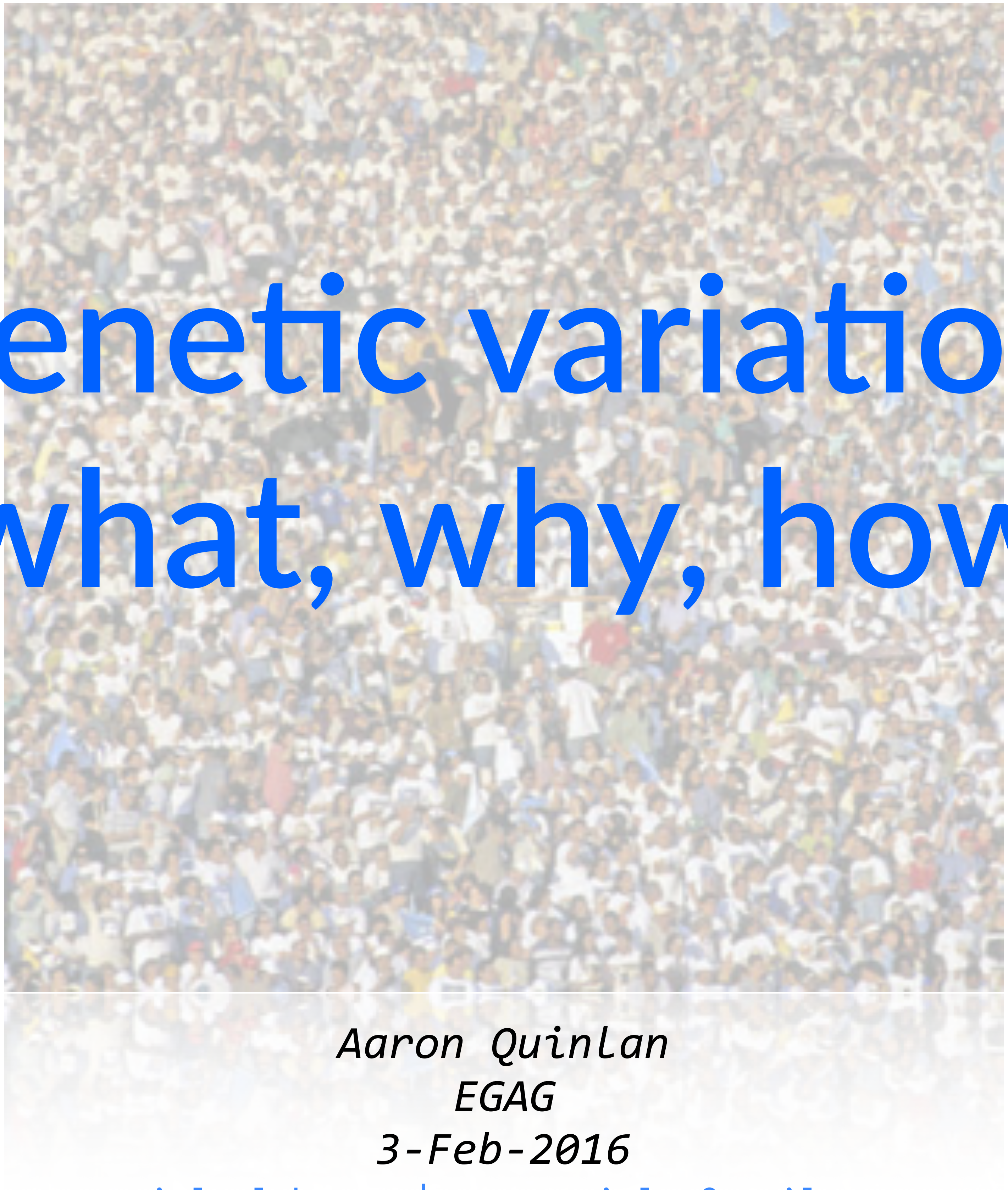


Genetic variation: what, why, how

Aaron Quinlan
EGAG
3-Feb-2016
quinlanlab.org | aaronquinlan@gmail.com

What is genetic variation?

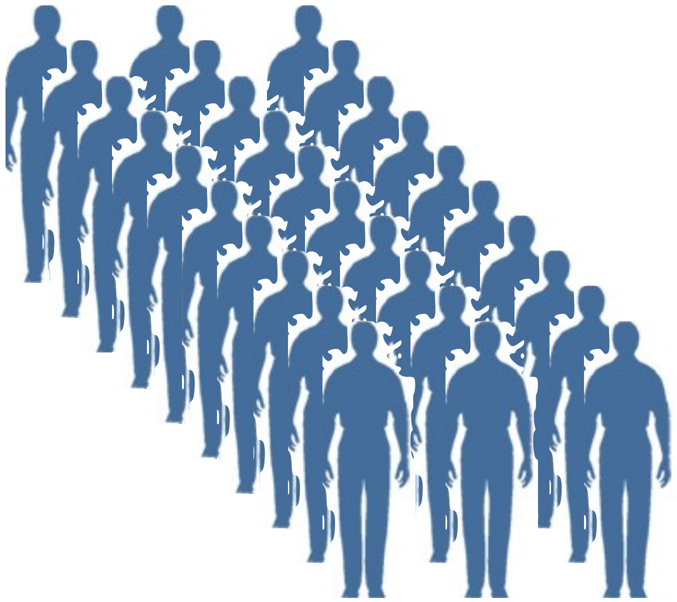
- Differences in DNA content or structure among individuals
- Any two individuals have ~99.5% identical DNA.
- But the human genome is big - each haploid set of 23 chromosomes has 3 billion nucleotides.
- The details matter.

~98-99% identical DNA

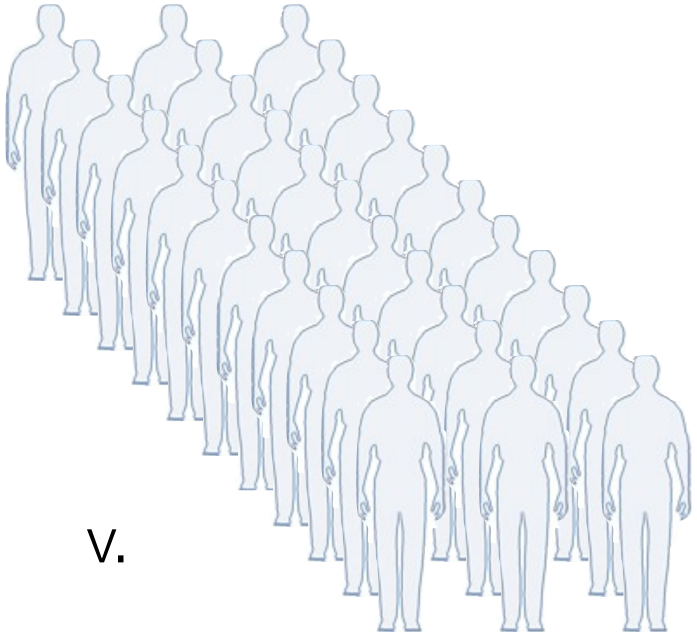
~99.5% identical DNA

Why do we care?

- Understand the relationship between genotype and phenotype.



v.



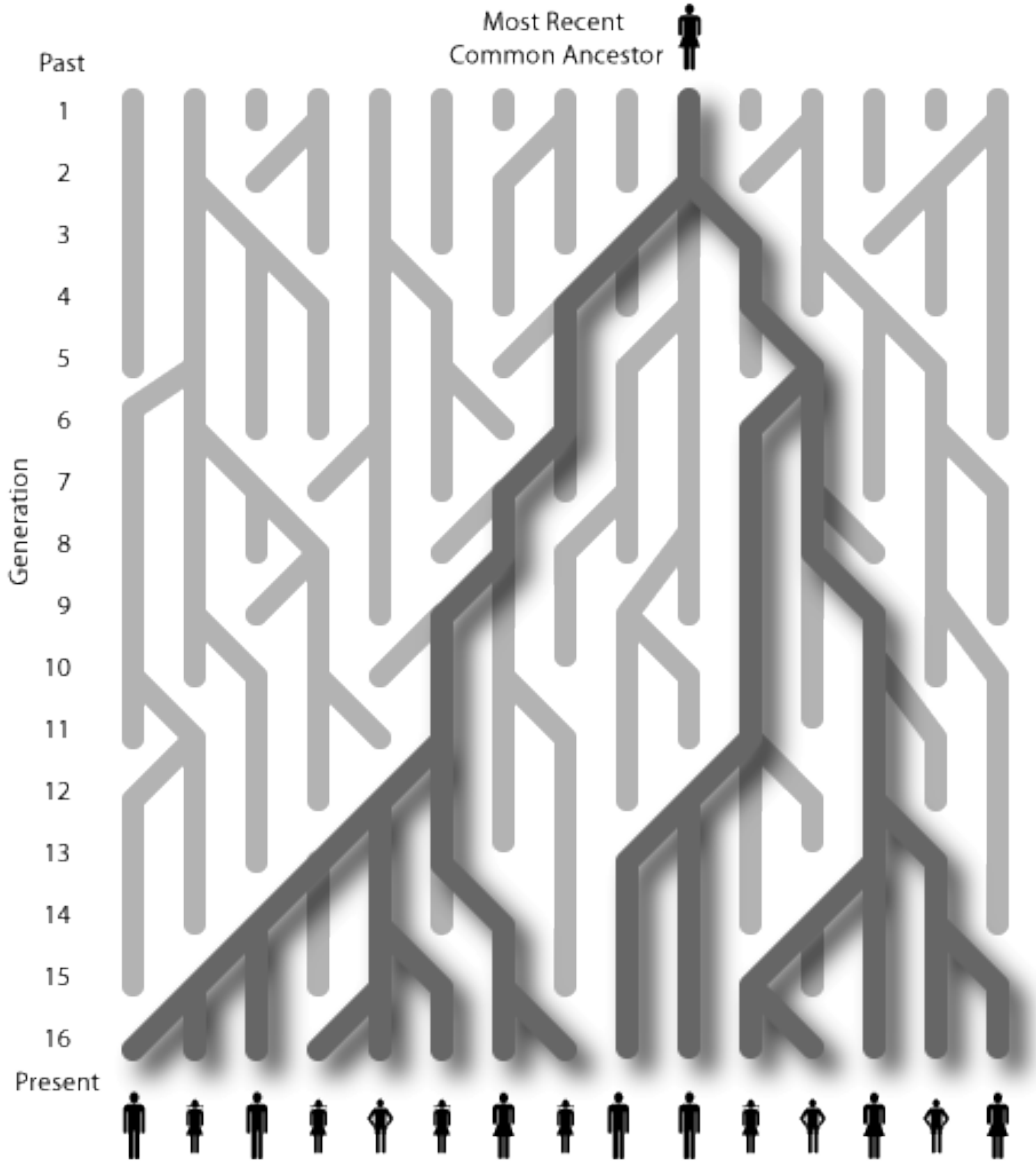
Cases
(have disease)

Controls
(no disease)

Complex diseases
(multiple genes contribute to risk)

Why do we care?

- Bread crumbs of evolution



http://www.fos.auckland.ac.nz/~howardross/RecreateTheResearch/RTR/images/Coalescence.png

Why do we care?

- How, when, where does our genome evolve?

Types of genetic variation

ctc**c**gag
ctc**t**gag

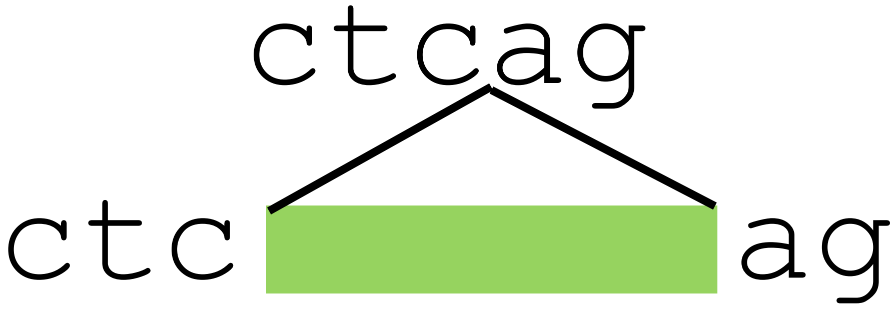
Single-nucleotide polymorphisms (**SNPs**)

“spelling mistakes”

ctc--ag
ctc**tg**ag

Insertion-deletion polymorphisms (**INDELs**)

“extra or missing letters”



ctcag
ctc ag

Structural variants (**SVs**)

“extra, missing or reordered chapters”

Properties of genetic variation

	Single-nucleotide (**SNPs**)	Insertion-deletions (**INDELs**)	Structural variants (**SVs**)
	ctc**c**gag ctc**t**gag	ctc**--**ag ctc**tg**ag	ctcag ctc 🟩 ag
Size	1bp	1-100bp	100bp-1Mb+
Frequency	3 million / genome	300K / genome	3,000 / genome
Detection Difficulty	Easy	Medium	Hard

How different are we?

How different are we?

The genomes of any two humans are ~99.5% identical

How different are we?

The genomes of any two humans are ~99.5% identical

But, the genome is big.

How different are we?

The genomes of any two humans are ~99.5% identical

But, the genome is big.

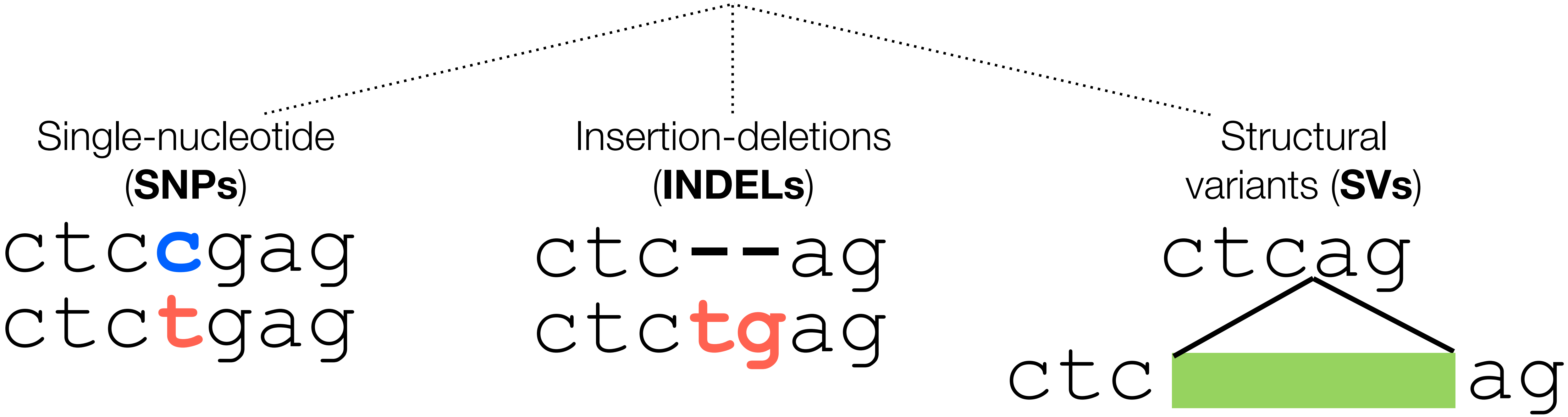
15,000,000 to 21,000,000 different base pairs.

How different are we?

The genomes of any two humans are ~99.5% identical

But, the genome is big.

15,000,000 to 21,000,000 different base pairs.



Single-nucleotide (**SNPs**)

ctc**c**gag
ctc**t**gag

Insertion-deletions (**INDELs**)

ctc--ag
ctc**tg**ag

Structural variants (**SVs**)

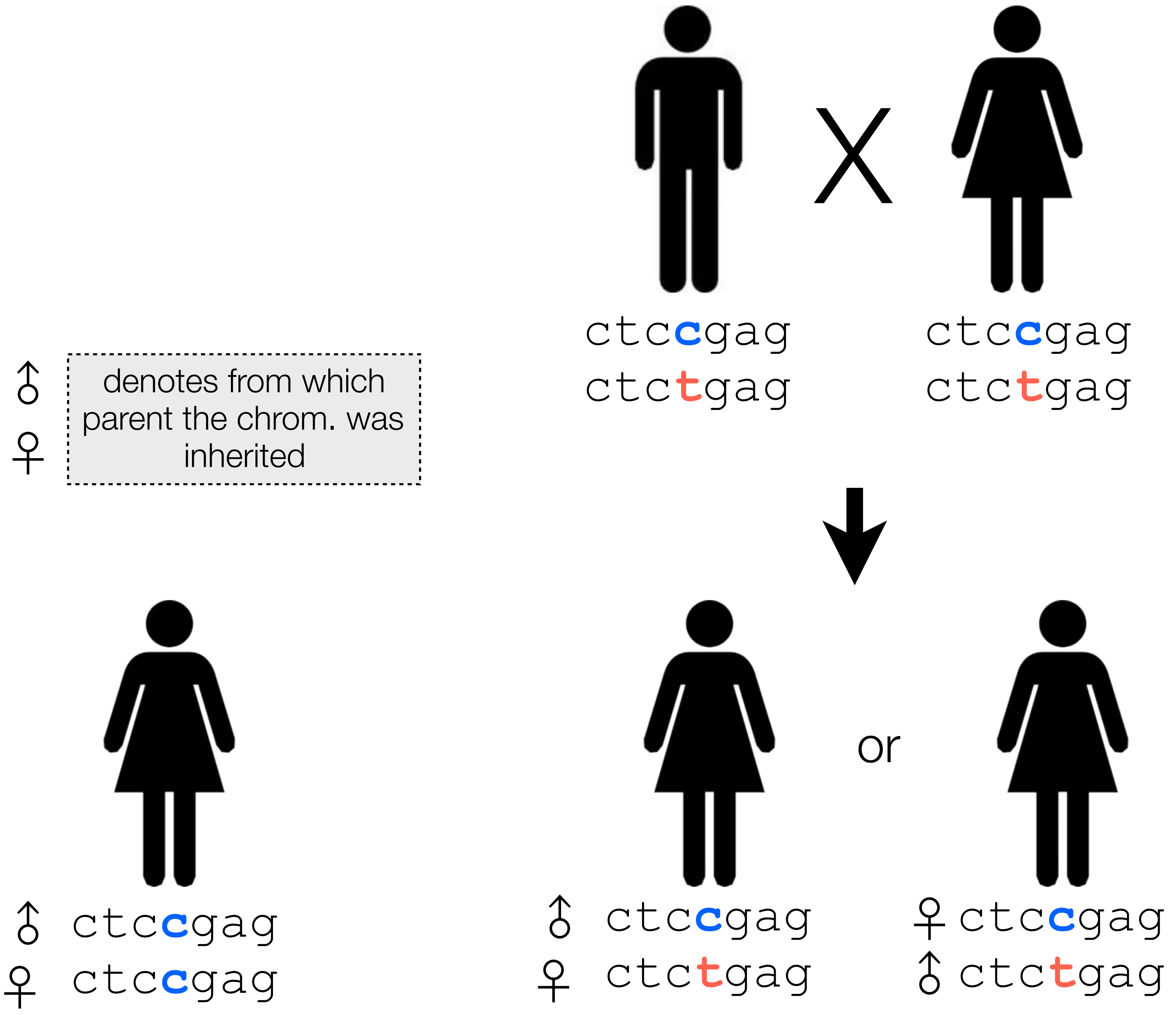
ctcag
ctc 🟩 ag

Detecting genetic variation

```
CGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTC
GTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAGCCAGACTTCCCGTGTCCTTTCC
ACCGGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGG
GTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTG
TTGTCTGCATGTAACTTAATACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCT
AGGCCAGTAAGTAGTGCTTGTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCT
GCTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGG
AGAGCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTGGCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCT
CTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGCCGCAGGCACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGAT
GTGTGGGCCCACCGGCCCCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCAGGTCTCCAGAG
CTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGA
GTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCC
TGTGATGTCATCTGGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAAGGCCTGGCAGAGTCTTTCCCAGGGA
AAGCTACAAGCAGCAAACAGTCTGCATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCAAGAAAGGCTCTGG
TGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAGG
GGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGAGGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTA
ACTCAAAGCCTTAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACCAGTGAT
ACACCCGGCACCCTGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGC
TGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCCAGT
TTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTG
TTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCA
CAGGCATTAGTGCCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGGGTCCTGGCCATCCGTGAGATCTTCCCAGGTGTGCCGT
TTTCTCTGGAAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCCATGGAGCACAGGCAGACAGAAGTCCCCGCCC
CAGCTGTGTGGCCTCAAGCCAGCCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGCTGGTTCCGTCACCCCCTCCCAAGGAAGTAGG
TCTGAGCAGCTTGTCCTGGCTGTGTCCATGTCAGAGCAACGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAGCCCCCTACGA
TTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCCT
CCGGGCCCCTCACCAGCCCCAGGTCCTTTCCCAGAGATGCCTGGAGGGAAAAGGCTGAGTGAGGGTGGTTGGTGGGAAACCCTGGTTC
CCCCAGCCCCCGGAGACTTAAATACAGGAAGAAAAAGGCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCTGCCTC
CTACCCTTGCGCCTCATGACCGGAGCCATAGCCCAGGCAGGAGGGCTGAGGACCTCTGGTGGCGGCCCAGGGCTTCCAGCATGTGCCC
TAGGGGAAGCAGGGGCCAGCTGGCAAGAGCAGGGGGTGGGCAGAAAGCACCCGGTGGACTCAGGGCTGGAGGGGAGGAGGCGATCTTG
CCCAAGGCCCTCCGACTGCAAGCTCCAGGGCCCGCTCACCTTGCTCCTGCTCCTTCTGCTGCTGCTTCTCCAGCTTTCGCTCCTTCAT
GCTGCGCAGCTTGGCCTTGCCGATGCCCCCAGCTTGGCGGATGGACTCTAGCAGAGTGGCCAGCCACCGGAGGGGTCAACCACTTCCC
```

How existing (germline) variation is inherited

***Example**: Mom and dad are heterozygous;*
that is, the zygote
from which they developed
was comprised of a sperm
and egg with that
gave two paternal alleles



♂ ♀ denotes from which parent the chrom. was inherited

ctc**c**gag
ctc**t**gag

X

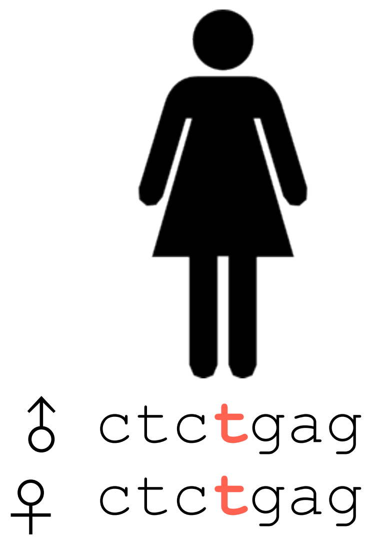
ctc**c**gag
ctc**t**gag

♂ ctc**c**gag
♀ ctc**c**gag

♂ ctc**c**gag
♀ ctc**t**gag

or

♀ ctc**c**gag
♂ ctc**t**gag



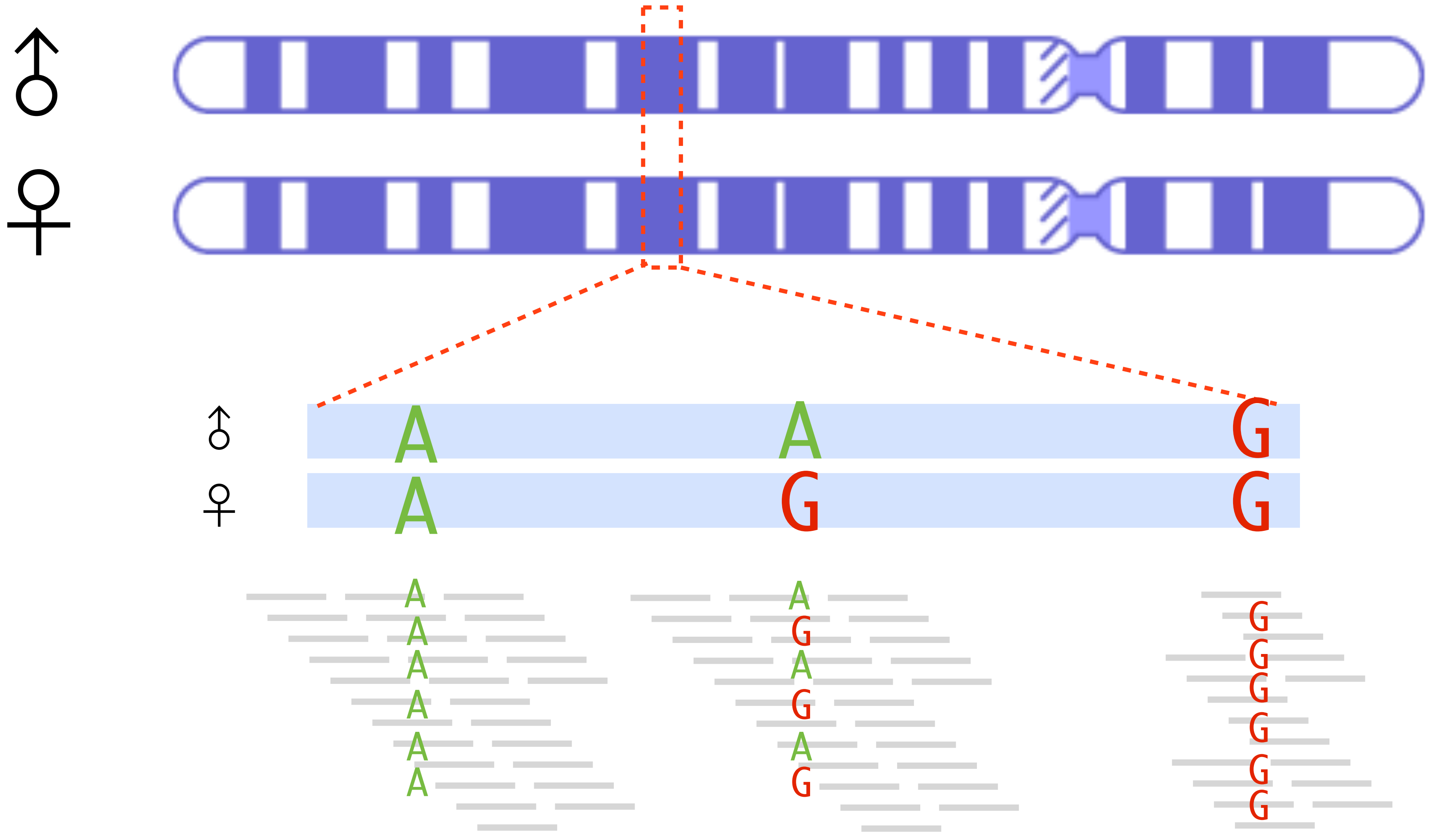
♂ ctc**t**gag
♀ ctc**t**gag

Kid is homozygous
(C/C)

Kid is heterozygous
(C/T)

Kid is homozygous
(T/T)

Recall: **we are diploid.**



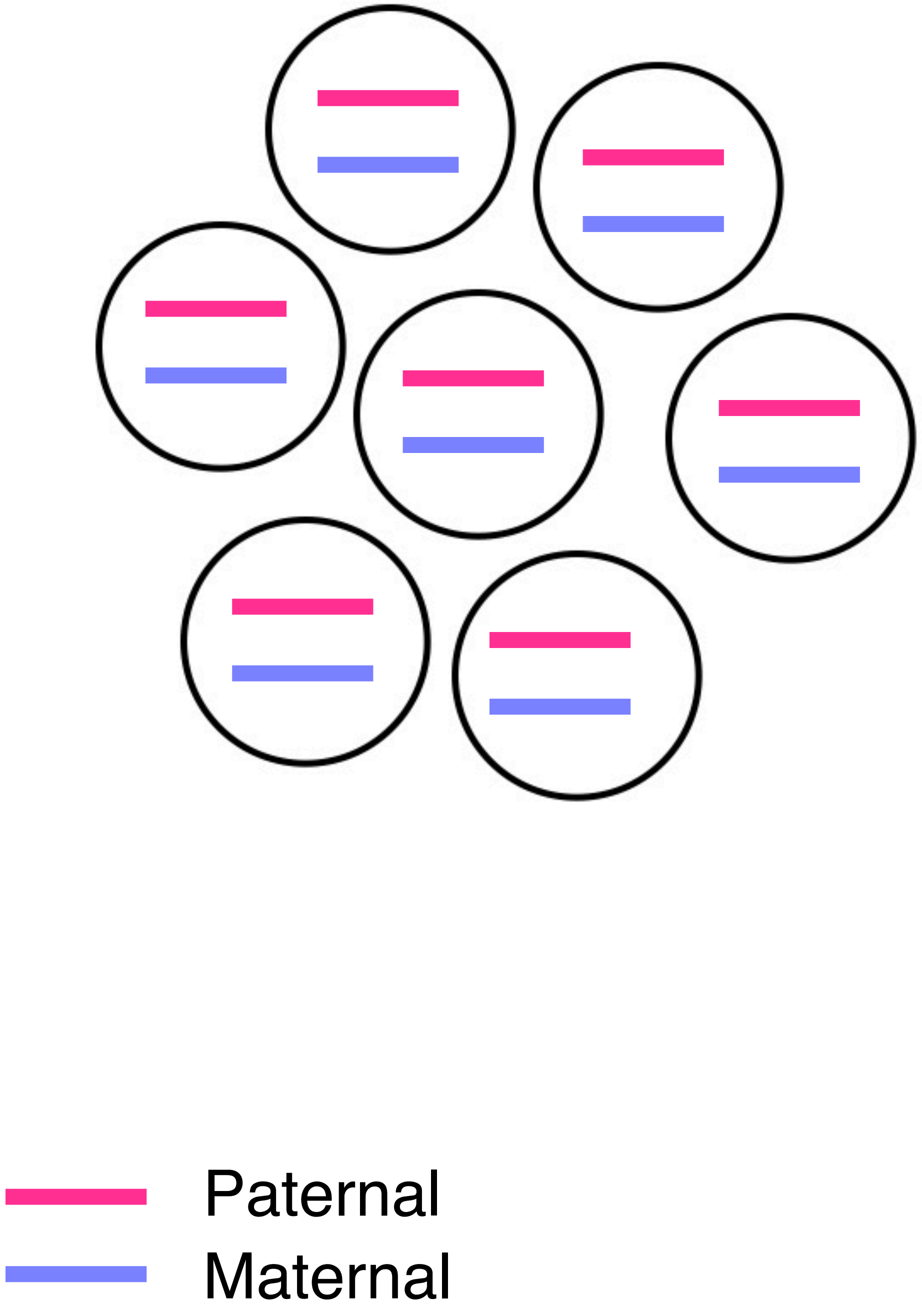
Aligned DNA sequence "reads"

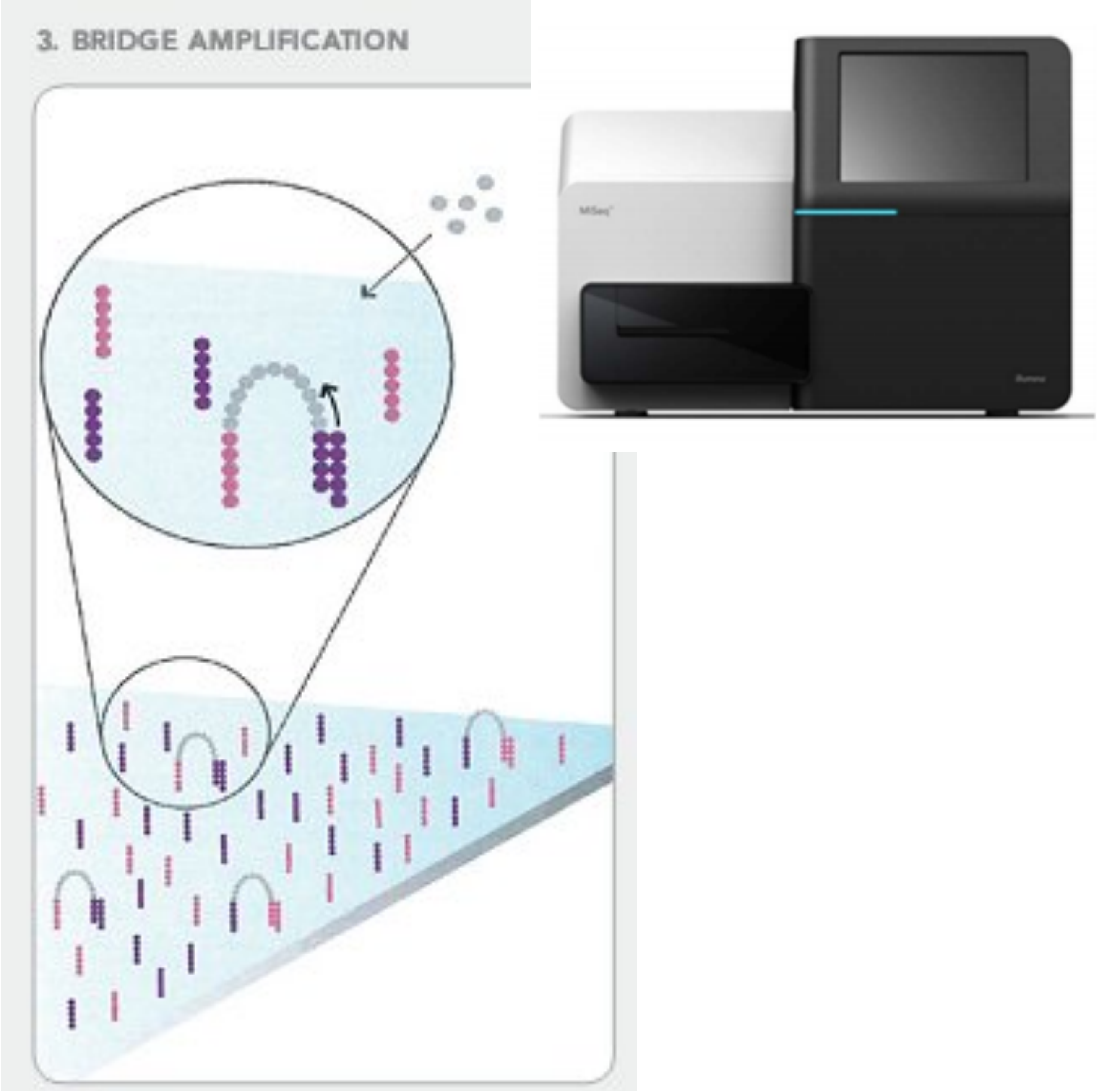
homozygous
A/A

heterozygous
A/G

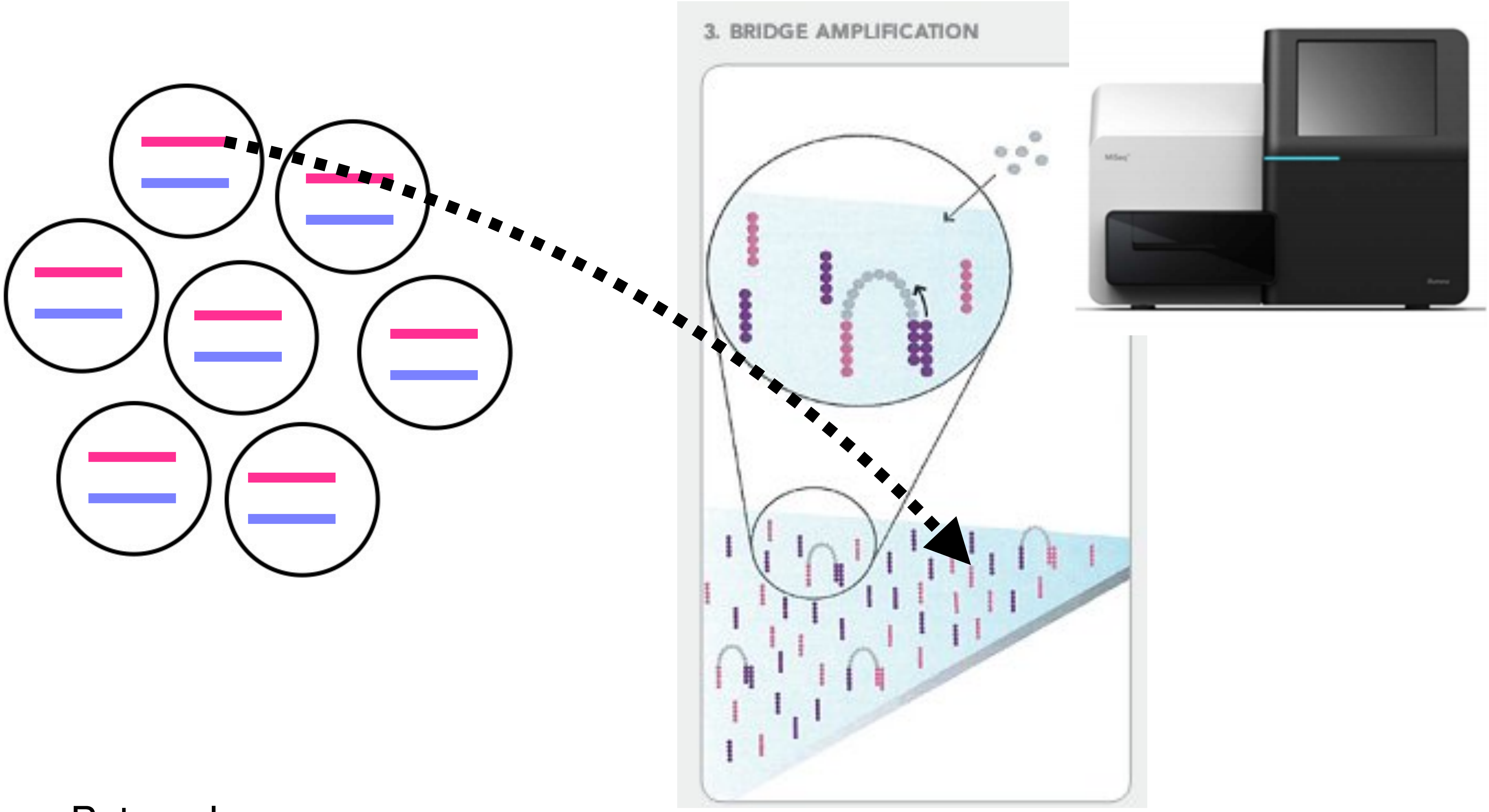
homozygous
G/G

Each sequencing read is a piece of a parental allele from a single cell





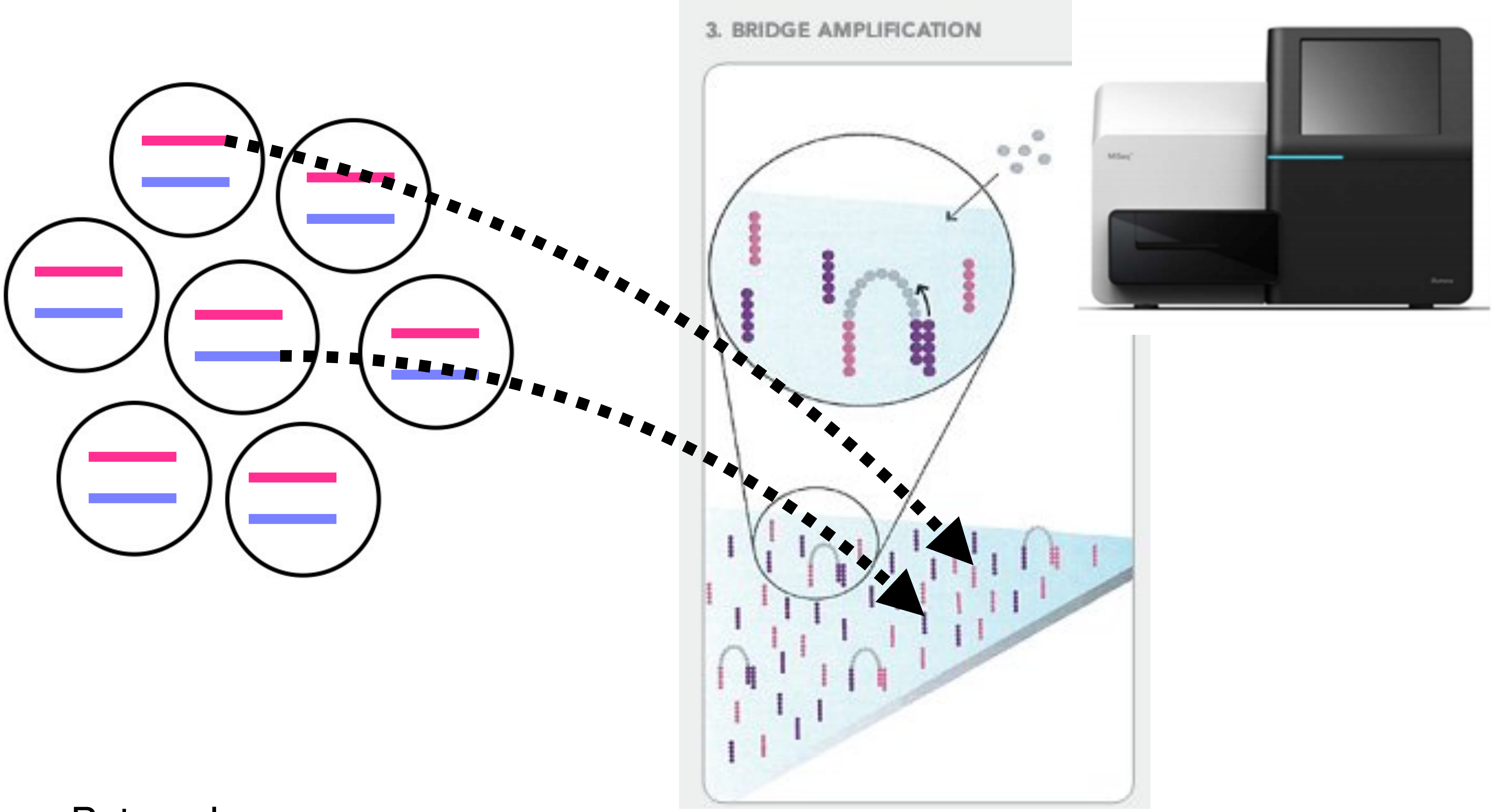
Each sequencing read is a piece of a parental allele from a single cell



▬ Paternal

▬ Maternal

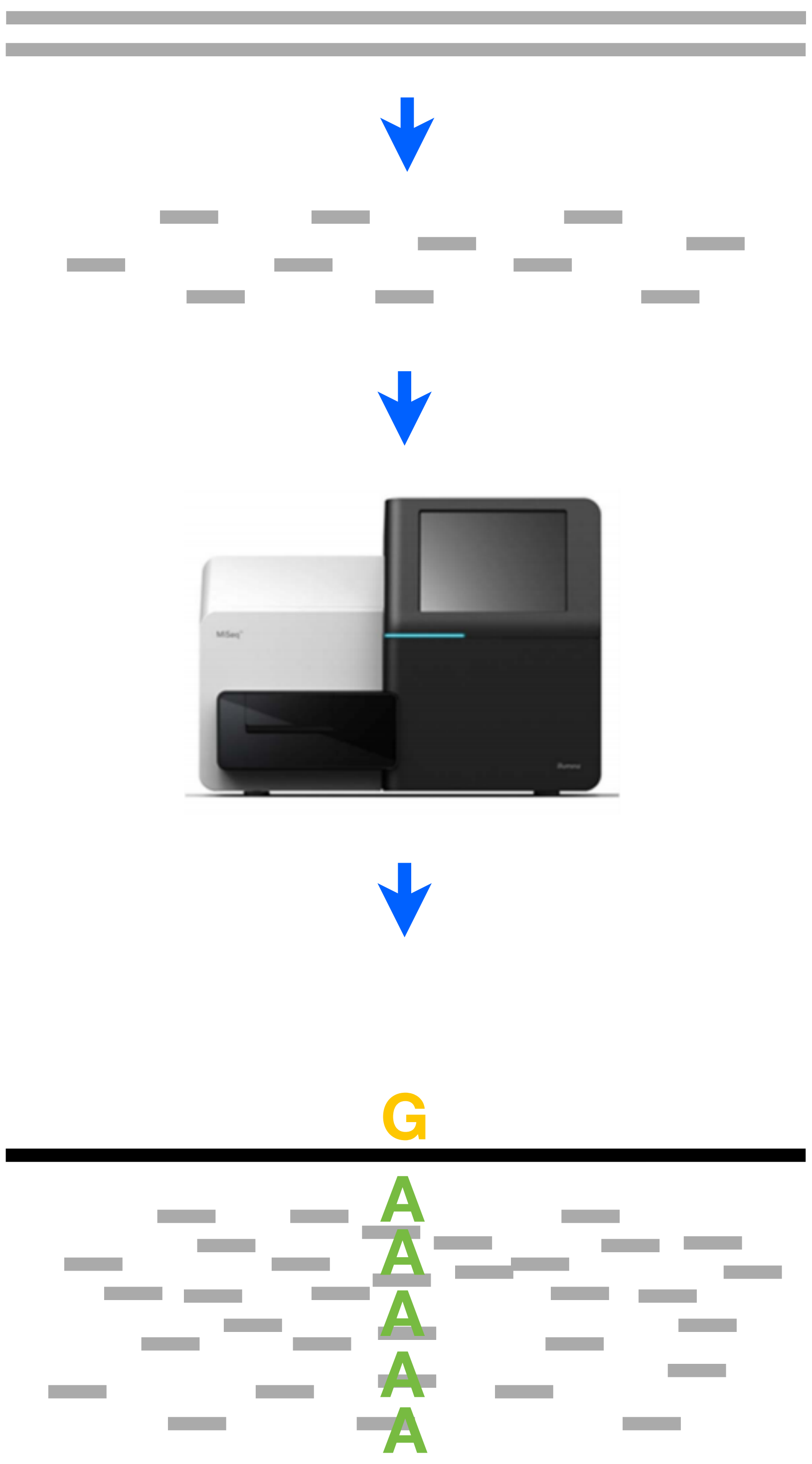
Each sequencing read is a piece of a parental allele from a single cell



🟥 Paternal

🟦 Maternal

A “pileup” of reads at a given chromosomal position is a sampling of the allels present in a population of cells.



Genomic DNA
from millions of cells

Fragmented DNA

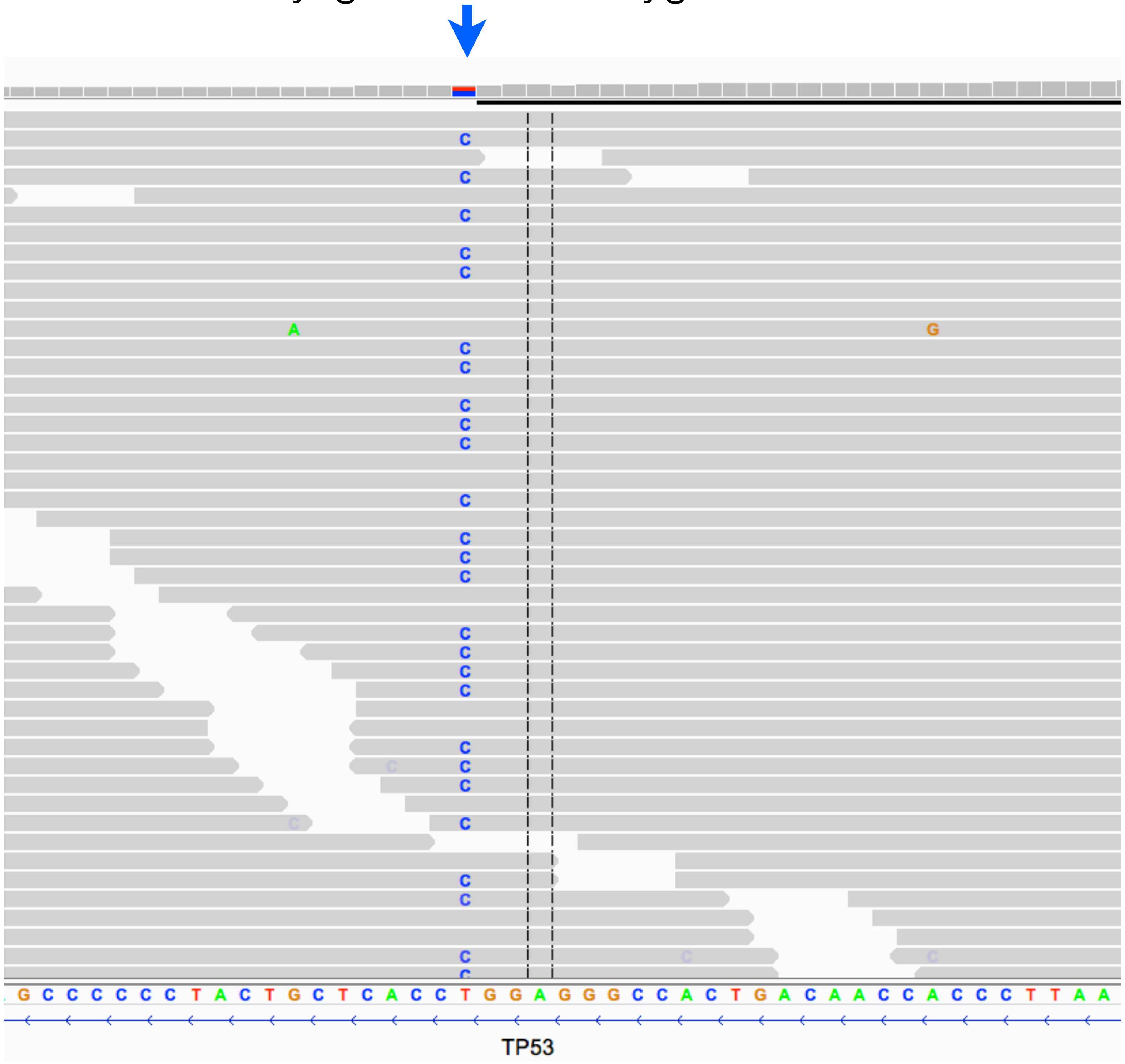
Sequence billions of
DNA fragments
from millions of cells.

Align DNA to a reference genome. Comparing sample DNA to reference reveals genetic differences.

Heterozygotes: expect ~50/50 allele balance (binomial expectation). There are biases…

True variation.
Homoyzgous or heterozygous?

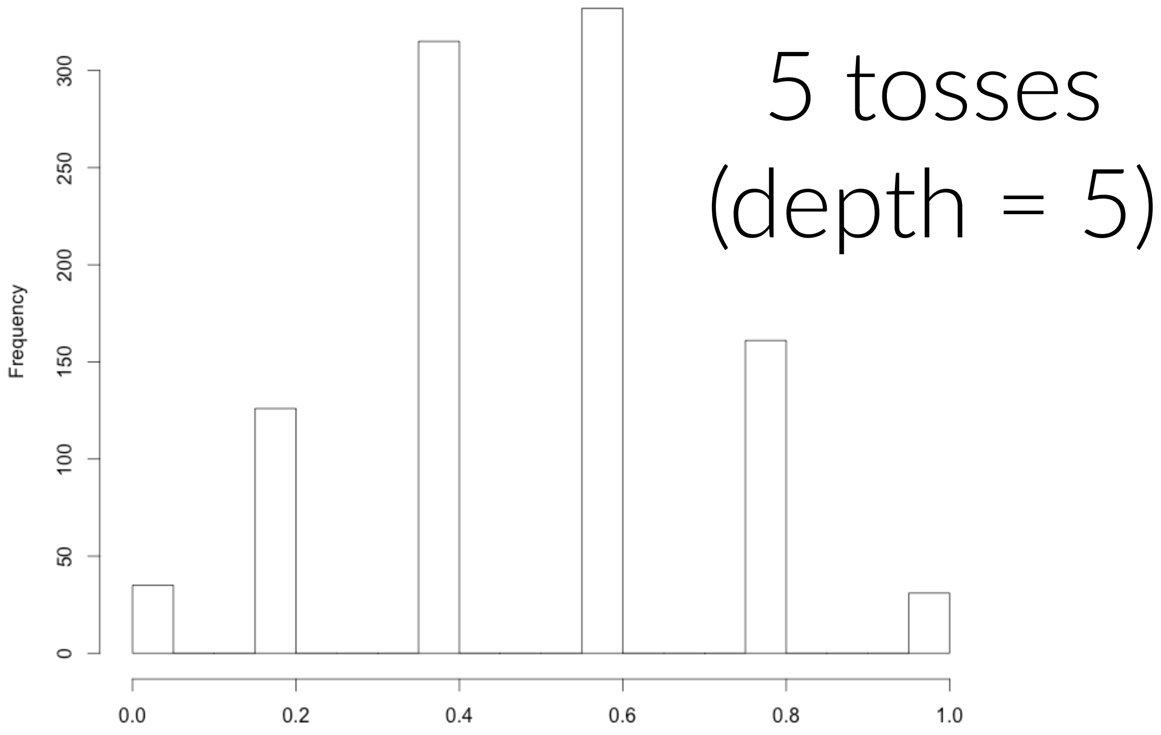
# of T = 19
# of C = 25


A
G
C
G C C C C C C T A C T G C T C A C C T G G A G G G C C A C T G A C A A C C A C C C T T A A
TP53

https://gist.github.com/arq5x/942a18105a0be1352f05

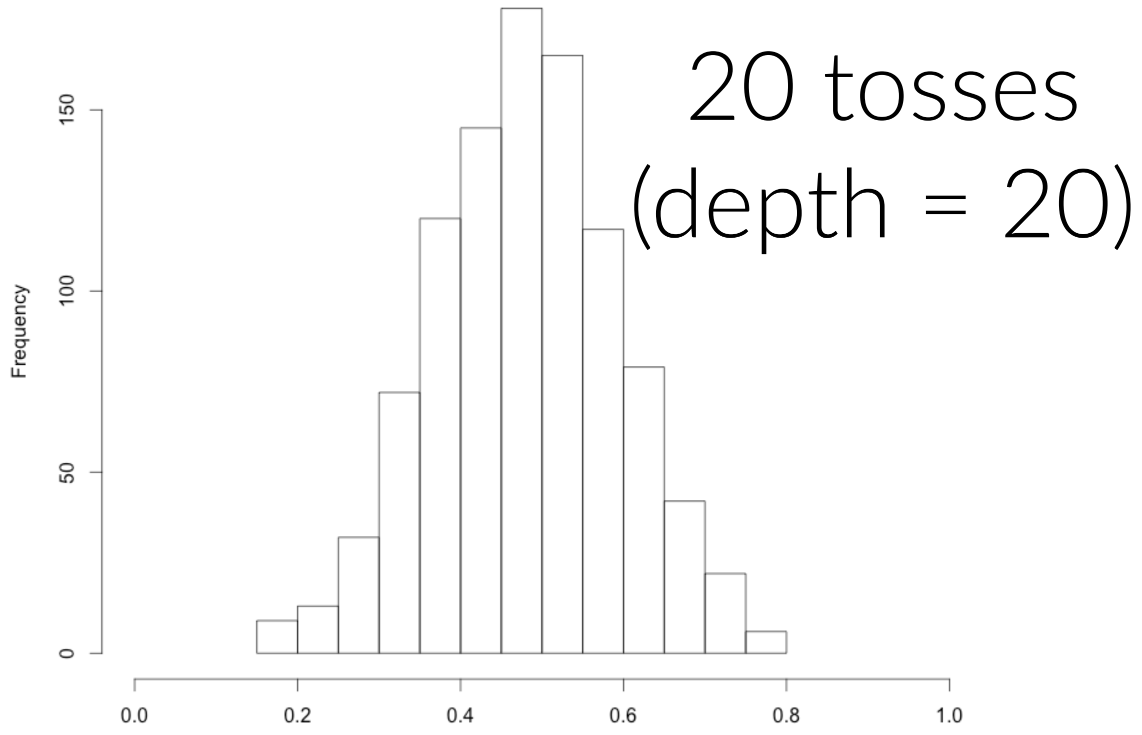
Allele sampling from NGS reads is akin to coin tosses. Deeper sampling "coverage" is better.

Distribution of % heads from 1000 experiments with 5 tosses each



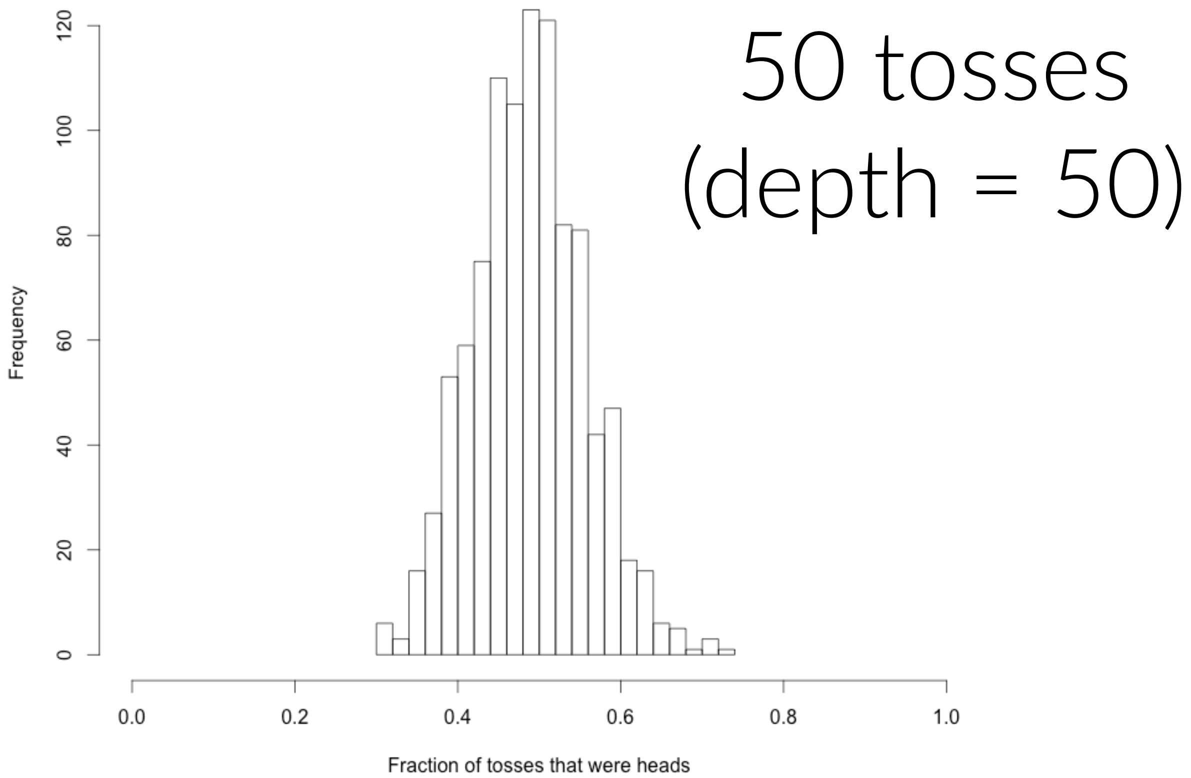
5 tosses
(depth = 5)

Distribution of % heads from 1000 experiments with 20 tosses each



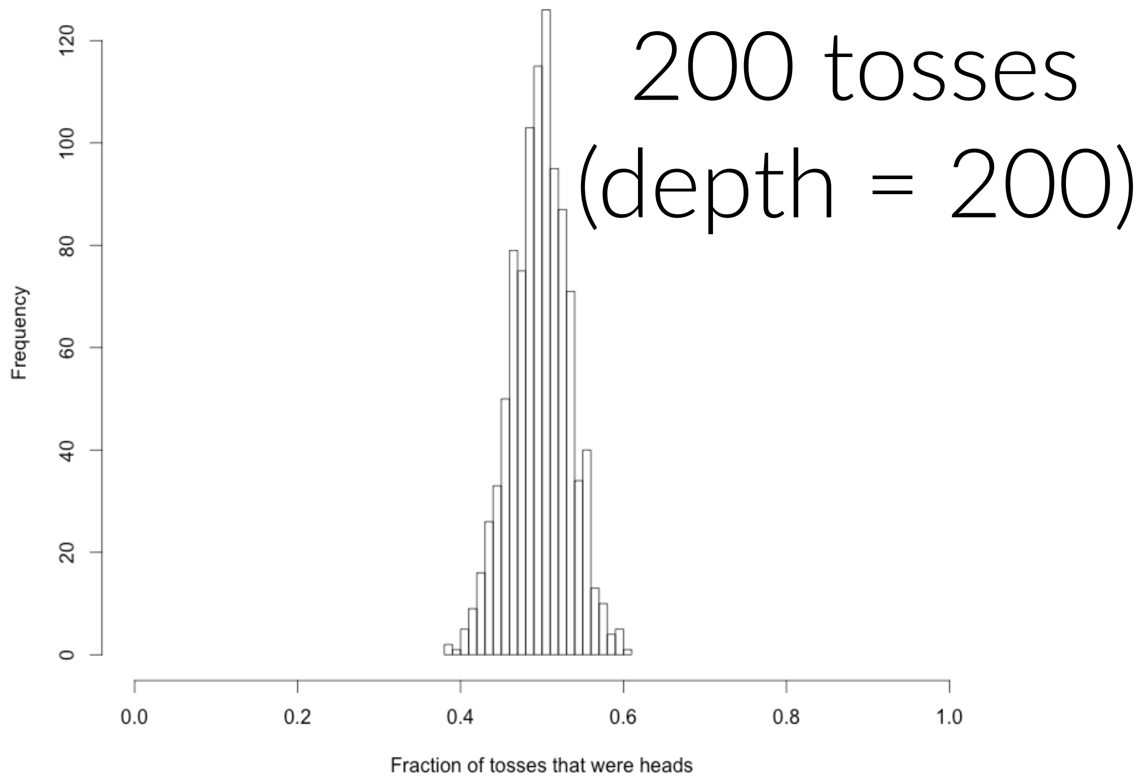
20 tosses
(depth = 20)

Distribution of % heads from 1000 experiments with 50 tosses each



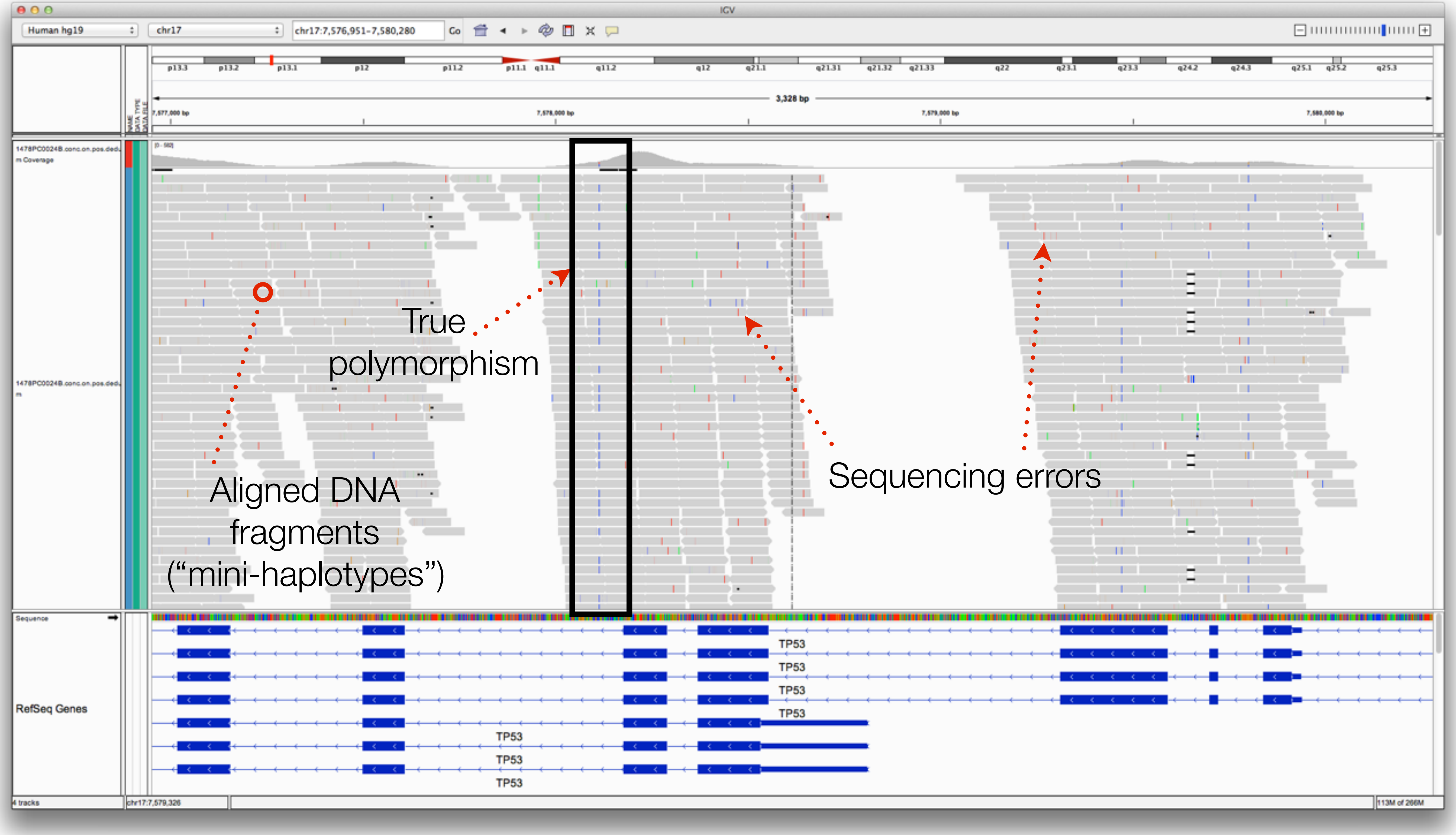
50 tosses
(depth = 50)

Distribution of % heads from 1000 experiments with 200 tosses each



200 tosses
(depth = 200)

Sequencers make mistakes!

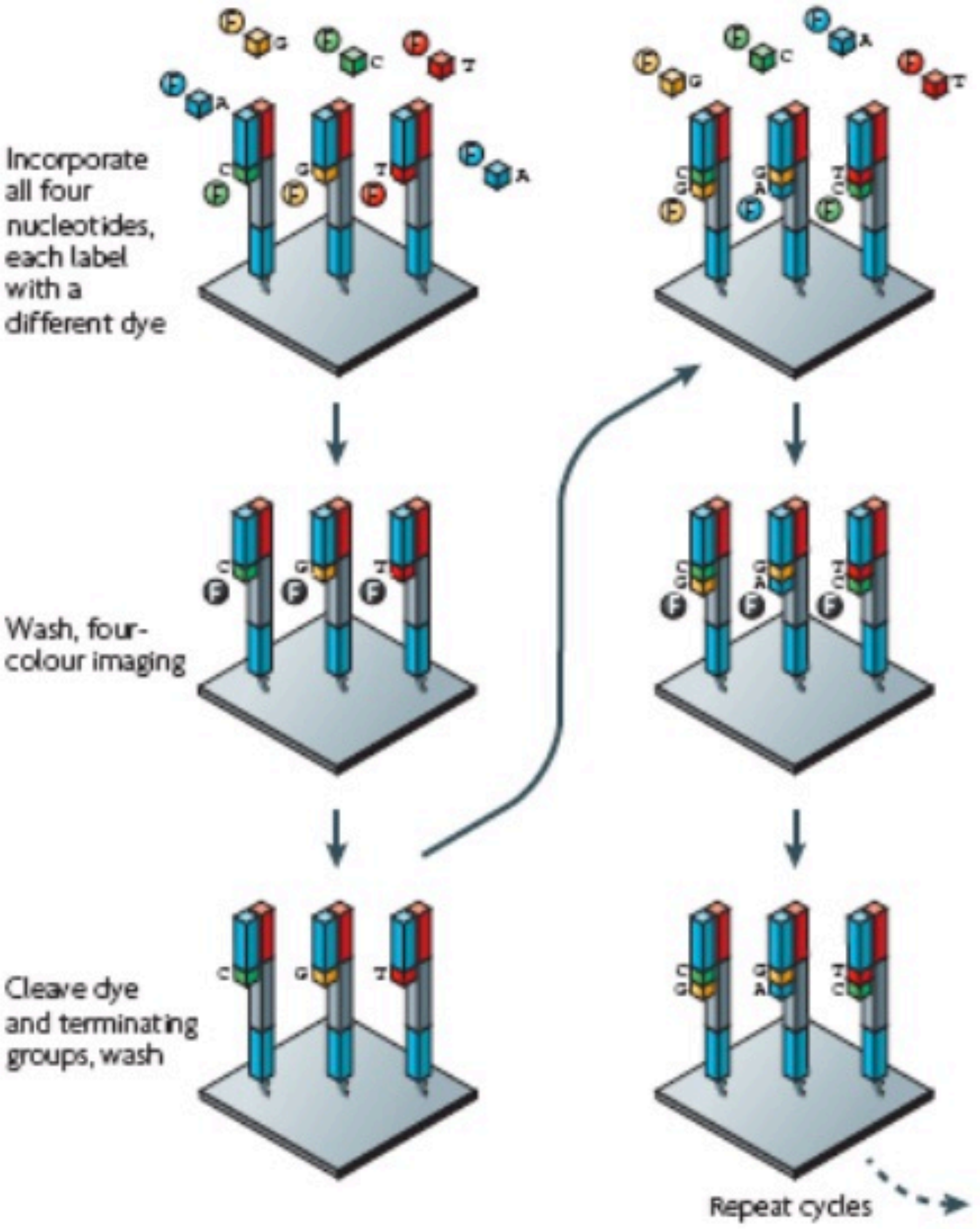


True polymorphism

Sequencing errors

Aligned DNA fragments ("mini-haplotypes")

Not surprising: Solexa sequencing is ~stargazing. Think about this. It is amazing. Error rate is ~0.1%!!!



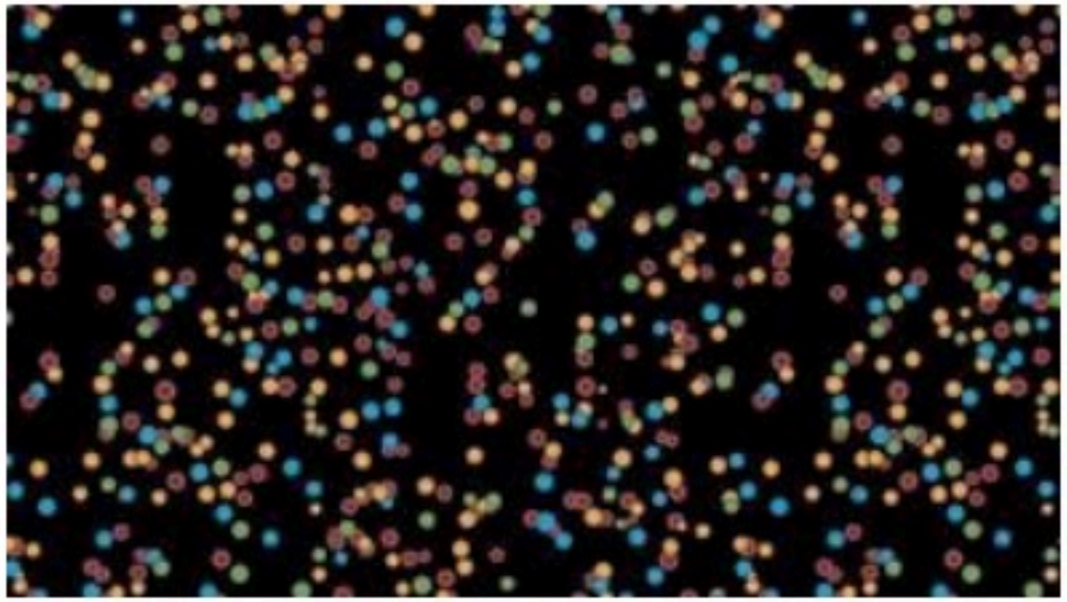
Incorporate all four nucleotides, each label with a different dye

Wash, four-colour imaging

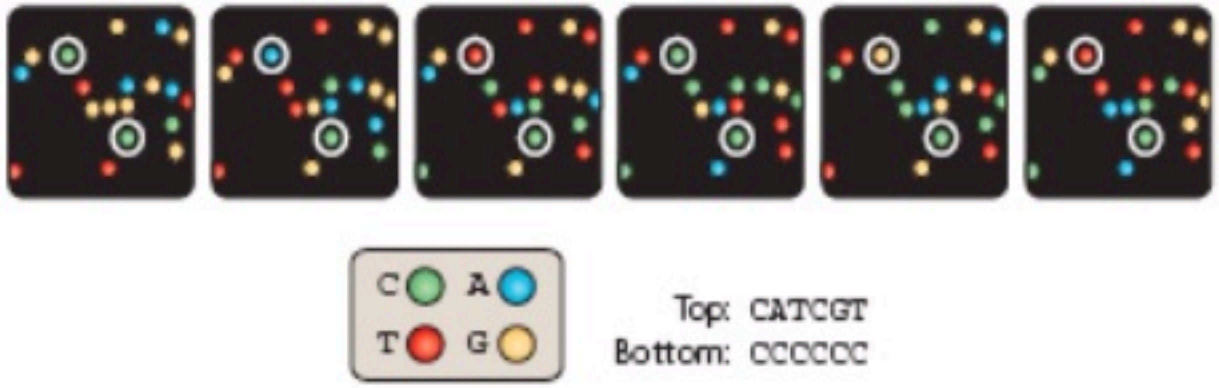
Cleave dye and terminating groups, wash

Repeat cycles

4 images merged



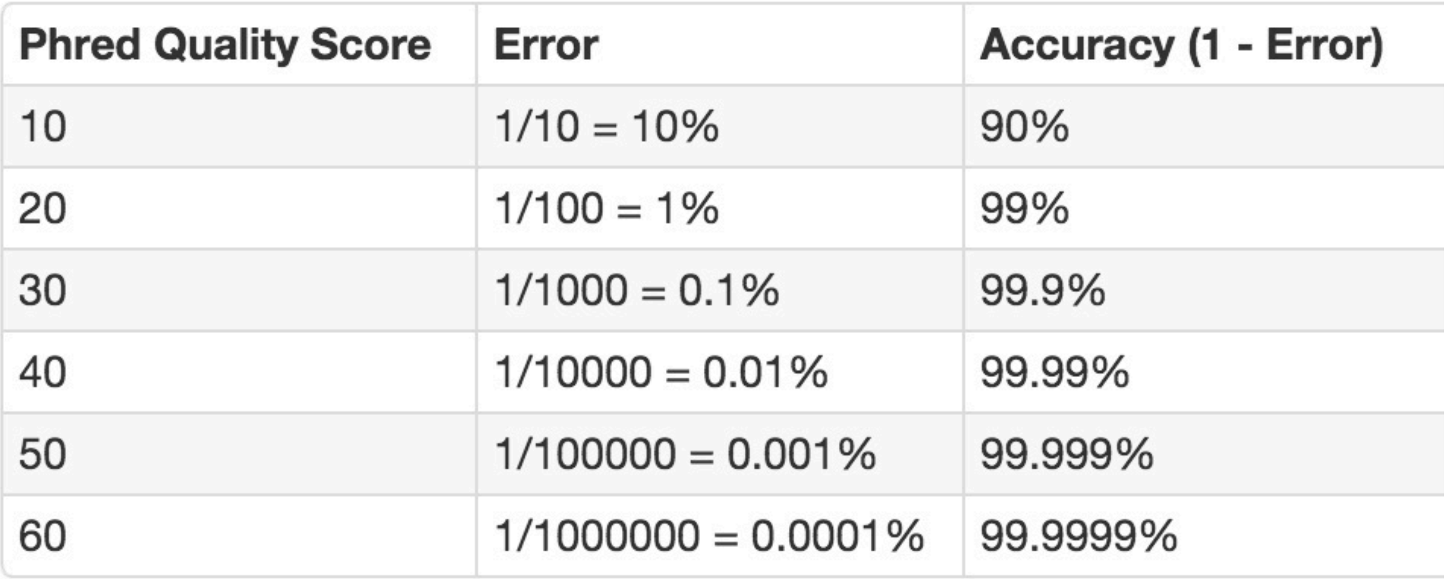
6 cycles w/ base-calling



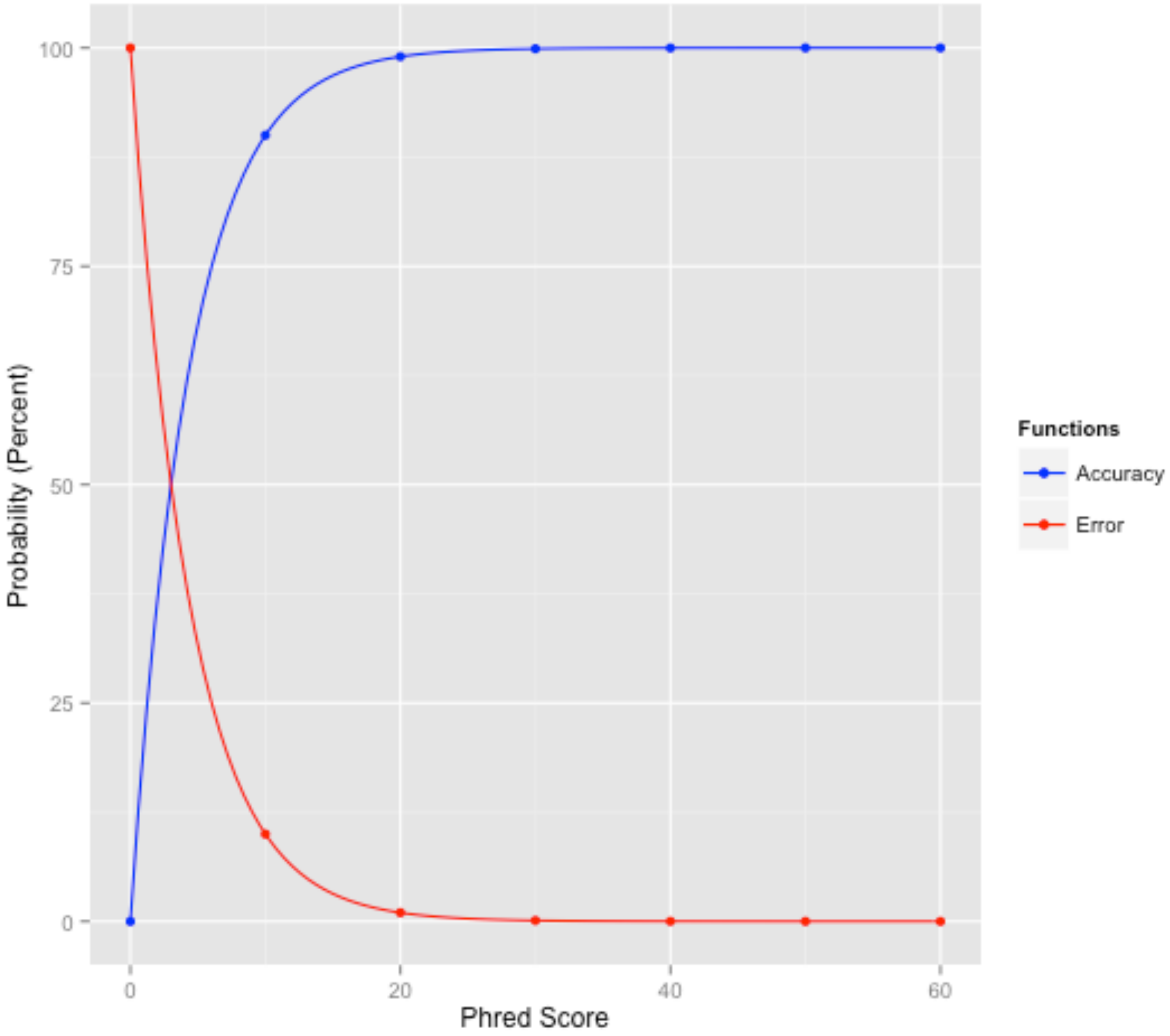
Top: CATCGT
Bottom: CCCCCC

“Phred-scaled” Quality Scores

$$Q = -10 * \log_{10}(P(\text{error}))$$

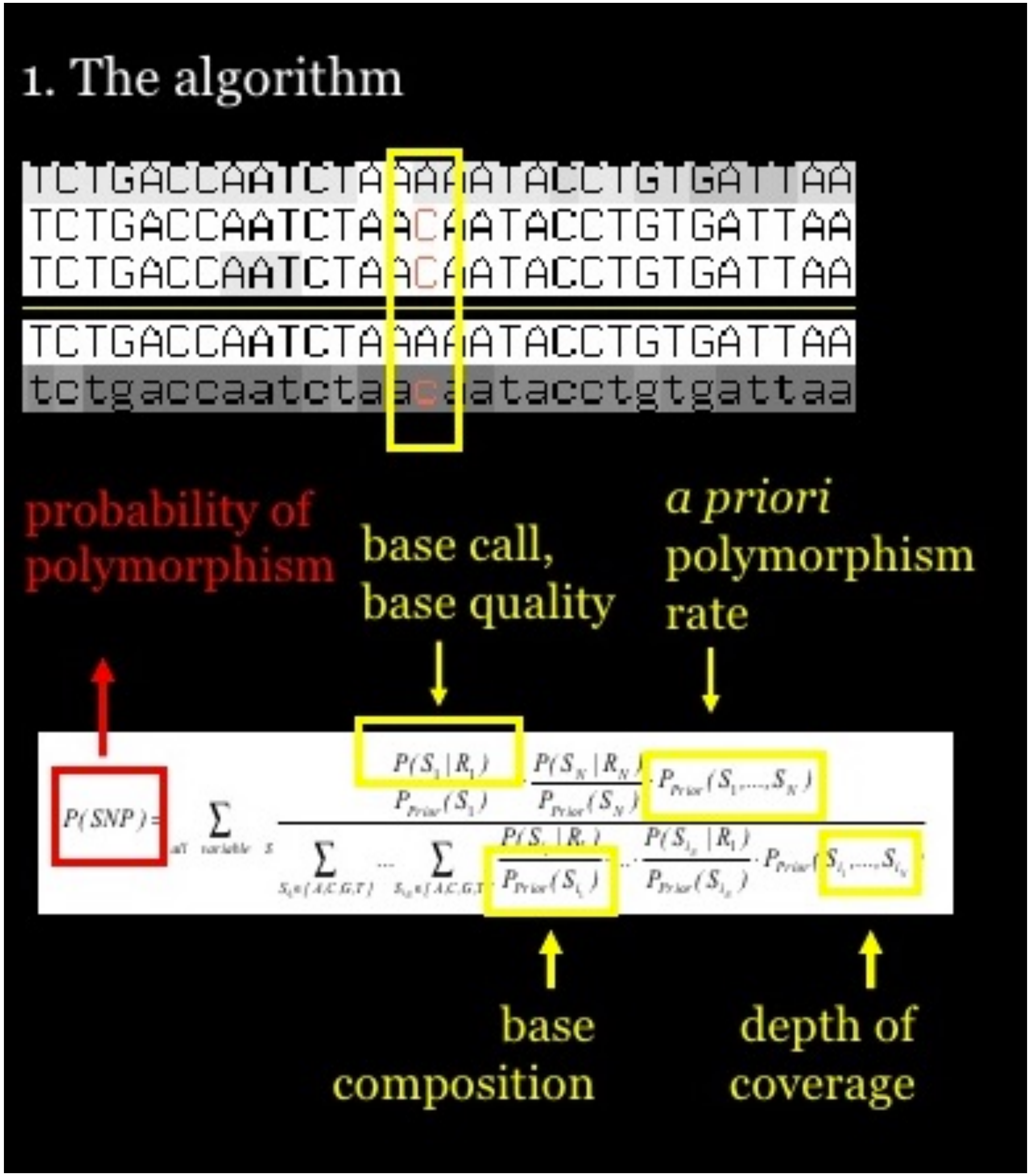


Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%



http://gatkforums.broadinstitute.org/gatk/discussion/4260/how-should-i-interpret-phred-scaled-quality-scores

PolyBayes: the first Bayesian approach to SNP discovery. Accounts for base quality. Predecessor to FreeBayes


1. The algorithm
TCTGACCAATCTAAAAATACCTGTGATTAA
TCTGACCAATCTAACAATACCTGTGATTAA
TCTGACCAATCTAACAATACCTGTGATTAA
TCTGACCAATCTAAAAATACCTGTGATTAA
tctgaccaatctaacaatacctgtgattaa
probability of polymorphism
base call, base quality
a priori polymorphism rate
base composition
depth of coverage

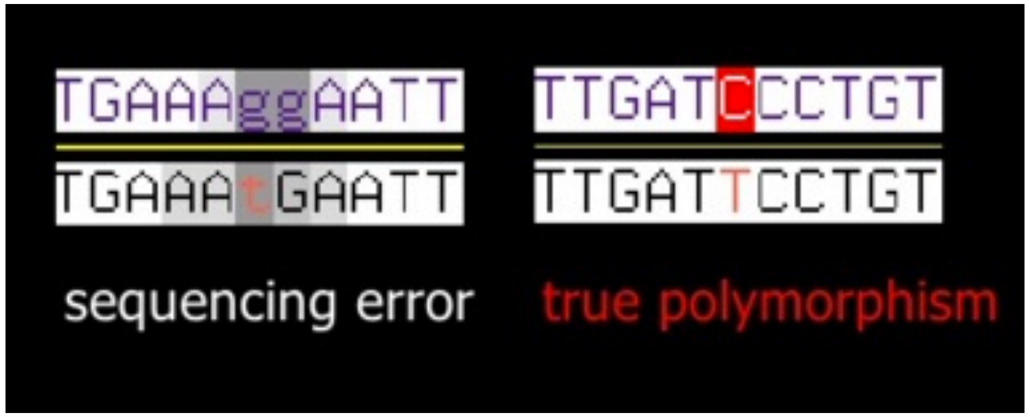
$$P(SNP)=\sum_{\text{all variable } S}\frac{\frac{P(S_1|R_1)}{P_{Prior}(S_1)}\cdots\frac{P(S_N|R_N)}{P_{Prior}(S_N)}\cdot P_{Prior}(S_1,\ldots,S_N)}{\sum_{S_{i_1}\in\{A,C,G,T\}}\cdots\sum_{S_{i_N}\in\{A,C,G,T\}}\frac{P(S_{i_1}|R_1)}{P_{Prior}(S_{i_1})}\cdots\frac{P(S_{i_N}|R_N)}{P_{Prior}(S_{i_N})}\cdot P_{Prior}(S_{i_1},\ldots,S_{i_N})}$$

A general approach to single-nucleotide polymorphism discovery

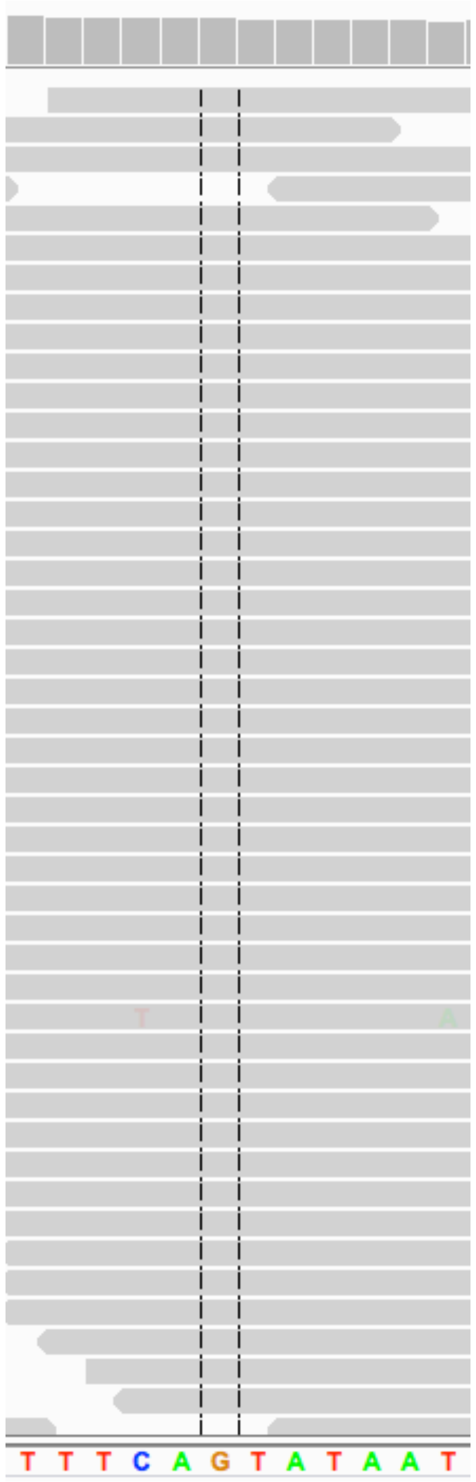
Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

identification and multiple alignment. We analyse these sequences with a novel, Bayesian inference engine, POLYBAYES, to calculate the probability that a given site is polymorphic. Rigorous treatment of base quality permits completely automated evaluation of the full length of all sequences, without limitations on alignment depth. We demonstrate this approach by accurate SNP predictions in human ESTs aligned to finished and working-draft quality genomic sequences, a data set representative of the typical challenges of sequence-based SNP discovery.

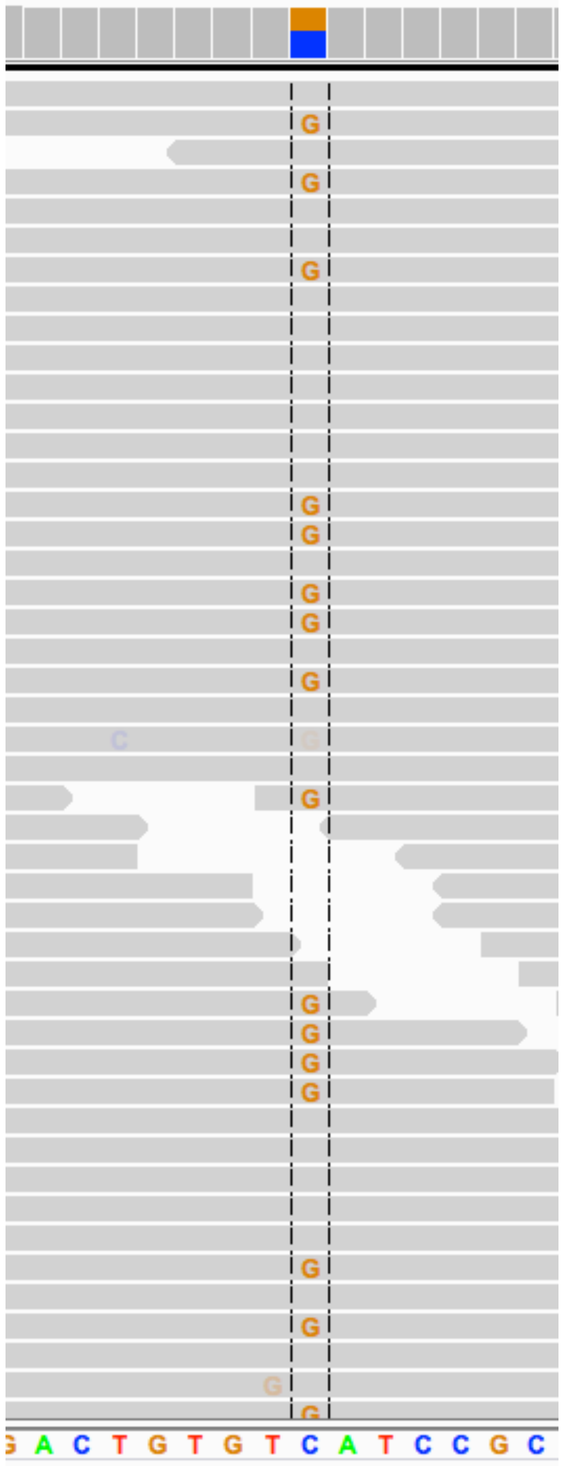
2. Use sequence quality information (base quality values) to distinguish true mismatches from sequencing errors



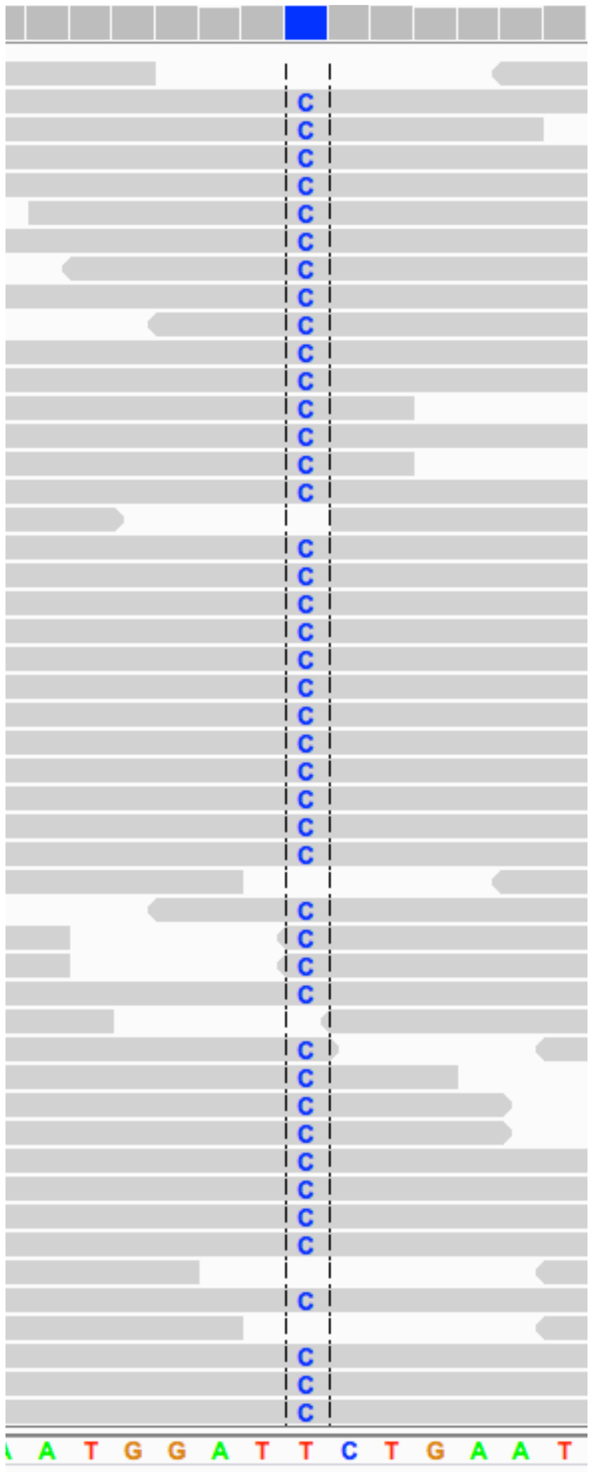
Different (diploid) SNP genotypes



Homozygous for reference
(i.e., both chroms same as ref)

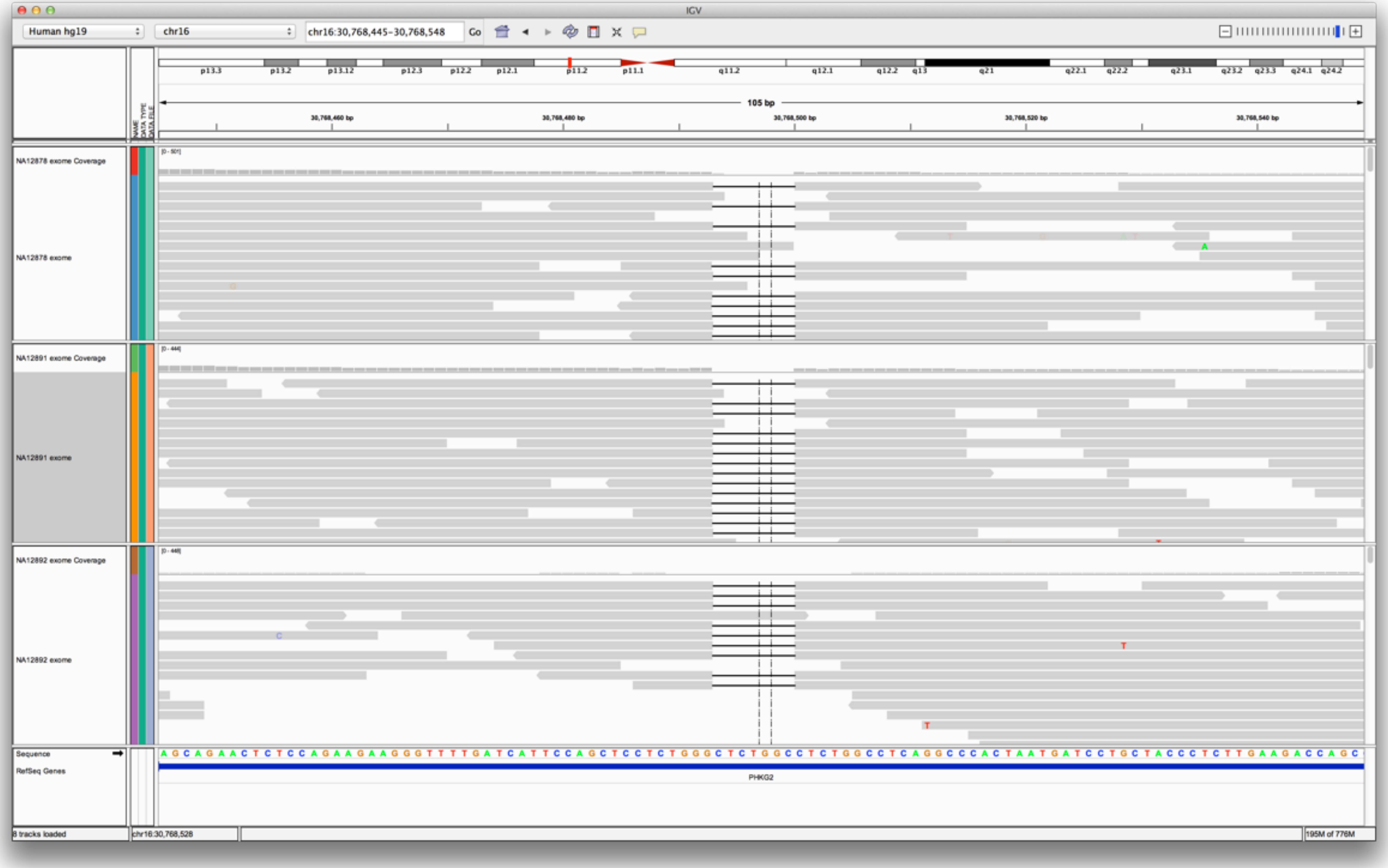


Heterozygous
(i.e., 1 chrom same as ref, 1 diff.)

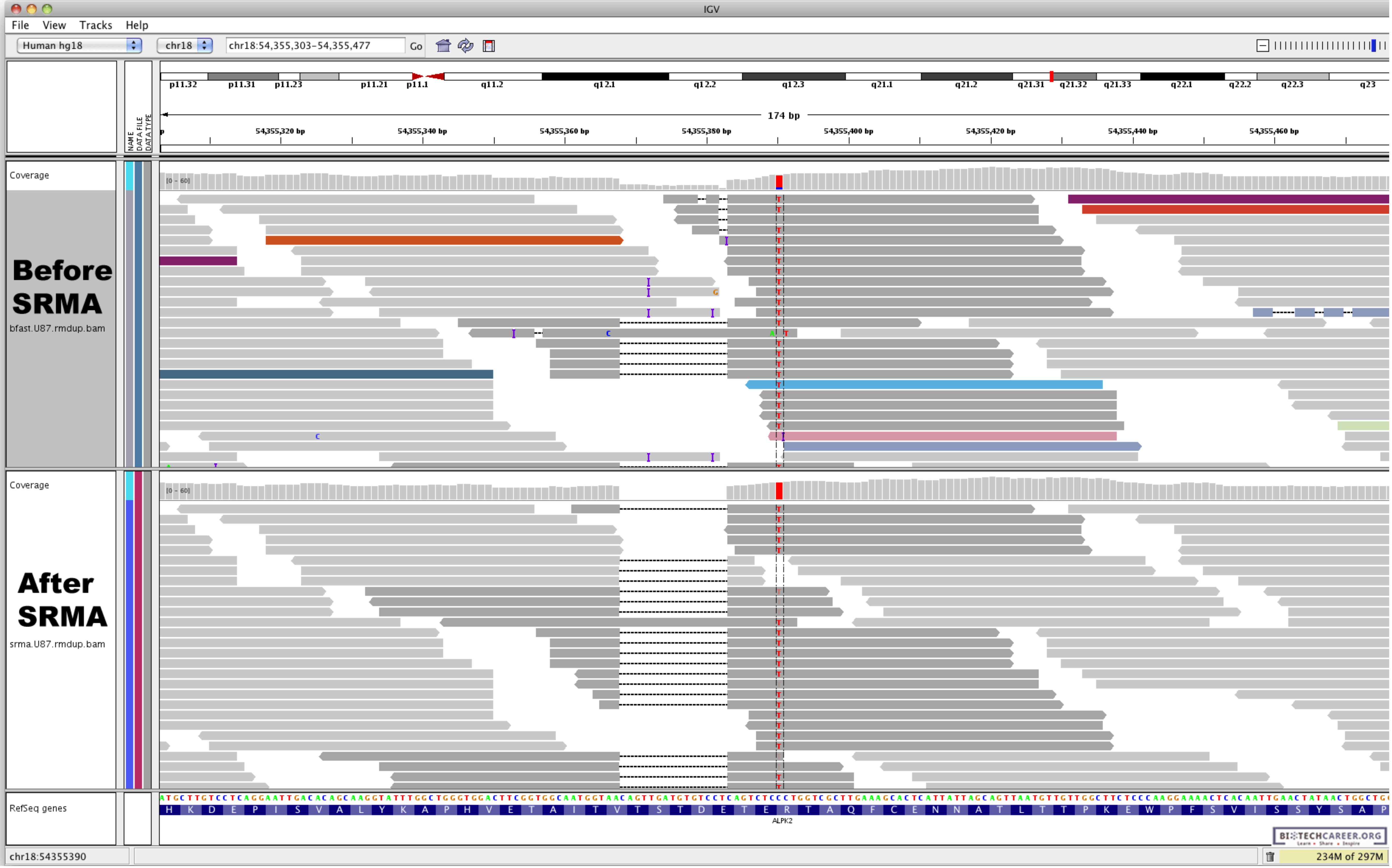


Homozygous for reference
(i.e., both chroms diff than ref)

Insertion-deletion polymorphisms (INDELs)

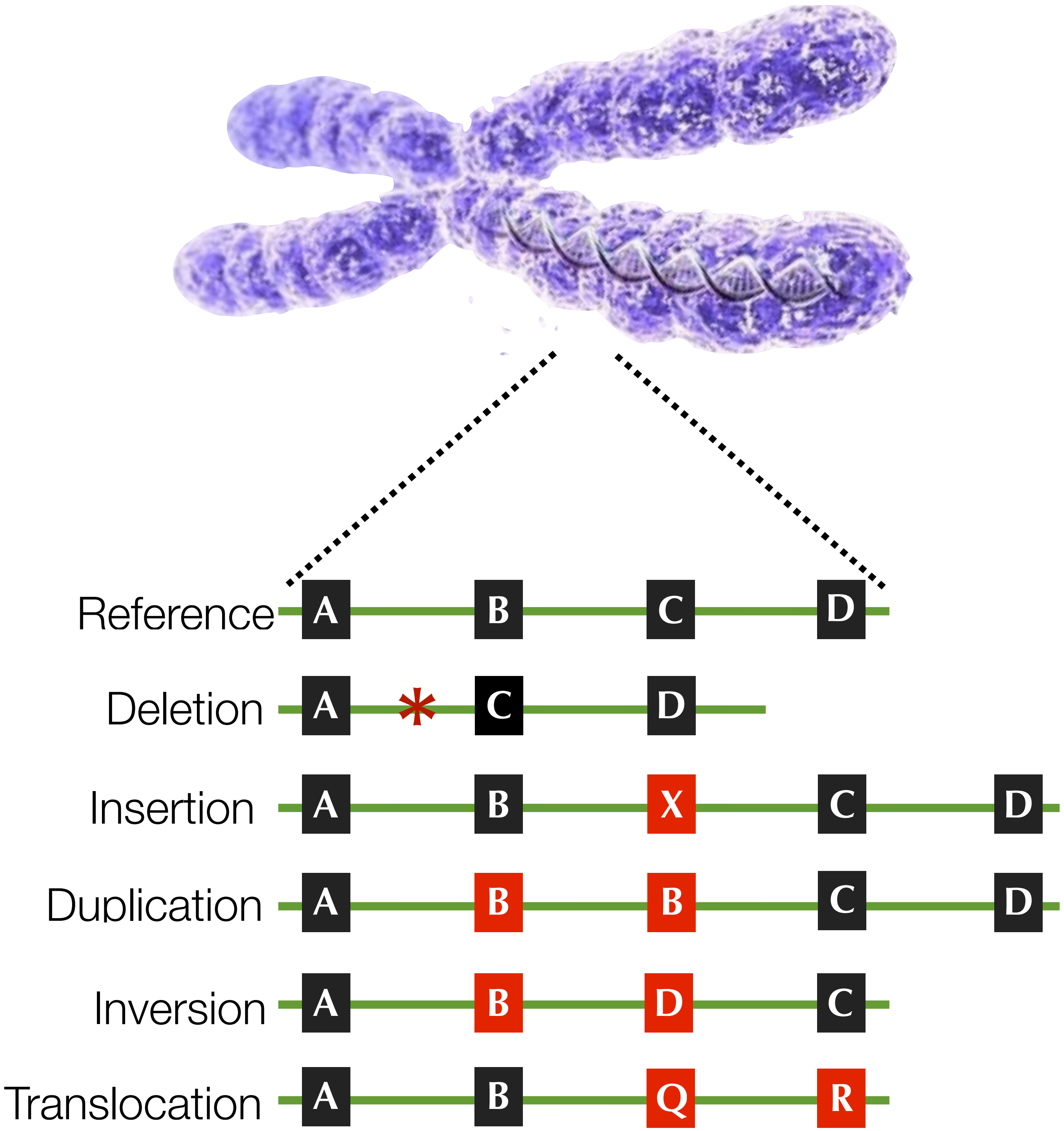


Not always so simple…



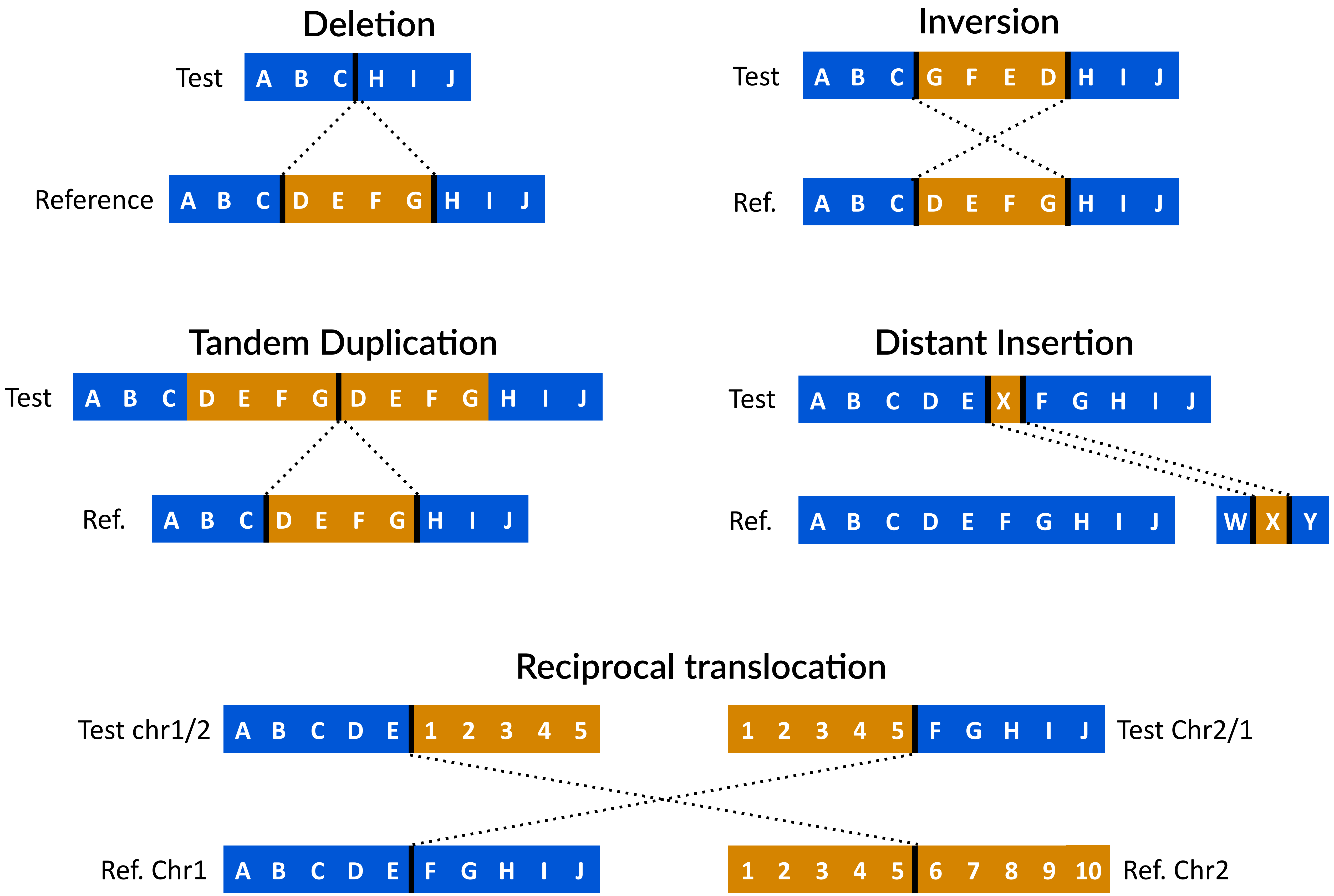
IGV
File View Tracks Help
Human hg18
chr18
chr18:54,355,303-54,355,477
Go
p11.32 p11.31 p11.23 p11.21 p11.1 q11.2 q12.1 q12.2 q12.3 q21.1 q21.2 q21.31 q21.32 q21.33 q22.1 q22.2 q22.3 q23
174 bp
54,355,320 bp
54,355,340 bp
54,355,360 bp
54,355,380 bp
54,355,400 bp
54,355,420 bp
54,355,440 bp
54,355,460 bp
NAME
DATA FILE
DATA TYPE
Coverage
[0 - 60]
Before SRMA
bfast.U87.rmdup.bam
Coverage
[0 - 60]
After SRMA
srma.U87.rmdup.bam
RefSeq genes
ATGCTTGTCCTCAGGAATTGACACAGCAAGGTATTTGGCTGGGTGGACTTCGGTGGCAATGGTAACAGTTGATGTGTCCTCAGTCTCCCTGGTCGCTTGAAAGCACTCATTATTAGCAGTTAATGTTGTTGGCTTCTCCCAAGGAAAACTCACAATTGAACTATAACTGGCTG
H K D E P I S V A L Y K A P H V E T A I T V T S T D E T E R T A Q F C E N N A T L T T P K E W P F S V I S S Y S A P
ALPK2
chr18:54355390
234M of 297M
BIOTECHCAREER.ORG
Learn • Share • Inspire

Structural Variation



- large (>1kb) differences that affect the copy number, orientation, or location of genomic segments
- Common in mammalian genomes (~3-5 thousand between two people)
- A hallmark of cancer
- A major cause of spontaneous disease
- more are functional than SNPs
- very challenging to identify

SV Breakpoints



Deletion

Inversion

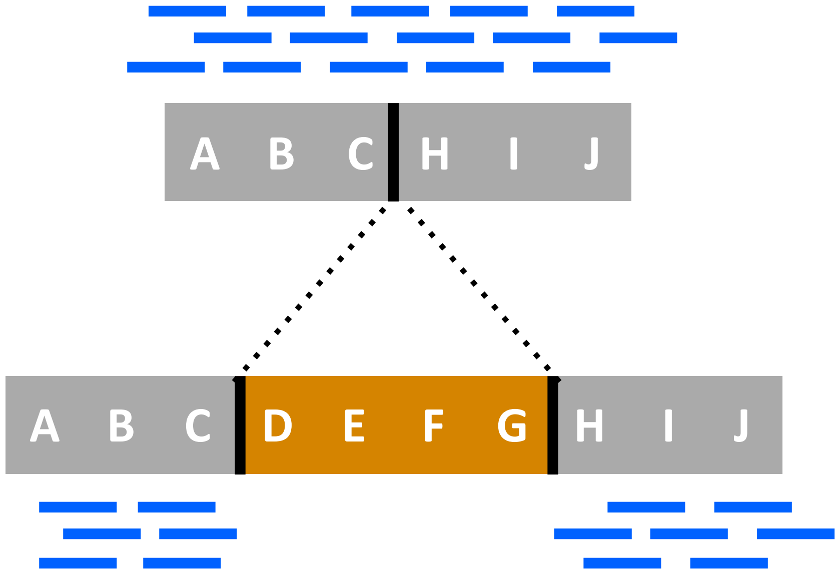
Tandem Duplication

Distant Insertion

Reciprocal translocation

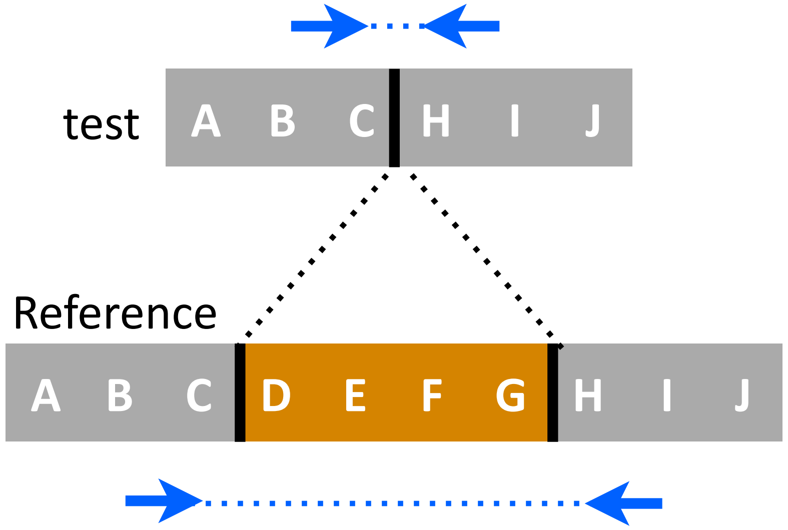
Detecting SVs from alignments

1) Read depth



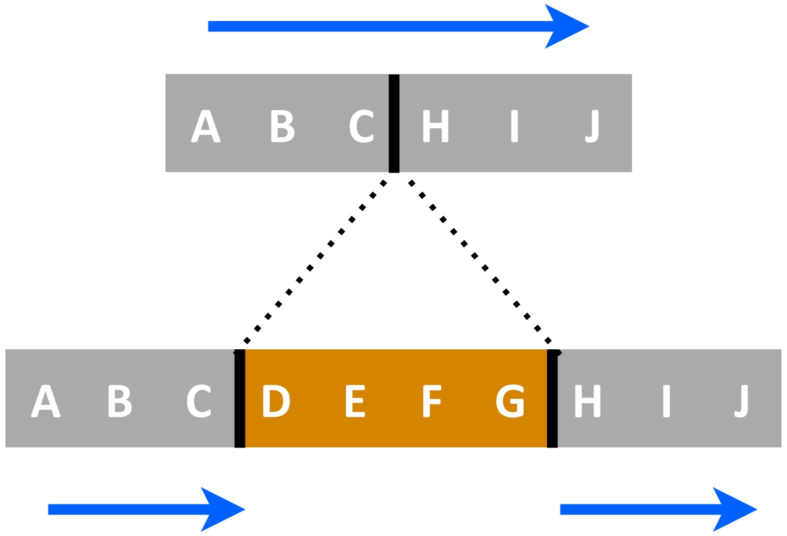
2) Paired-end mapping

paired-reads (or strobe)

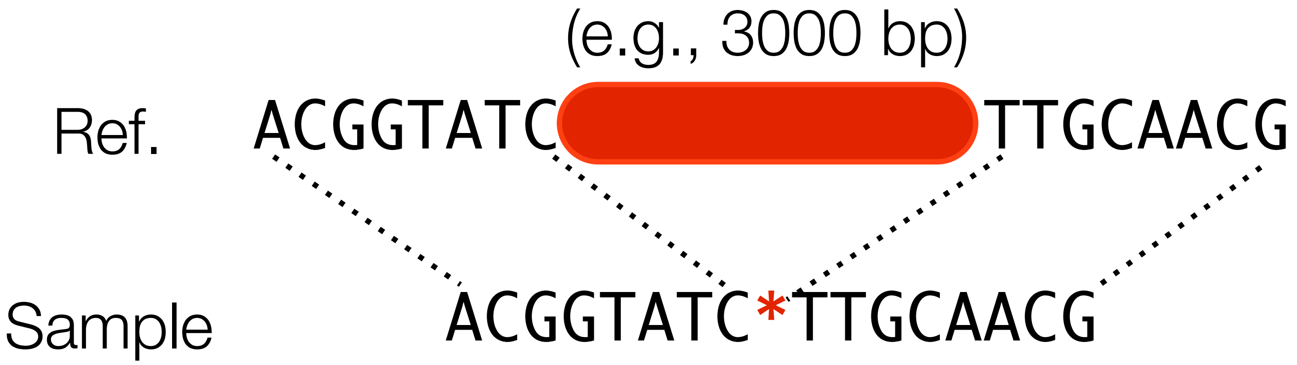


3) Split mapping

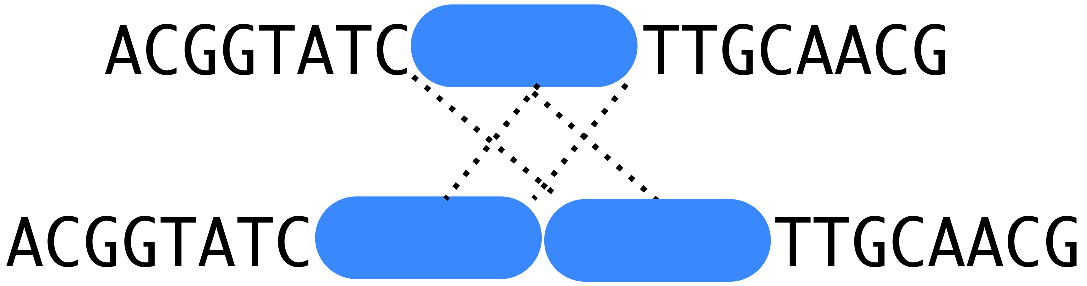
long reads or contigs



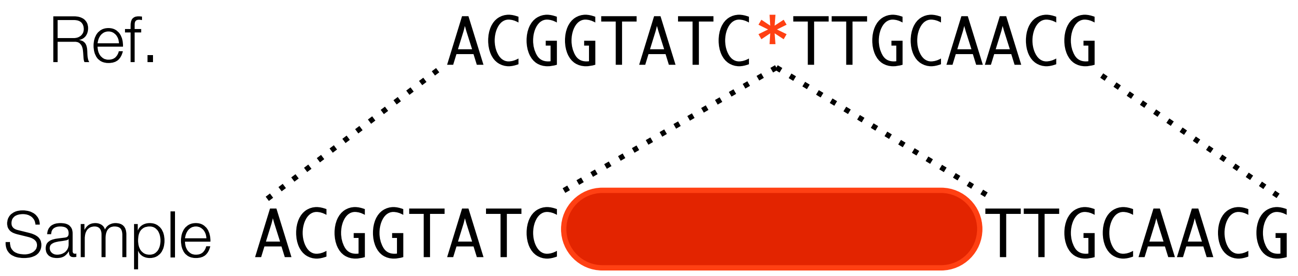
SV Alignment patterns



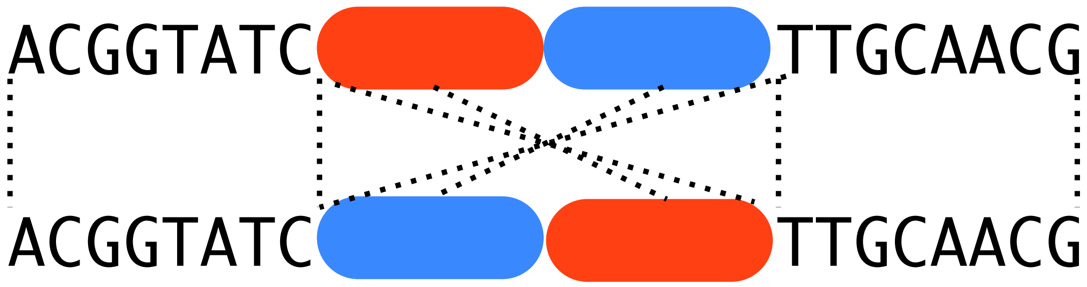
Deletion in sample



Duplication in sample

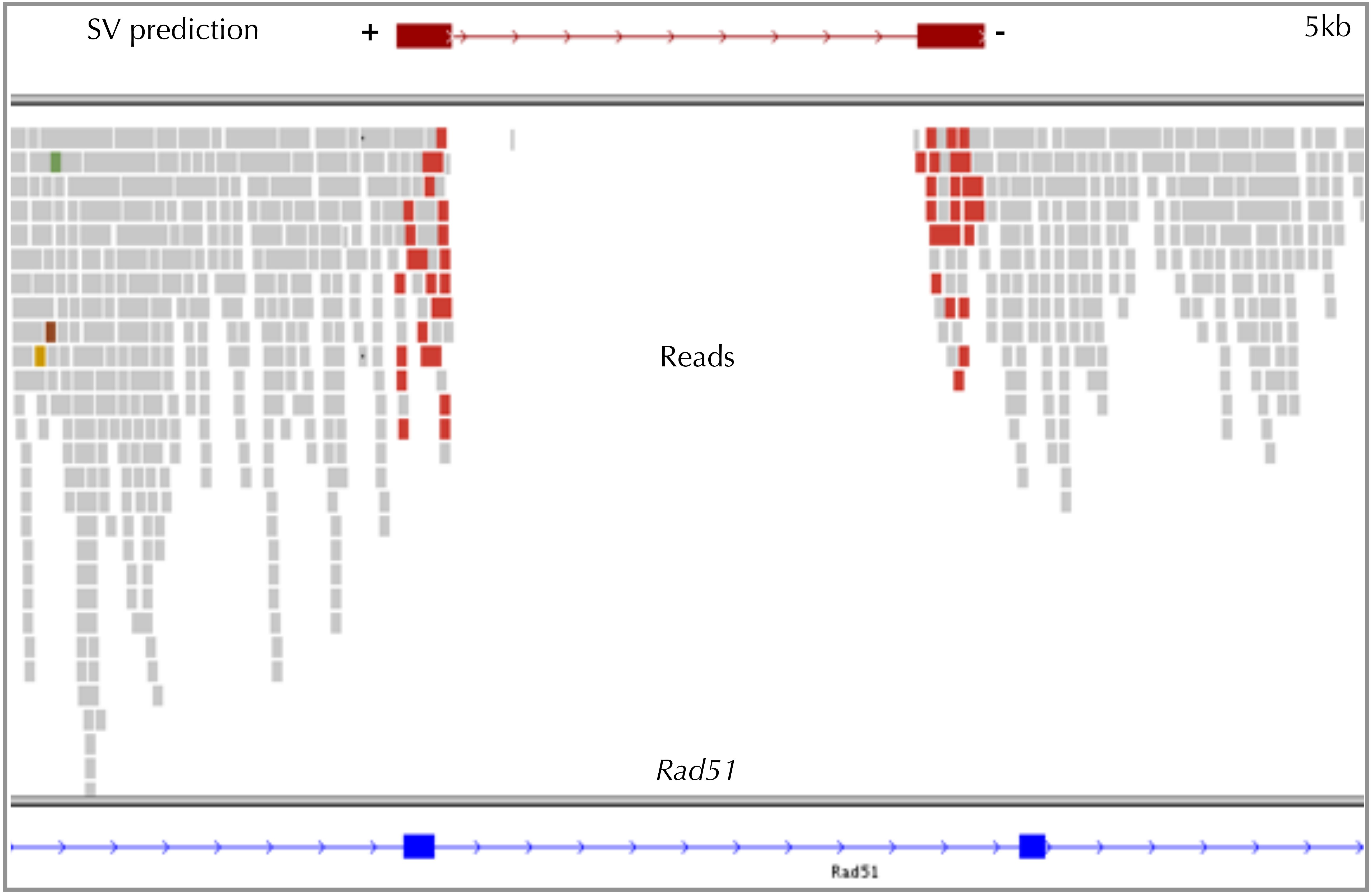


Insertion in sample

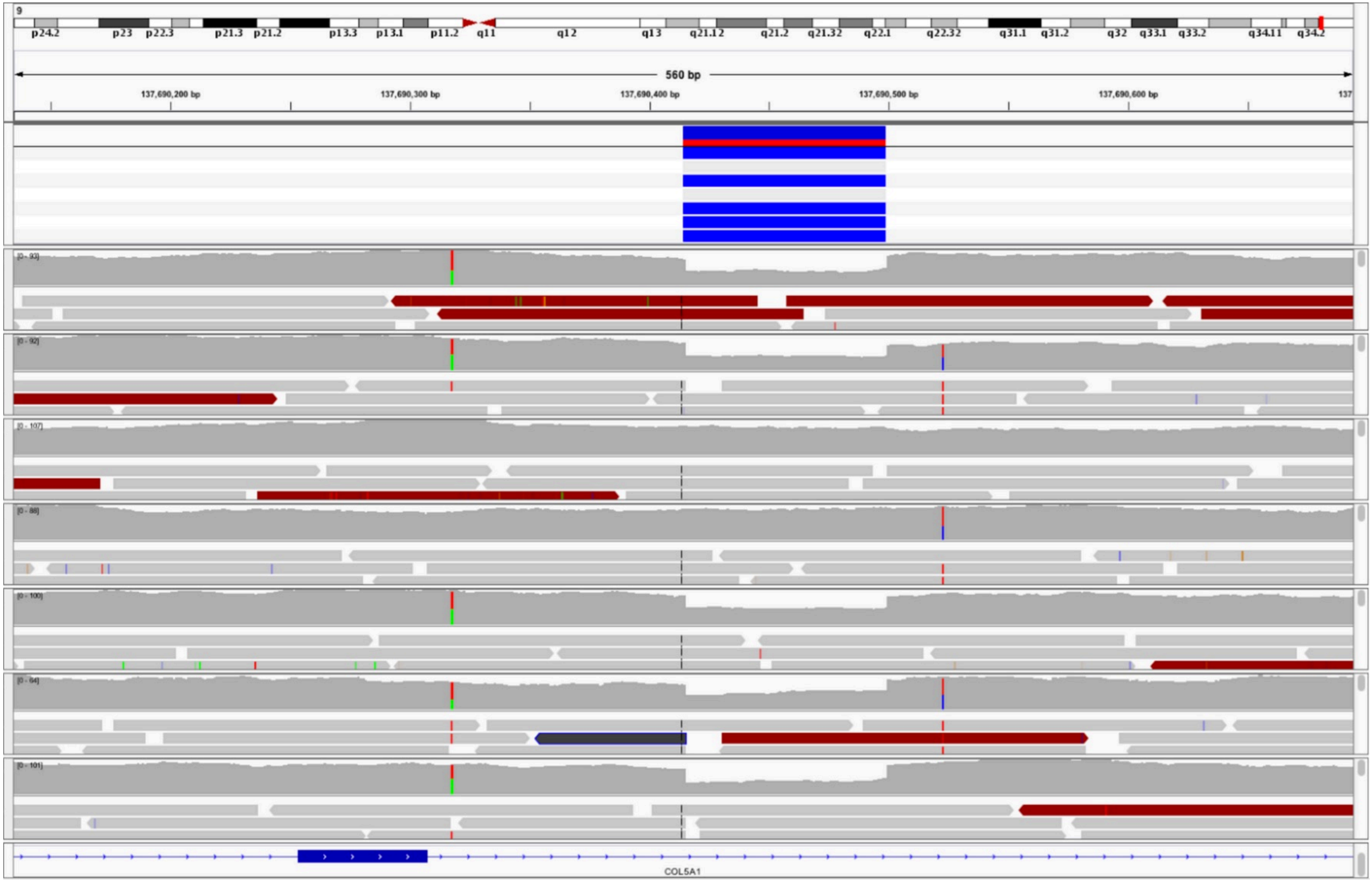


Duplication in sample

A real deletion detected with sequencing



A deletion segregating in a family



Summary

- Four major classes of genetic variation: single-nucleotide polymorphisms (SNPs), insertion-deletion polymorphisms (INDELs), structural variants (SVs), and mobile-element insertions (not discussed).
- Modern sequencing technologies provide an excellent substrate for detecting all forms of genetic variations
- However, sufficient sequencing depth and per-base accuracy are necessary for **comprehensive** and **accurate** variant discovery.
- Improved sequencing technologies (e.g. longer, more accurate reads) and better algorithms (e.g., modeling error, phase-aware) are the path forward.