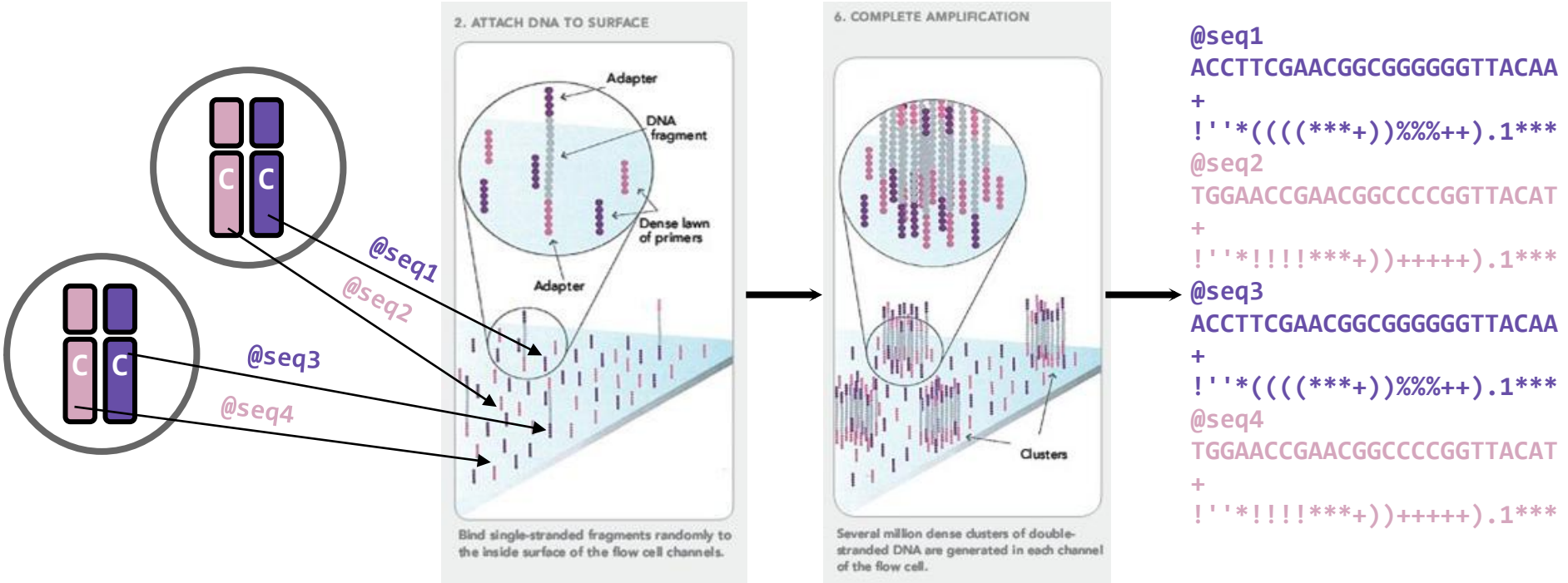
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



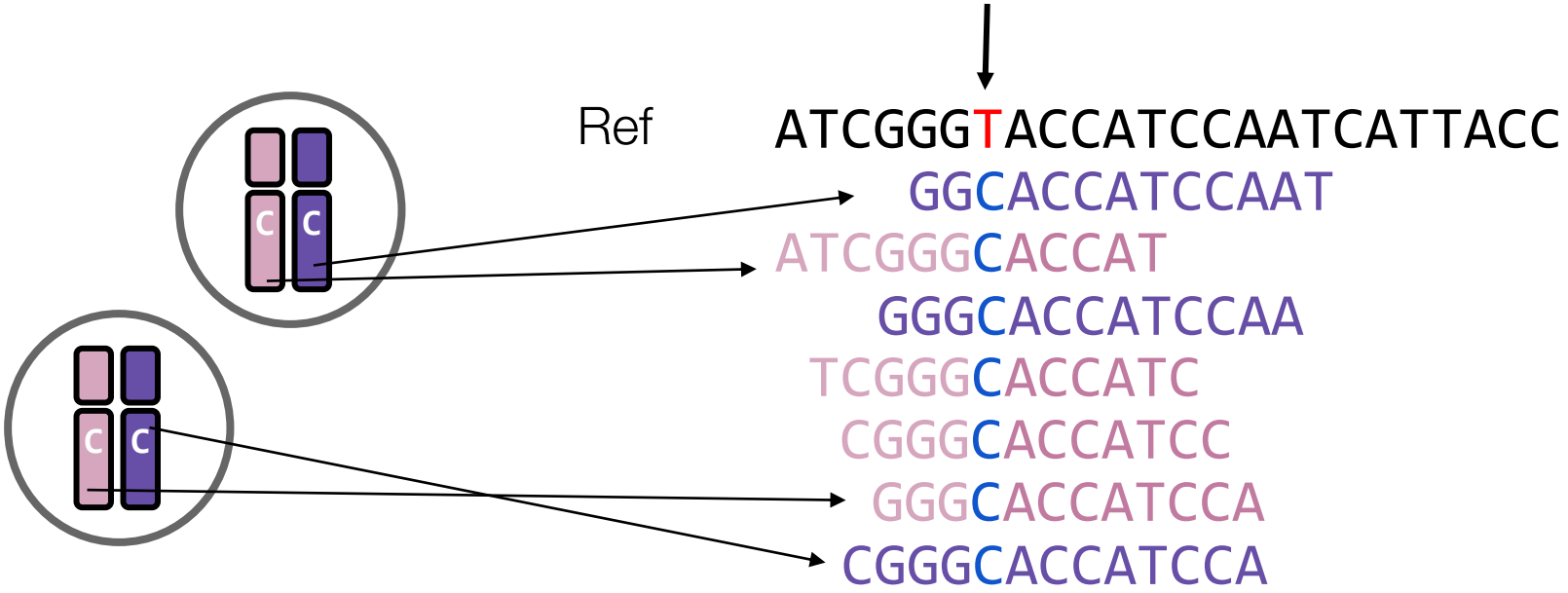
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



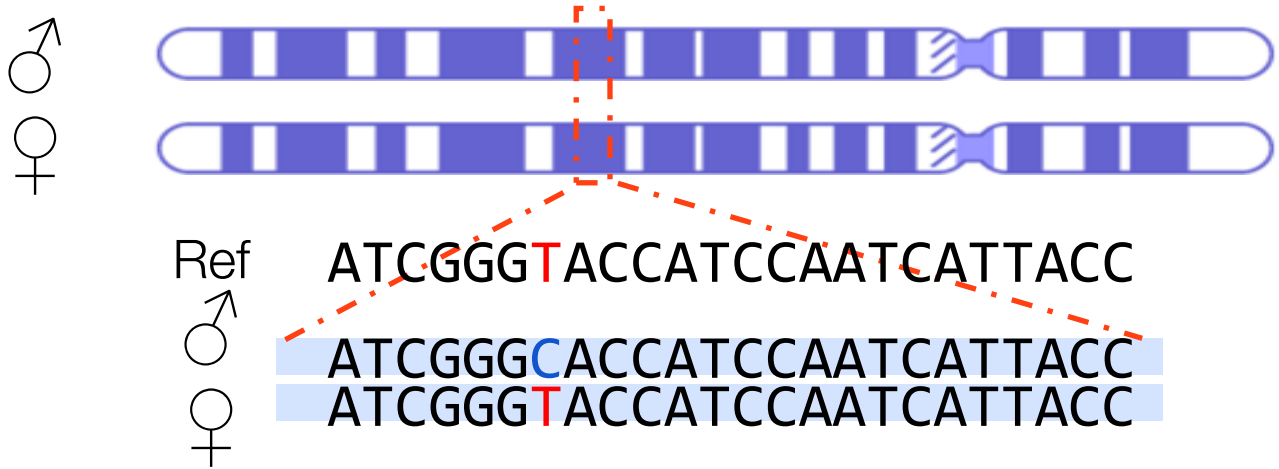
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



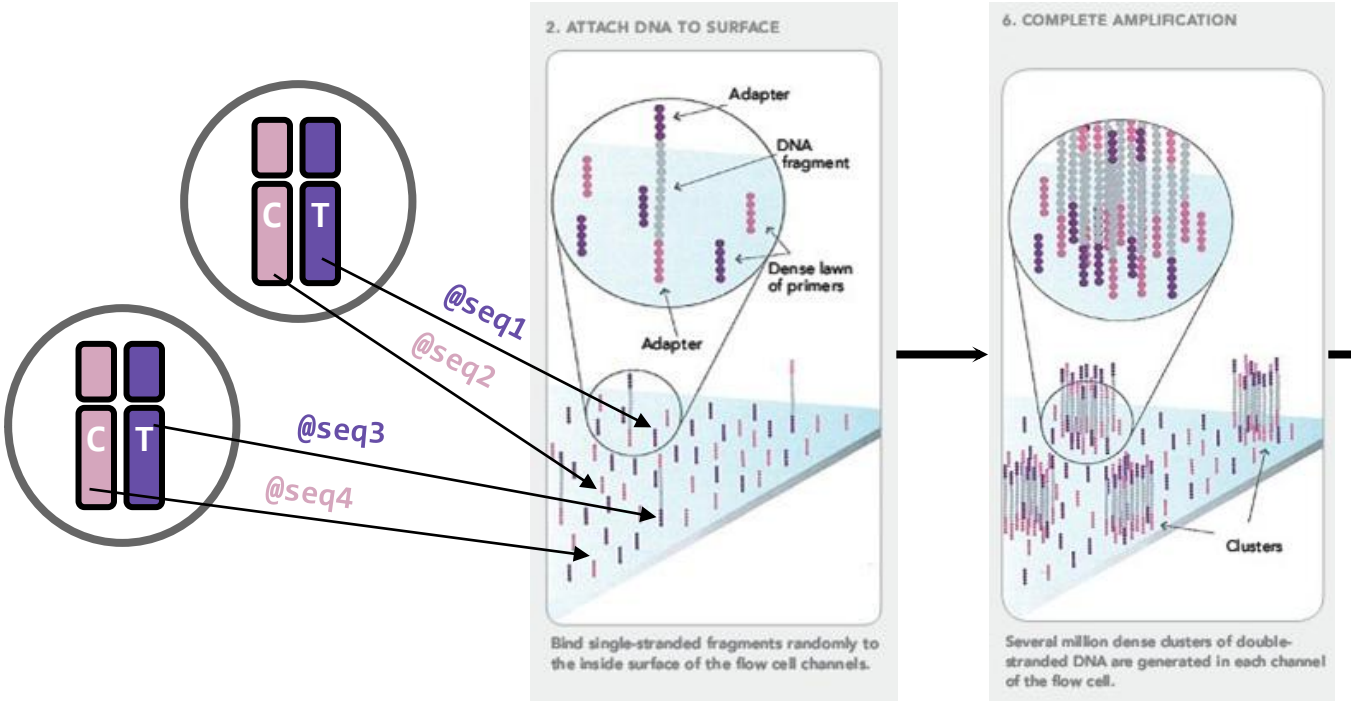
Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

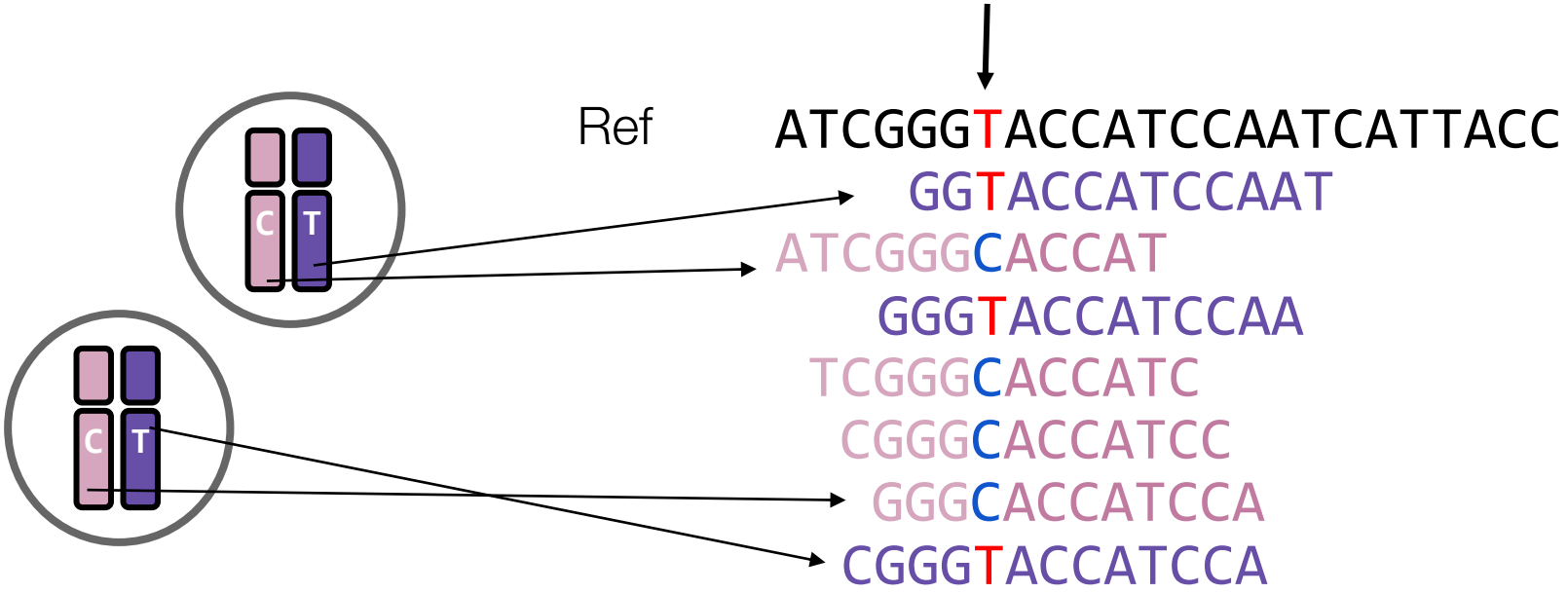
Scenario 3: An individual is heterozygous for an "alternate" allele.




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

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Why might finding heterozygous variants be harder?

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

The binomial distribution: adventures in coin flipping

Sequencing read's position ?

T

P(heads) = 0.5

C

P(tails) = 0.5

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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:

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

Notes:

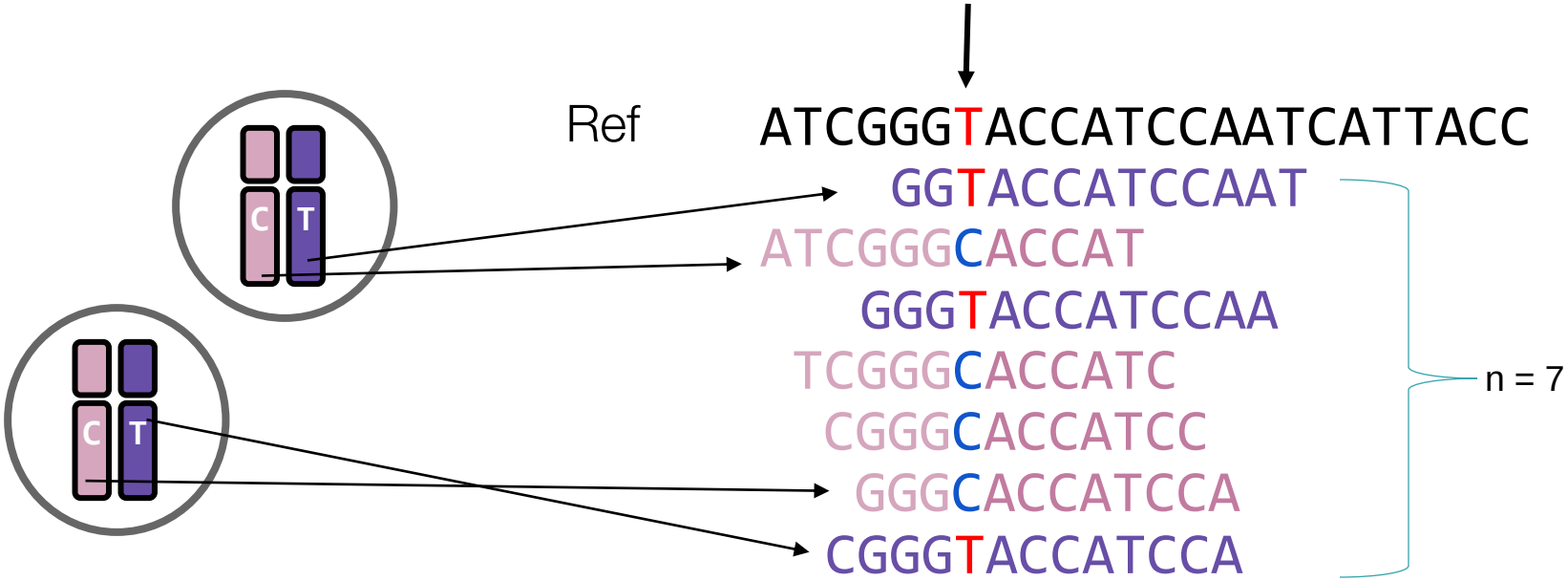
n: number of reads that cover a particular heterozygous position

If **X** is a binomial random variable for modelling the number of seeing **T/C** at this particular position with respect to the whole number of reads that cover this position. Then:

Pr(X=k): resprsents the probability of seeing exactly **k'Cs** or **k'Ts** depending on the defining random variable **X**.

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



n = 7

Pr [X =2], where X represents C's

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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:

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

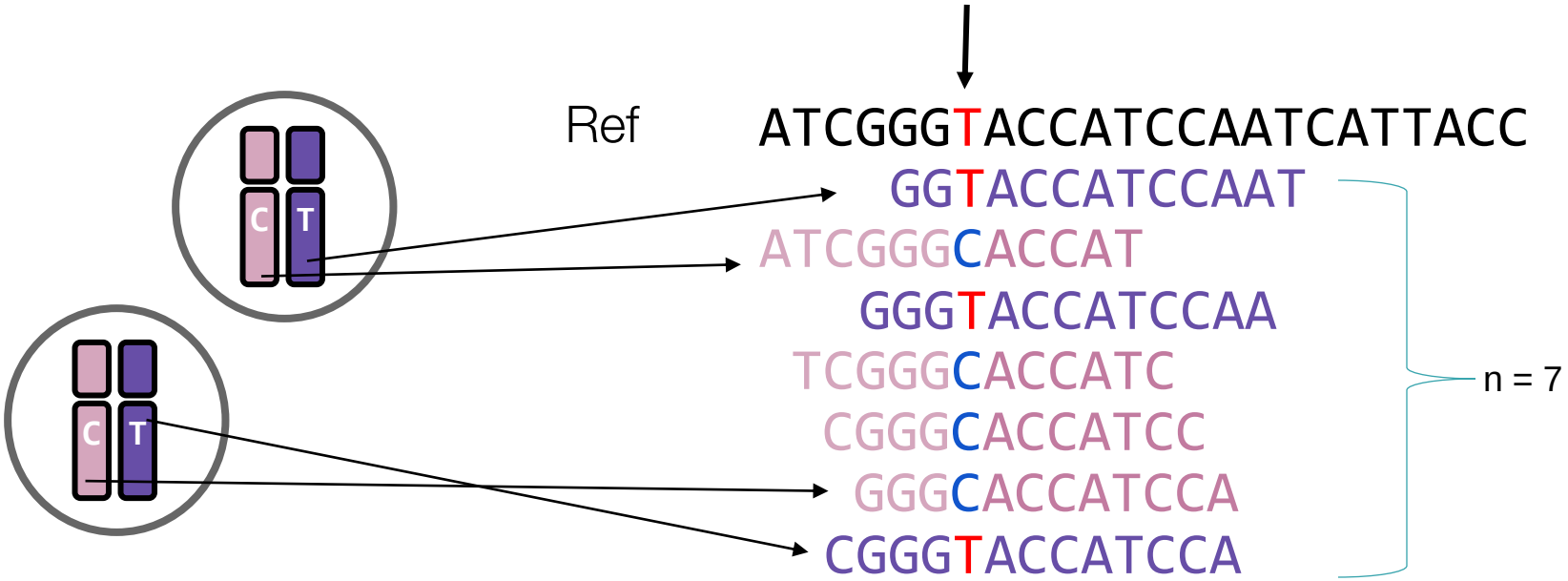
What is the probability of seeing k=1 tails in n=3 flips of a fair coin with the probability of a tail (p) = 0.5?

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)`

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



n = 7

Pr [X =2], where X represents C's, p= probability of success 0.57

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

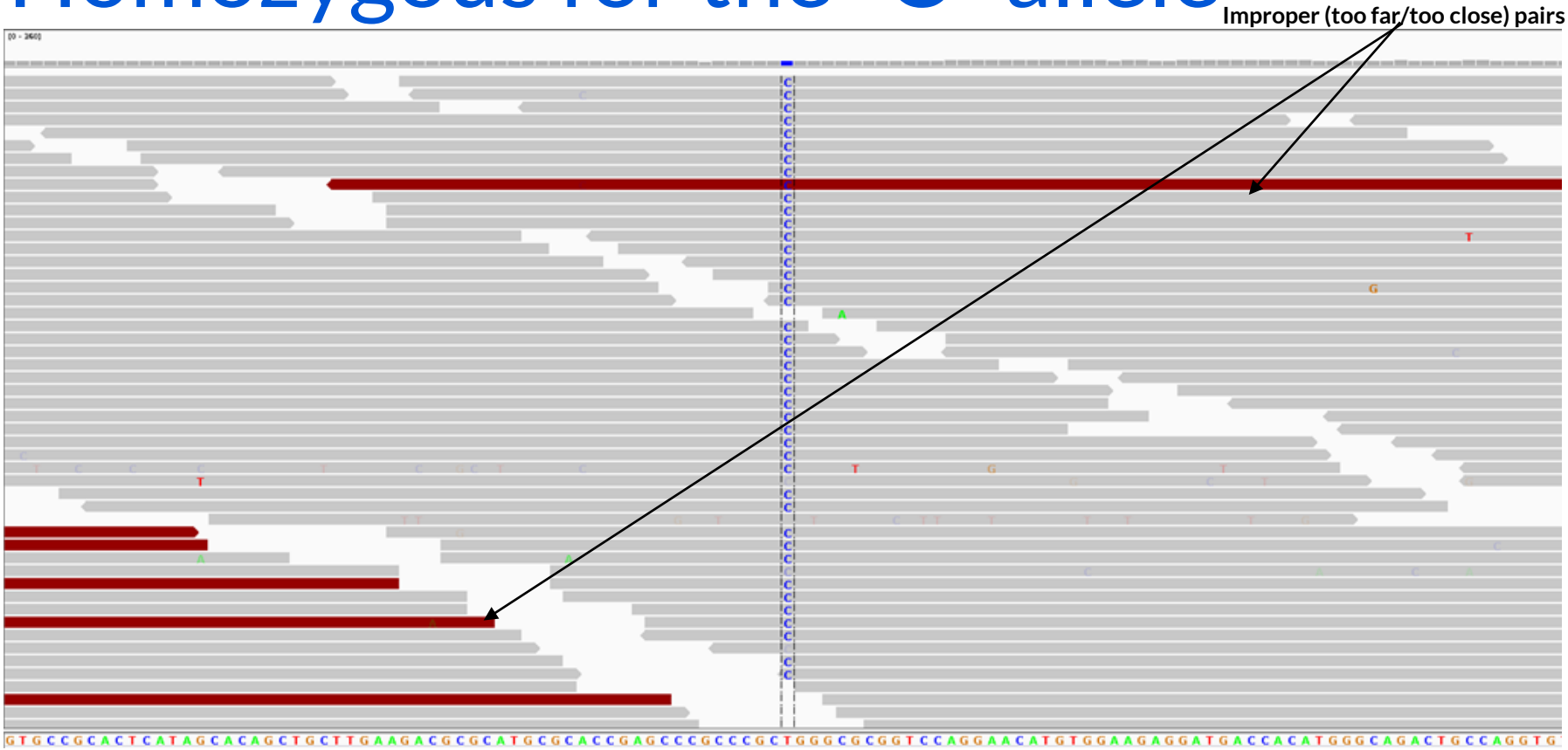
This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to find the majority of heterozygous alleles

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Some real examples of SNPs in IGV

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Homozygous for the "C" allele

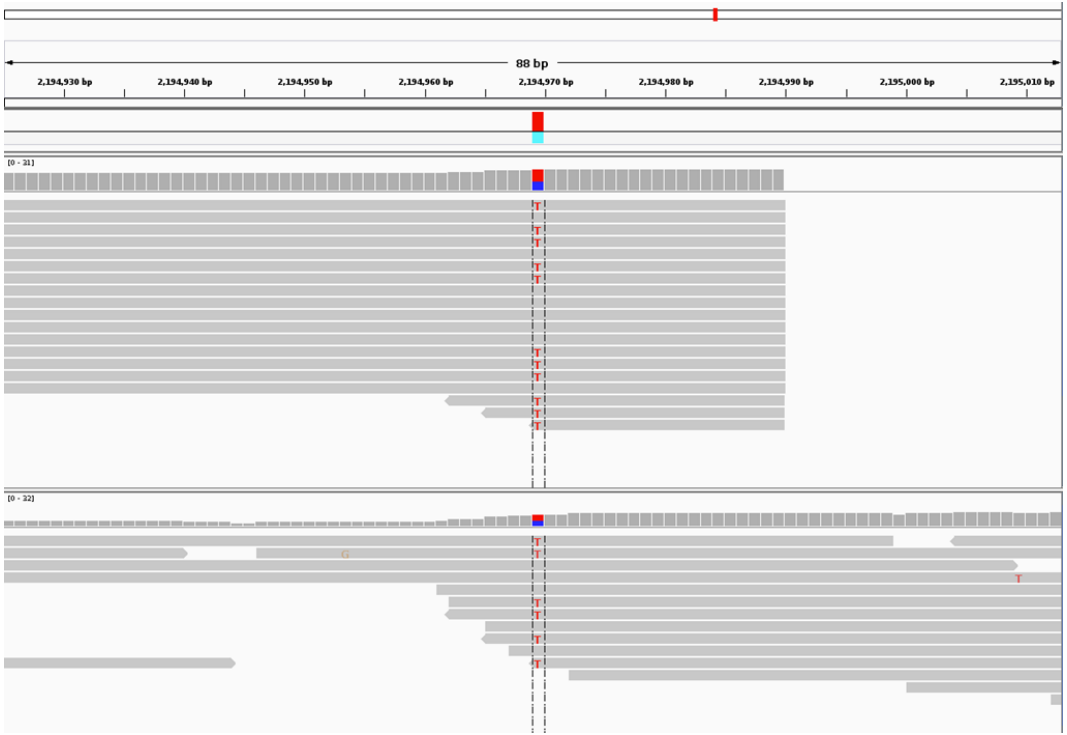


Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Heterozygous for the alternate allele

Individual 1

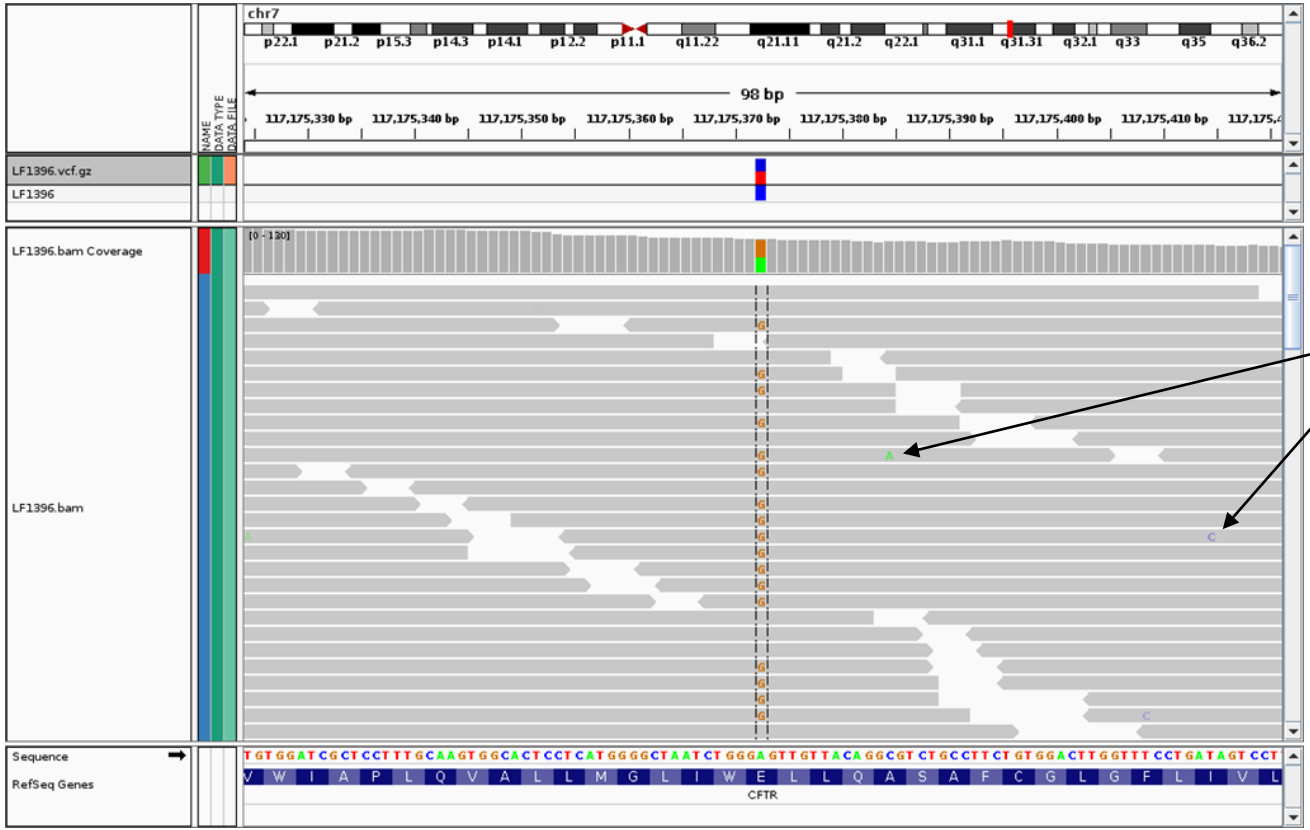
Individual 2



Which genotype prediction would you have more confidence in?

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

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



Sequencing errors

Error rate in this window= total number of errors / total numbers of bases
= 2 / (50 *98)=0.00041

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

Random versus systematic error

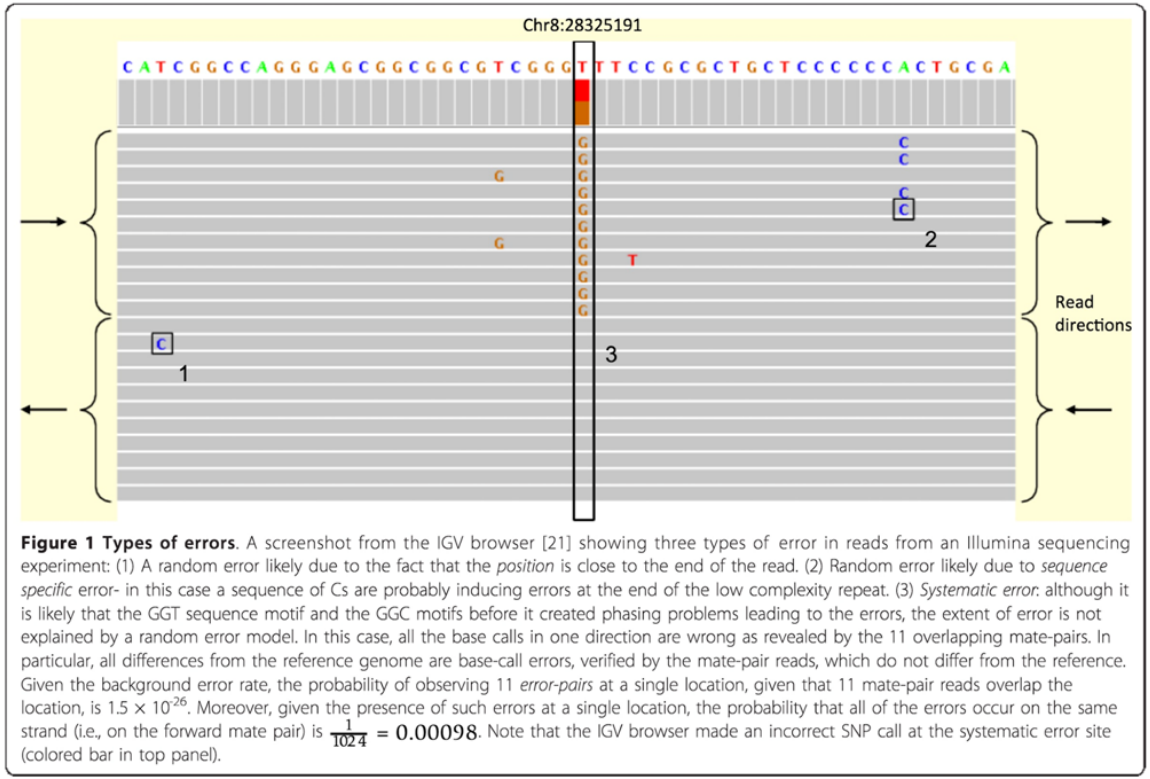
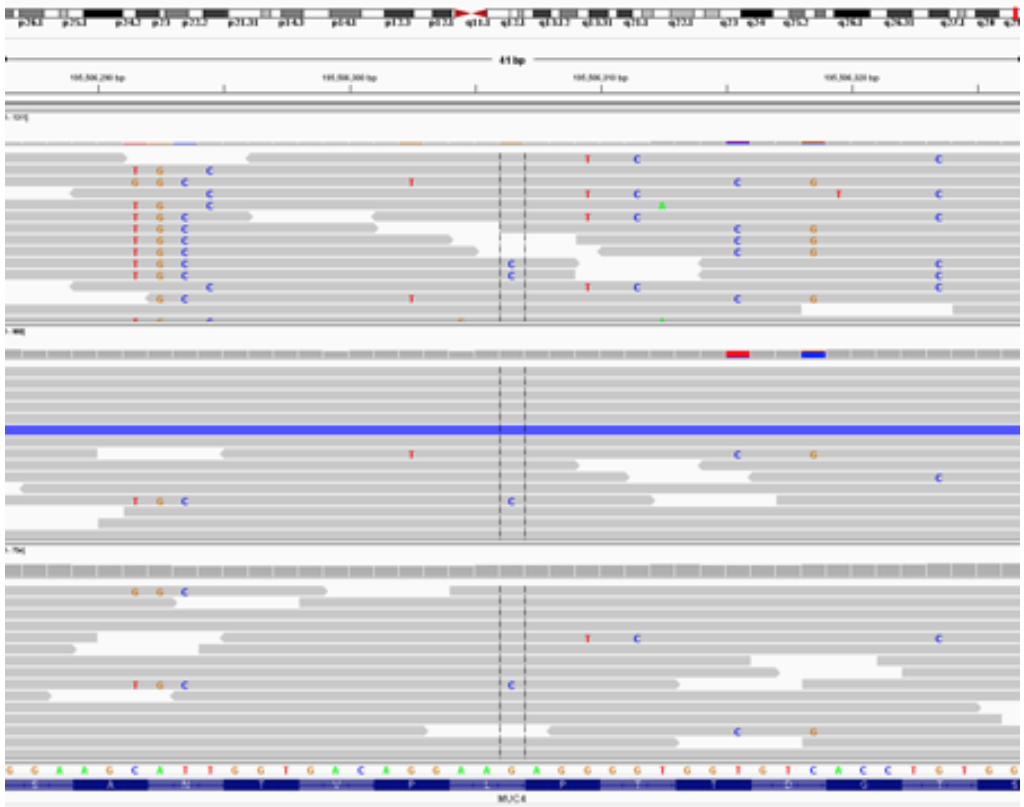


Figure 1 Types of errors. A screenshot from the IGV browser [21] showing three types of error in reads from an Illumina sequencing experiment: (1) A random error likely due to the fact that the *position* is close to the end of the read. (2) Random error likely due to *sequence specific* error- in this 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).

https://www.researchgate.net/publication/51814240 Identification and correction of systematic error in high-throughput sequence data

Pileups of many differences from paralogy



http://massgenomics.org/2013/06/ngs-false-positives.html

Calling INDELs is _much_ harder than SNPs