

Prior to 2005, genomic datasets were not commonly large

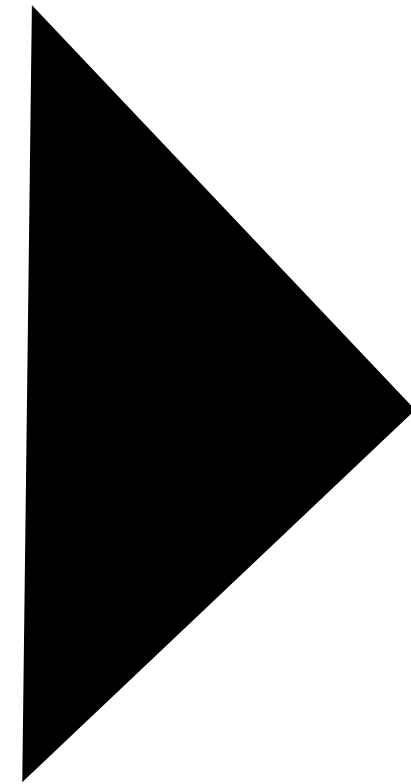
Allozymes

AFLPs

RFLPs

Sanger Sequencing

Microsatellites



All analyzable on personal computers
mostly with programs with GUIs; surveying
a handful to dozens of loci

Next-generation DNA sequencing

~2005 the first companies came out with “next-generation sequencing” technologies

Whole Genome “resequencing” — sequencing the whole genome with the goal to map to a known reference

“Low coverage” resequencing — Coverage is too low to allow for confident genotype calls at the individual level, so downstream analyses rely on genotype uncertainty information, not genotype calls

“Reduced representation” genomic sequencing — RADseq / GBS / ddRAD

Next-generation DNA sequencing

~2005 the first companies came out with “next-generation sequencing” technologies

Whole Genome “resequencing” — sequencing the whole genome with short reads and mapping them to a known reference

“Low coverage”

calls a low number of reads to allow for confident genotype calls

inform

genotype calls

These approaches have been revolutionary for the study of non-model systems in evolutionary biology!

“Reduced representation” genomic sequencing — RADseq / GBS / ddRAD

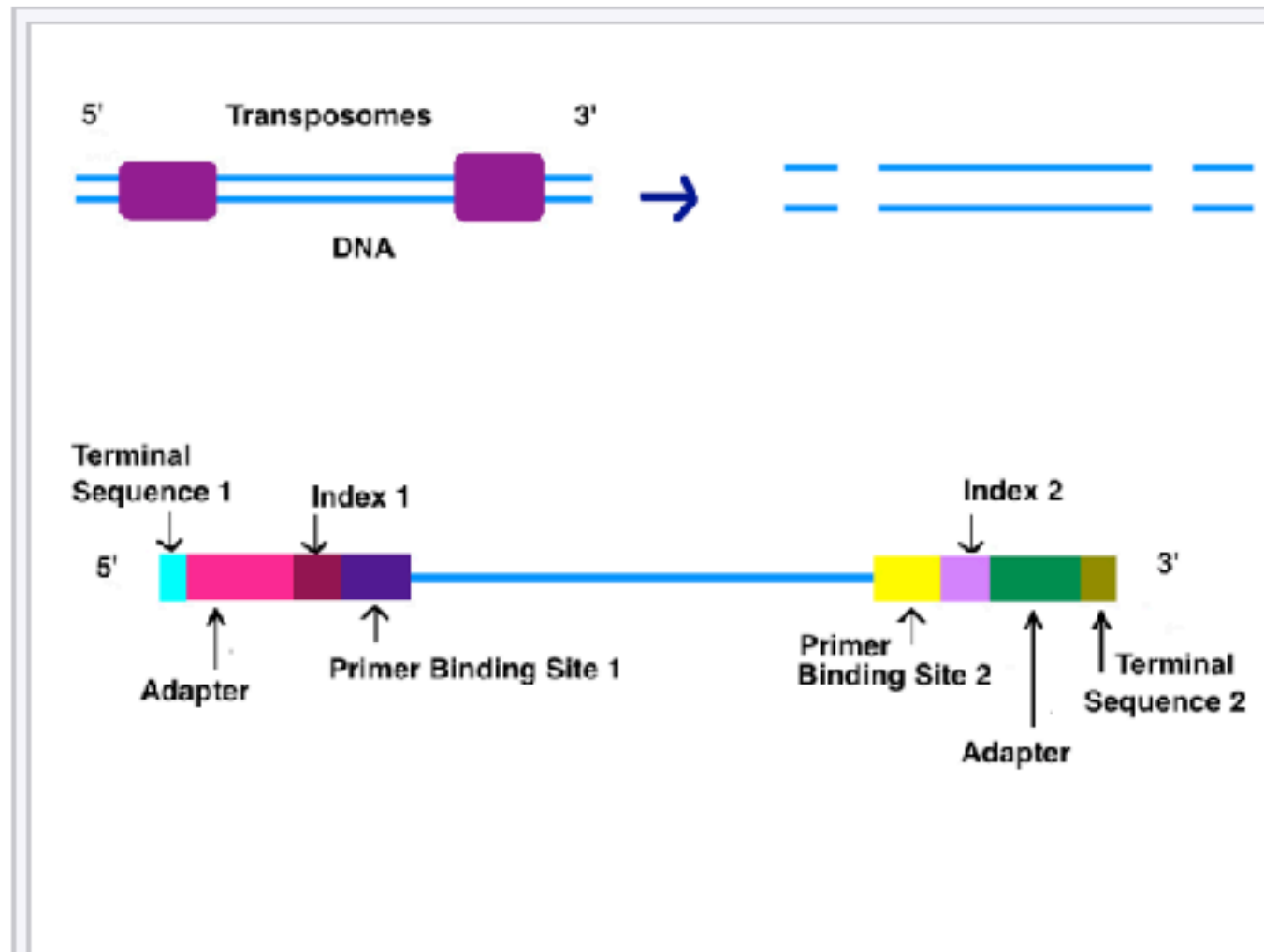
Next-generation DNA sequencing

~2005 the first companies came out with “next-generation sequencing” technologies

Illumina’s technology was the “winner” by far

<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

1) Sample (library) prep:



Double stranded DNA is cleaved by transposomes. The cut ends are repaired and adapters, indices, primer binding sites, and terminal sites are added to each strand of the DNA. Image based in part on illumina's sequencing video^[7]

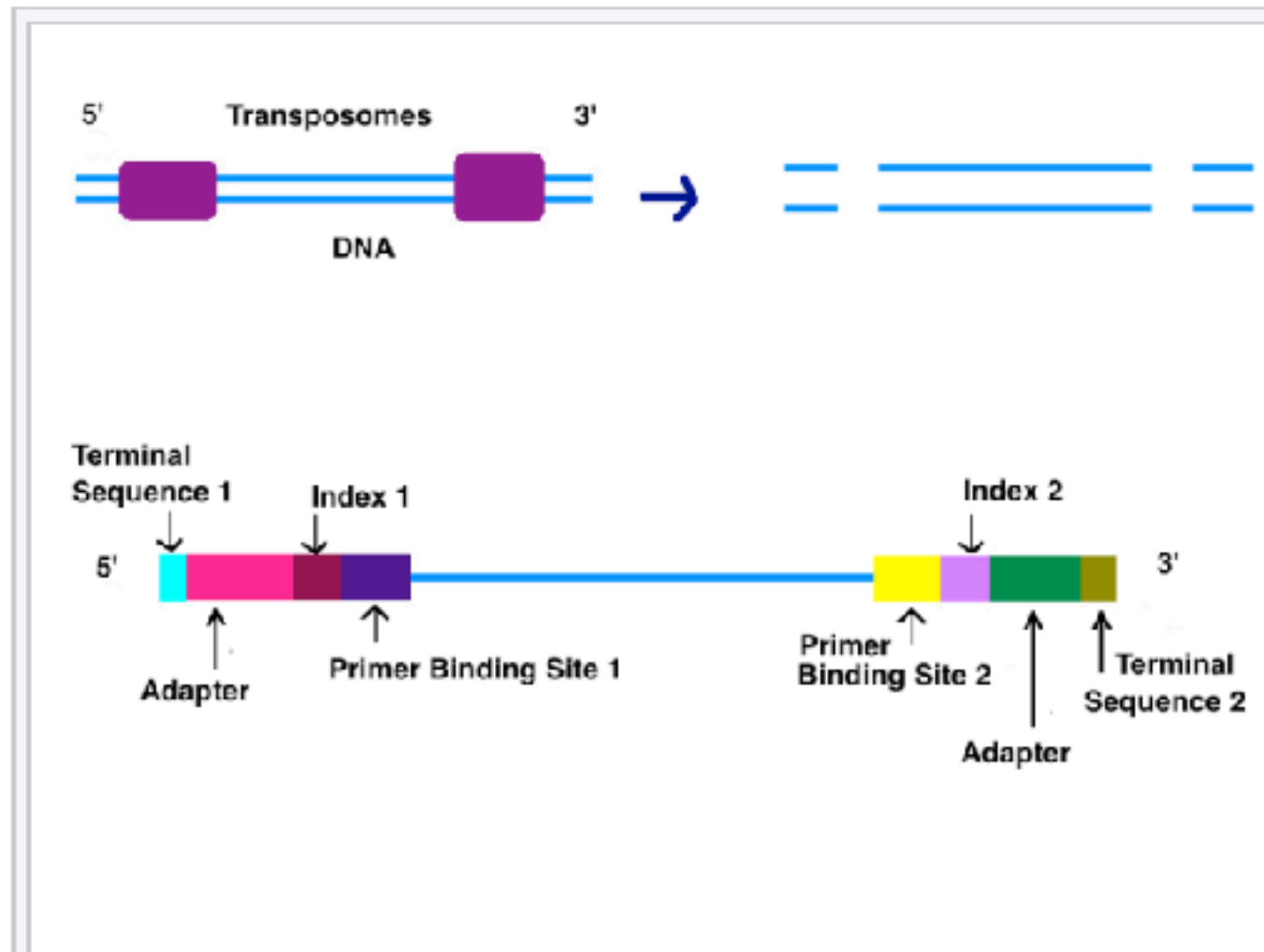
Next-generation DNA sequencing

~2005 the first companies came out with “next-generation sequencing” technologies

Illumina’s technology was the “winner” by far

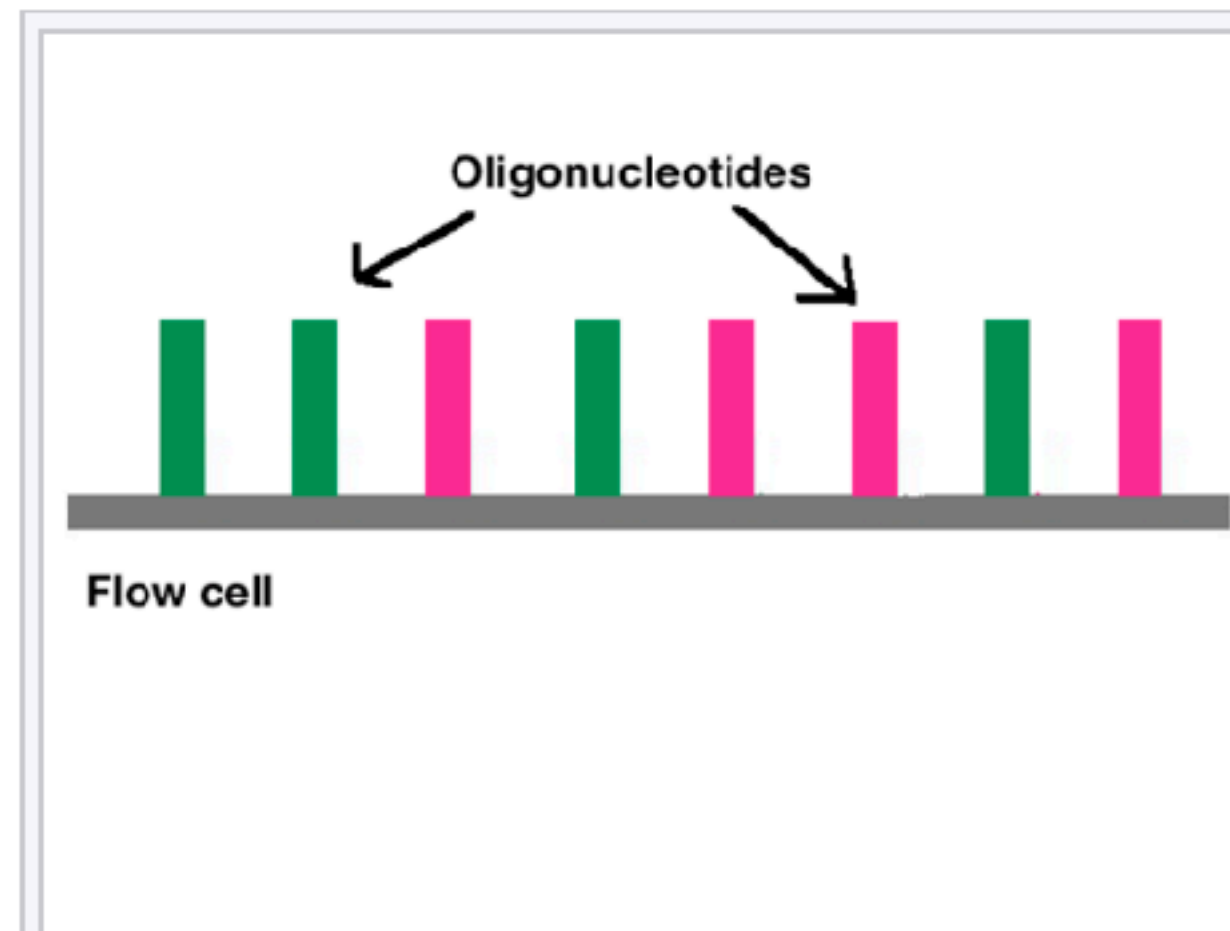
<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

1) Sample (library) prep:

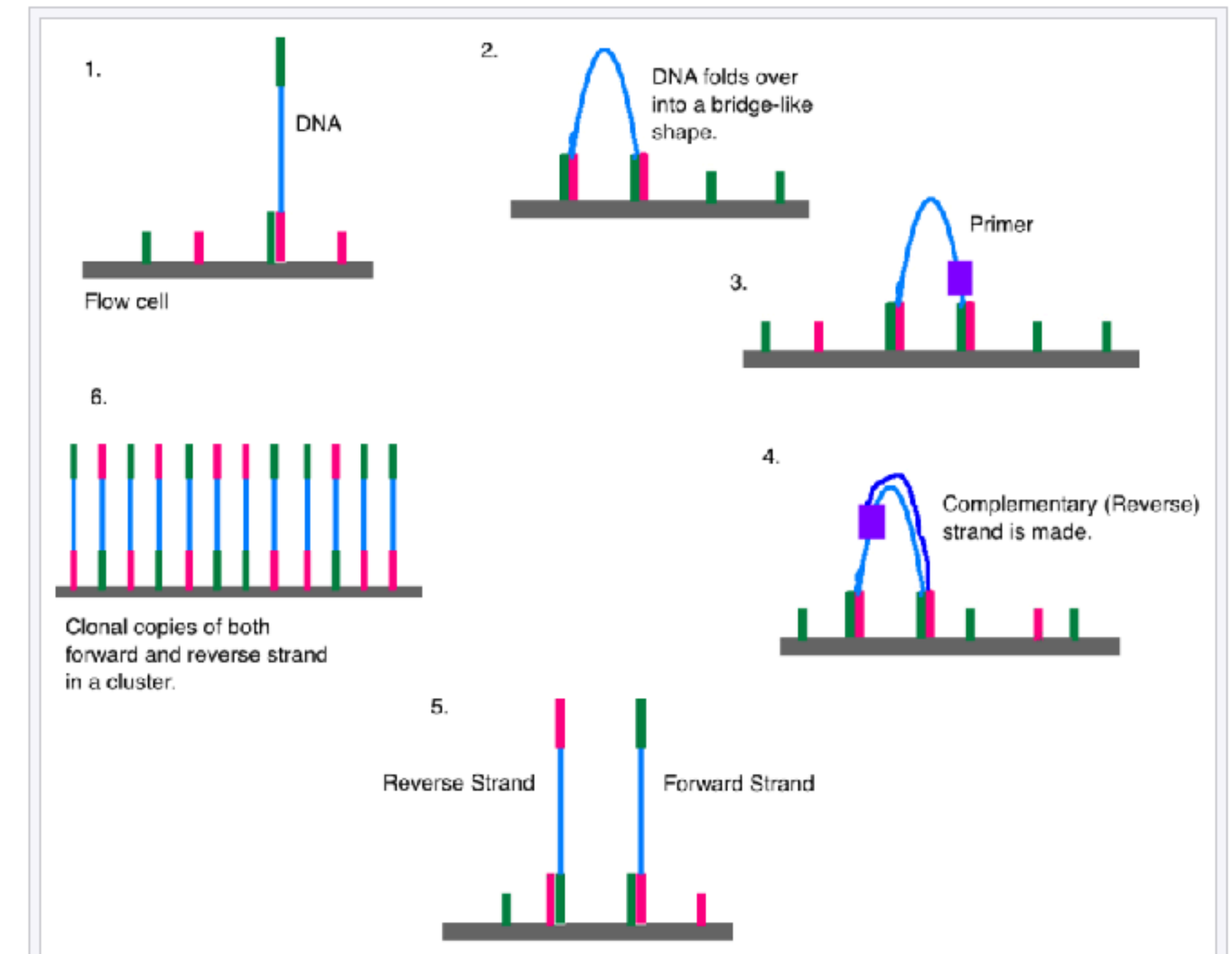


Double stranded DNA is cleaved by transposomes. The cut ends are repaired and adapters, indices, primer binding sites, and terminal sites are added to each strand of the DNA. Image based in part on illumina's sequencing video^[7]

2) Cluster generation:



Millions of oligos line the bottom of each flow cell lane.



The DNA attaches to the flow cell via complementary sequences. The strand bends over and attaches to a second oligo forming a bridge. A polymerase synthesizes the reverse strand. The two strands release and straighten. Each forms a new bridge (bridge amplification). The result is a cluster of DNA forward and reverse strand clones.

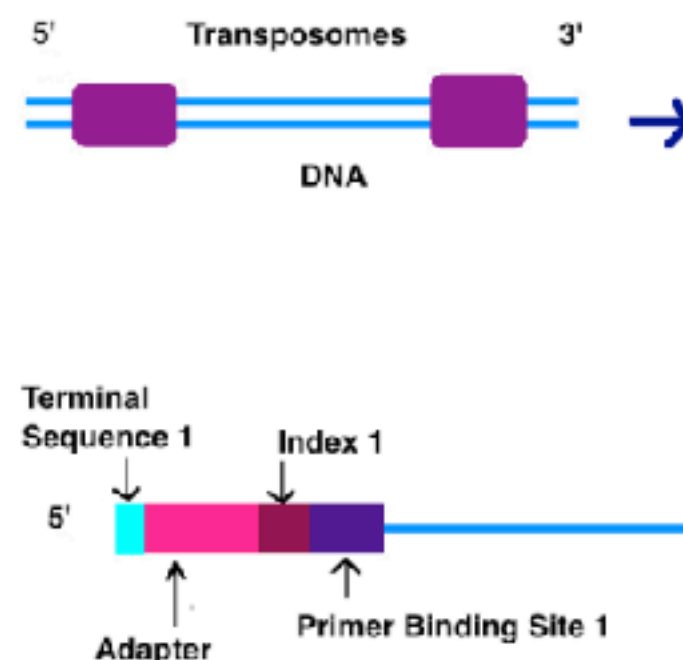
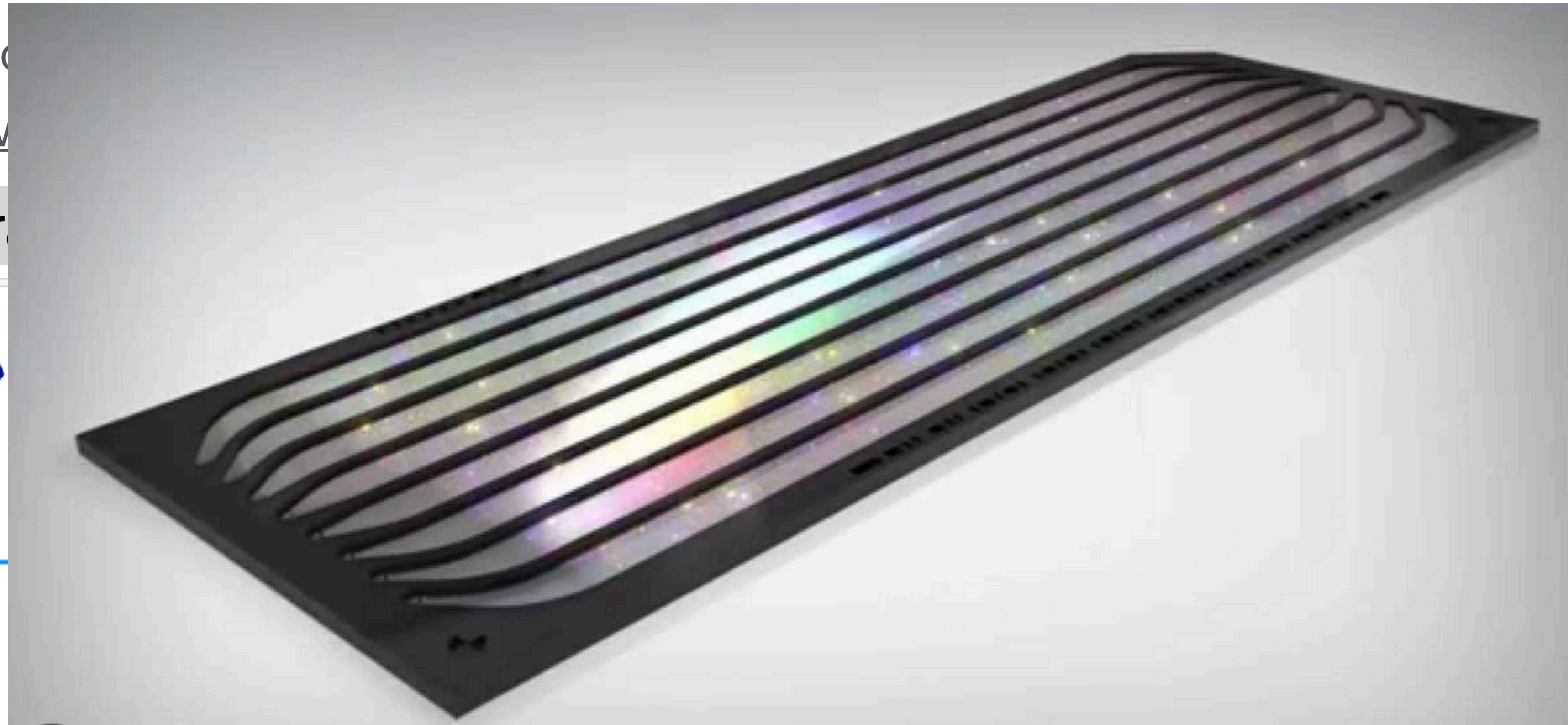
Next-generation DNA sequencing

~2005 the first companies came out with “next-generation sequencing” technologies

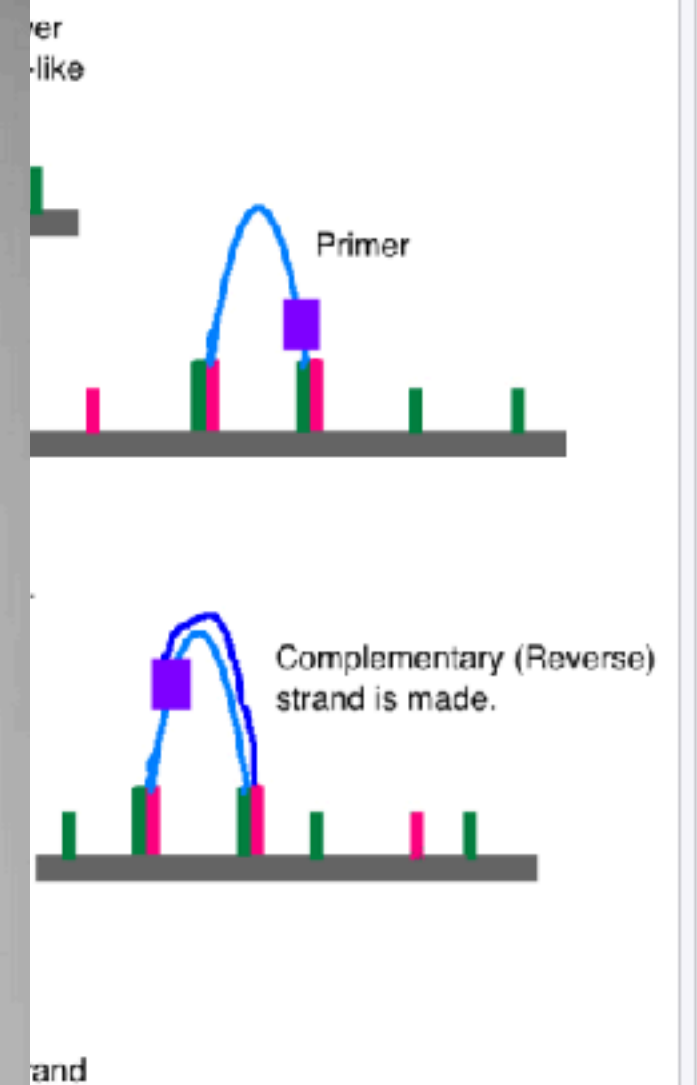
Illumina's tech

<https://www.illumina.com>

1) Sample (library)



Double stranded DNA is cleaved by transposomes. The cut ends are repaired and adapters, indices, primer binding sites, and terminal sites are added to each strand of the DNA. Image based in part on illumina's sequencing video^[7]



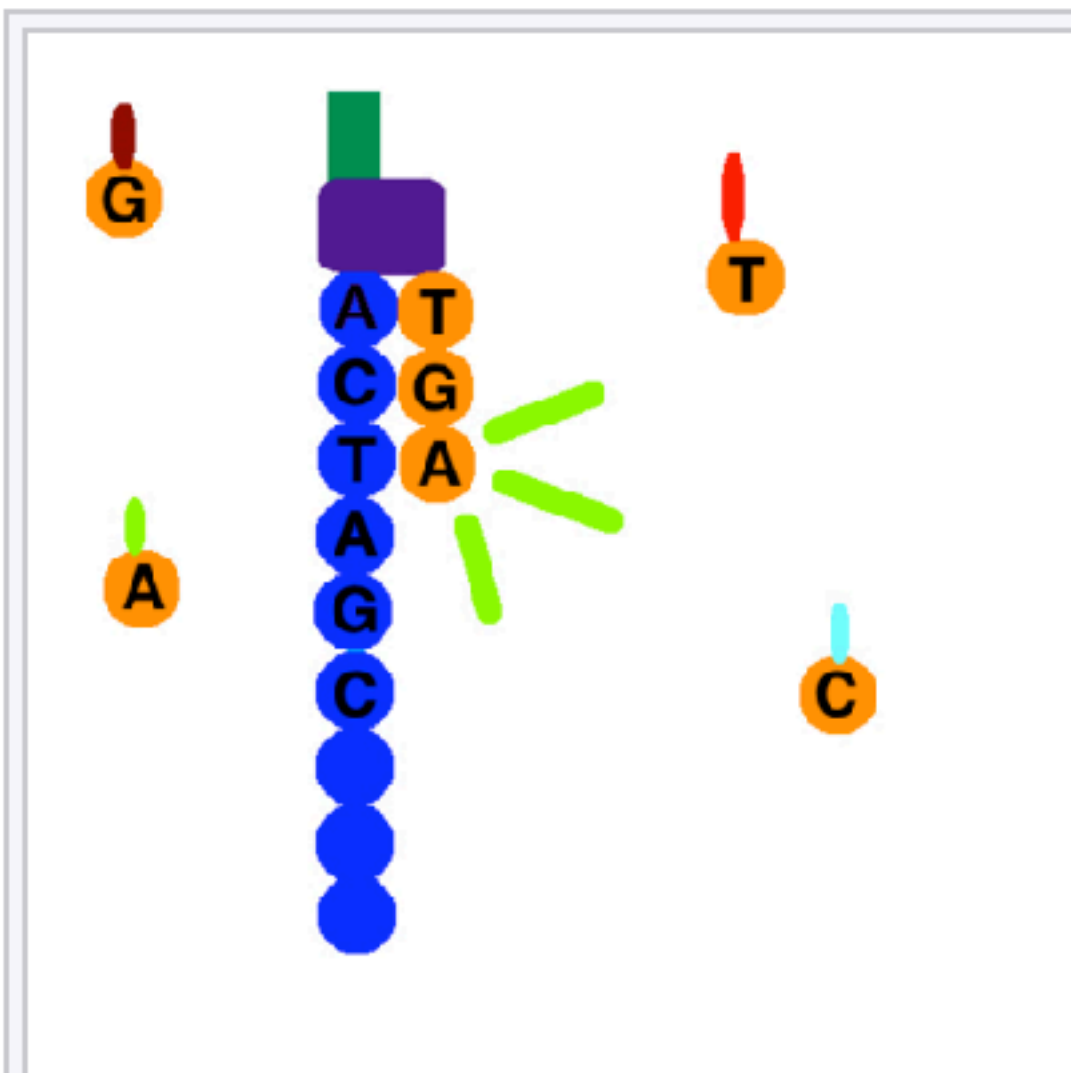
Millions of oligos line the bottom of each flow cell lane.

The strand bends over and attaches to a second oligo forming a bridge. A polymerase synthesizes the reverse strand. The two strands release and straighten. Each forms a new bridge (bridge amplification). The result is a cluster of DNA forward and reverse strand clones.

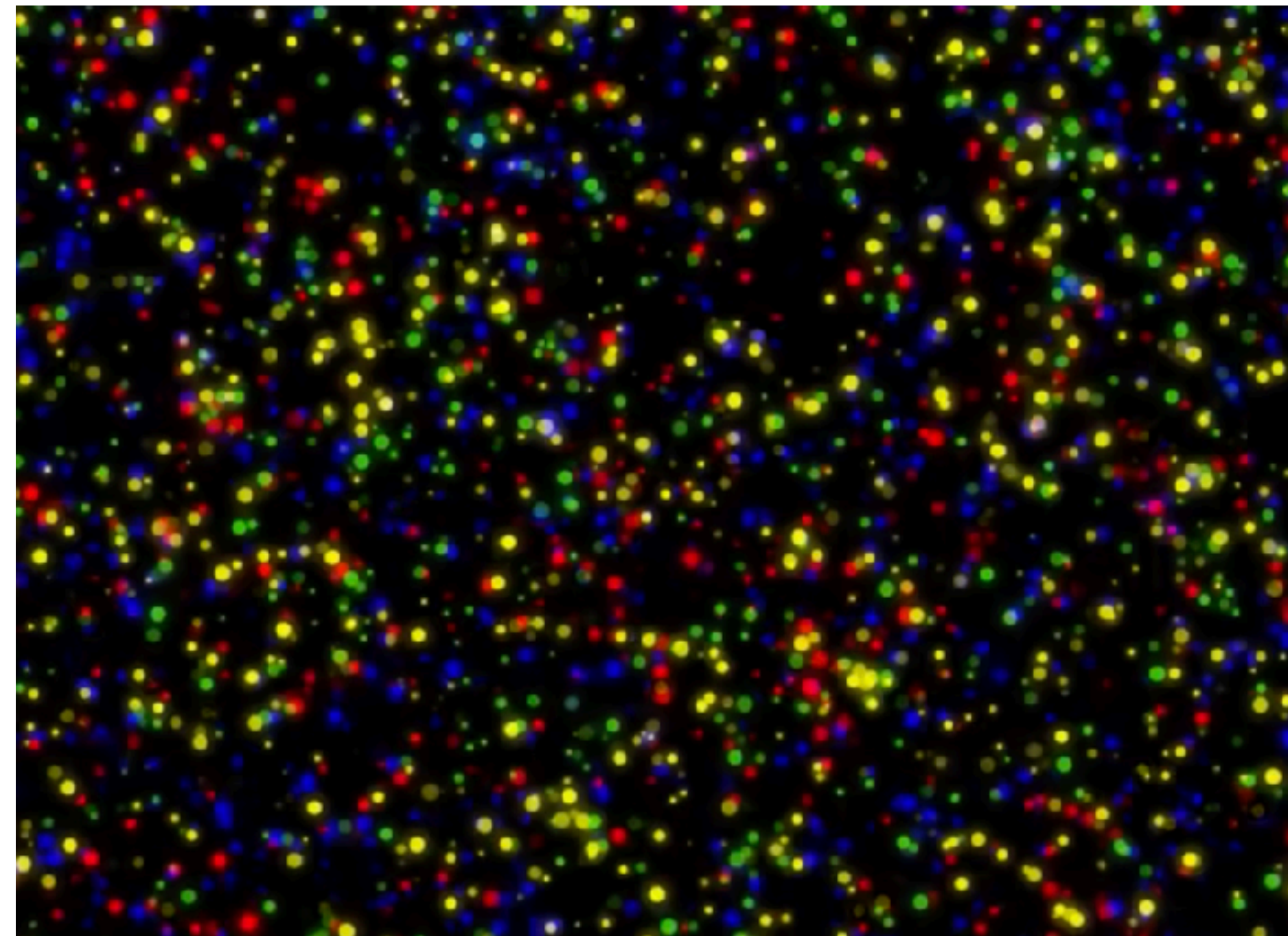
Next-generation DNA sequencing

Illumina continued:

3) Sequencing



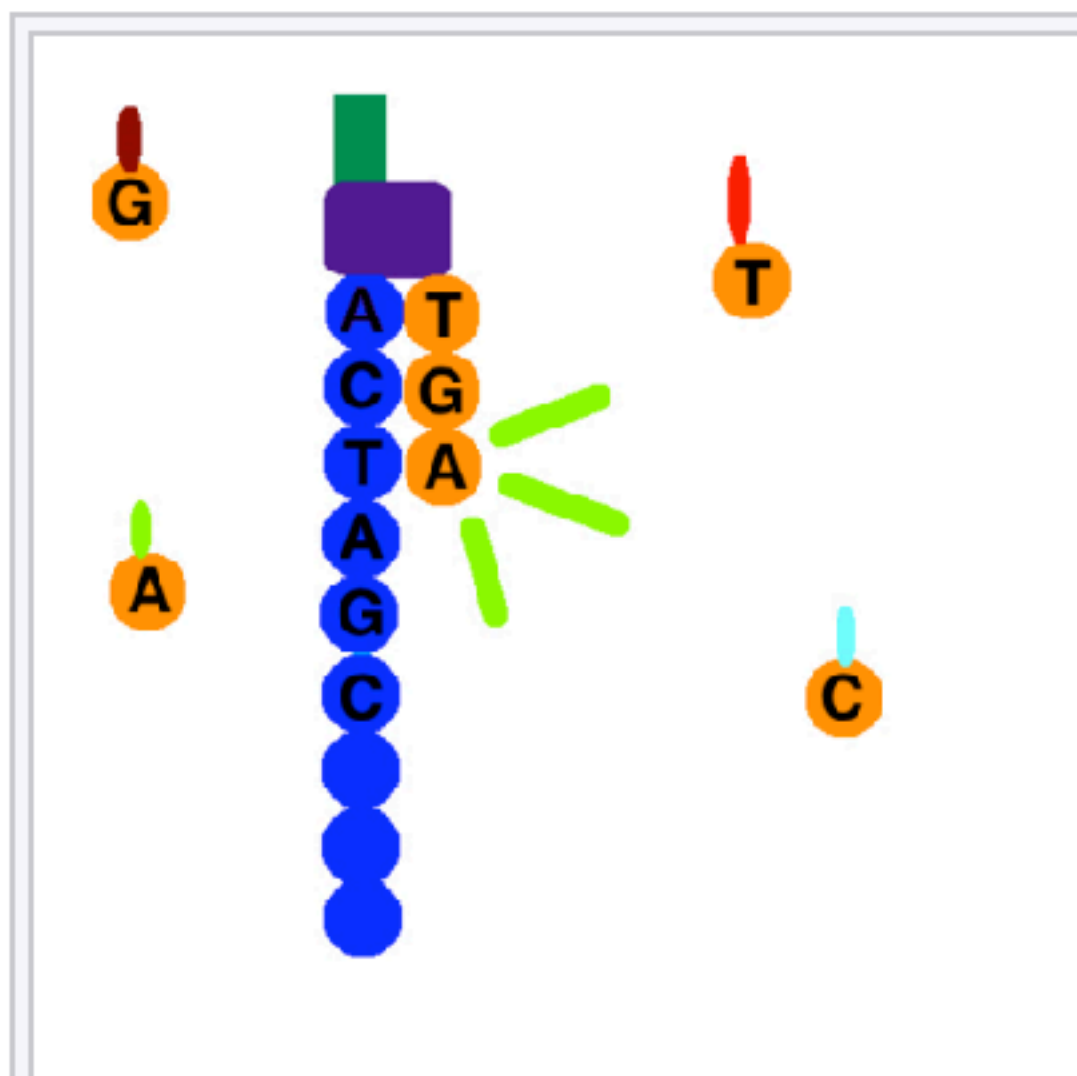
Tagged nucleotides are added in order to the DNA strand. Each of the four nucleotides have an identifying label that can be excited to emit a characteristic wavelength. A computer records all of the emissions, and from this data, base calls are made.



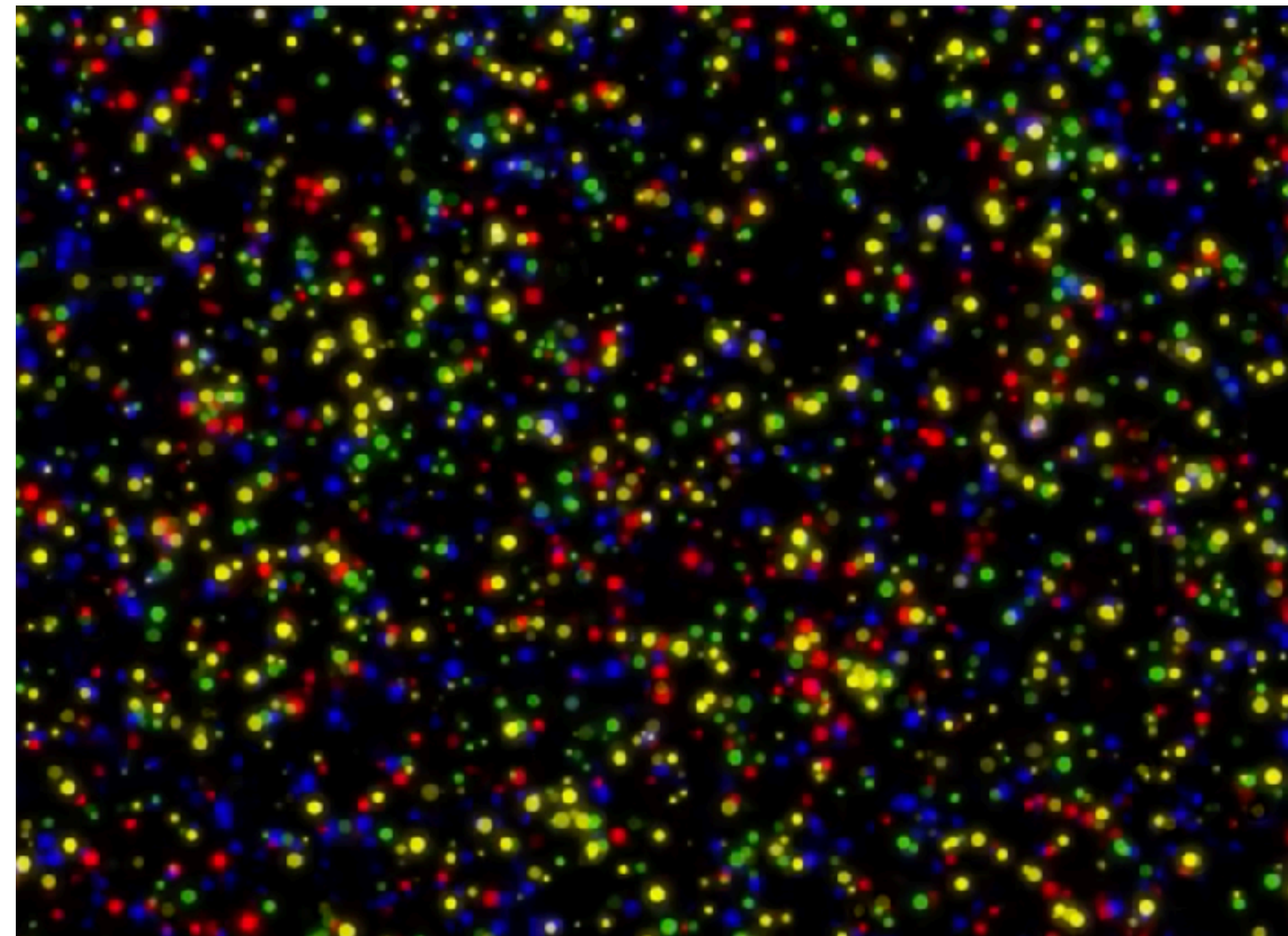
Next-generation DNA sequencing

Illumina continued:

3) Sequencing



Tagged nucleotides are added in order to the DNA strand. Each of the four nucleotides have an identifying label that can be excited to emit a characteristic wavelength. A computer records all of the emissions, and from this data, base calls are made.



3) Data analysis

Novaseq X+ can produce 52 billion reads per run @ 150bp reads; 16TB of data

The first human genome project took about 4 years, cost \$2.7 billion, and generated about 23 Gb of data (~7X coverage).

costs ~\$20k;
\$200 per genome!

We could today do the first HGP 200 times on one Novaseq X+ run.

Next-generation DNA sequencing

Illumina continued:

3) Data analysis

This process gives us a whole giant pile of reads. What next?

Identifying individuals

Mapping to a reference genome (or creating a reference genome denovo)

Calling genotypes given mapped reads

...all to just understand where there is genetic variation in the genome

Next-generation DNA sequencing

Illumina continued:

3) Data analysis

This process gives us a whole giant pile of reads. What next?

Identifying individuals

Mapping to a reference genome (or creating a reference genome denovo)

Calling genotypes given mapped reads

...all to just understand where there is genetic variation in the genome

We will dig into this process this week / next week

Next-generation DNA sequencing

Approaches other than Illumina:

Long read sequencing —

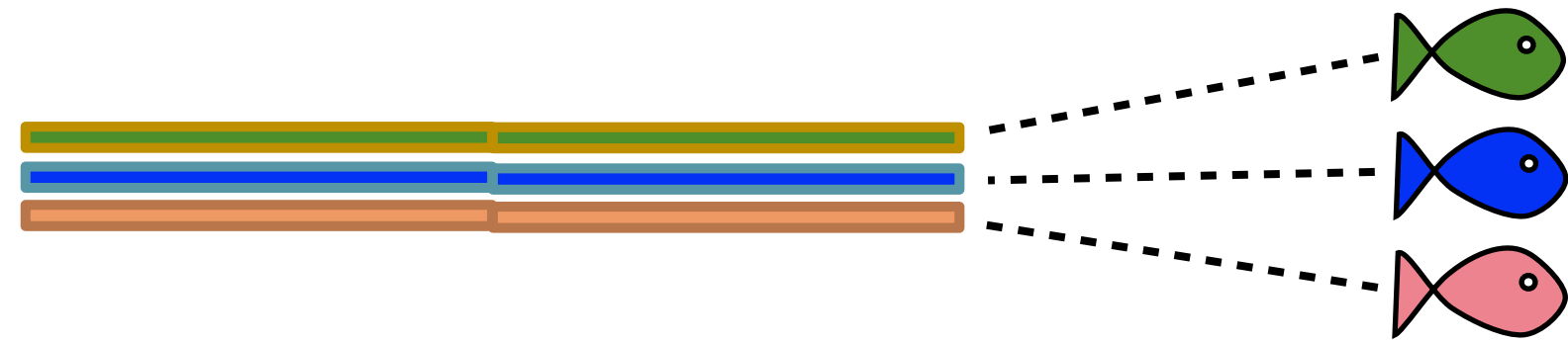
PacBio

Oxford Nanopore — portable!

Great choices for genome assembly and for detection of structural variants

Higher error rates (10-15% versus 0.1-0.5% for Illumina)

Next-generation DNA sequencing

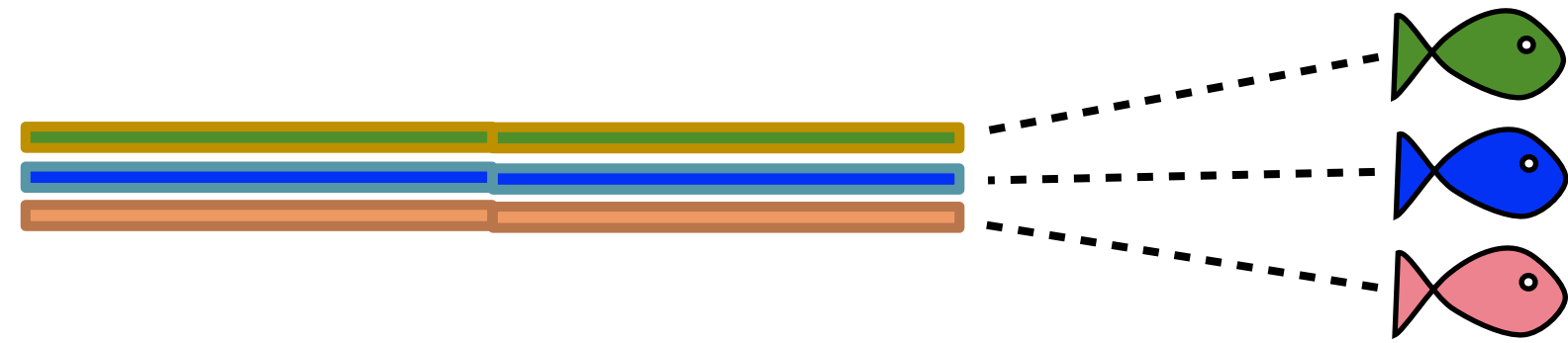


Calling genotypes for diploid organisms


Regardless of sequencing platform, going from data to genotypes is similar:

- 1) Determine reference genome to map to, or create a denovo assembly for your data
- 2) Map reads
- 3) Call genotypes

Next-generation DNA sequencing



Calling genotypes for diploid organisms

	A	G	T	C	A	A	A	G	G	G	A	A	A	G	G	A	A	G	A
	A	G	T	C	T	A	A	G	G	G	A	A	A	G	G	A	T	G	A
	A	G	T	C	T	A	A	G	G	C	A	A	A	G	G	A	A	G	A
	A	G	T	C	A	A	A	G	G	G	A	A	A	G	G	A	A	G	A
	A	G	T	C	T/A	A	A	G	G	G/C	A	A	A	G	G	A	A/T	G	A

← Called
genotype

Next-generation DNA sequencing

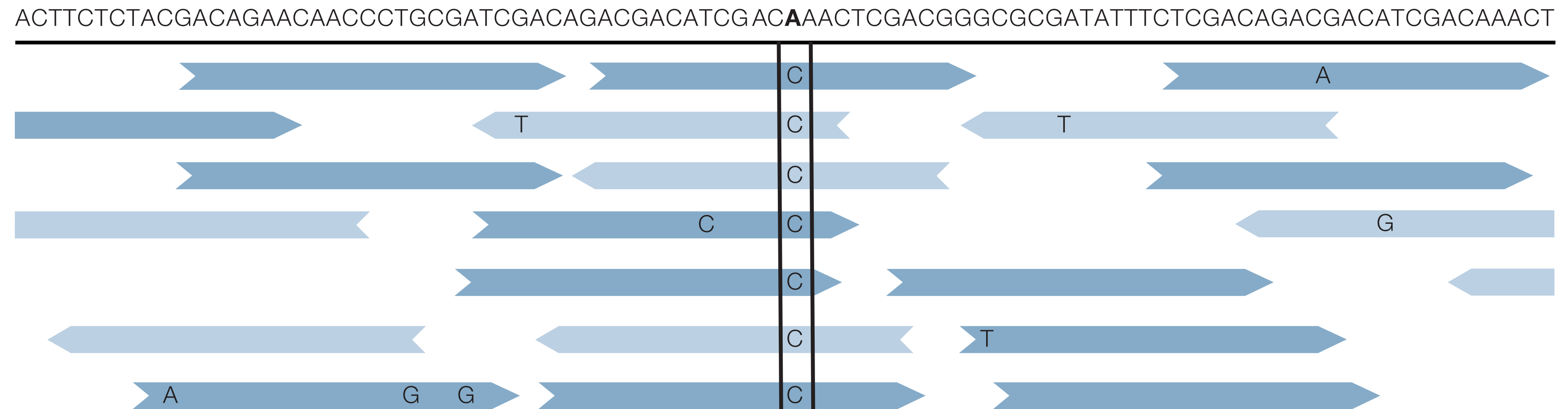


Figure 2.4

Next-generation DNA sequencing

Illumina:

The sequencing error rate is not zero! (but it's better for Illumina than other platforms)

Our confidence in any particular sequence variants depends on the sequencing effort — which depends on the flow cell size you choose, and the genome size of the target species (or size of the target portion of the genome), and the number of individuals

Next-generation DNA sequencing

“The random sequencing of the input DNA (“shotgun sequencing”) results in a wide distribution of read depths for each position in the genome (FIGURE 2.3). The read depth at each position—in other words, the number of sequence reads originating from a specific portion of the genome—is expected to be Poisson-distributed under simplifying assumptions (Lander and Waterman 1988).” Hahn, pg 33

Next-generation DNA sequencing

“The random sequencing of the input DNA (“shotgun sequencing”) results in a wide distribution of read depths for each position in the genome (FIGURE 2.3). The read depth at each position—in other words, the number of sequence reads originating from a specific portion of the genome—is expected to be Poisson-distributed under simplifying assumptions (Lander and Waterman 1988).” Hahn, pg 33

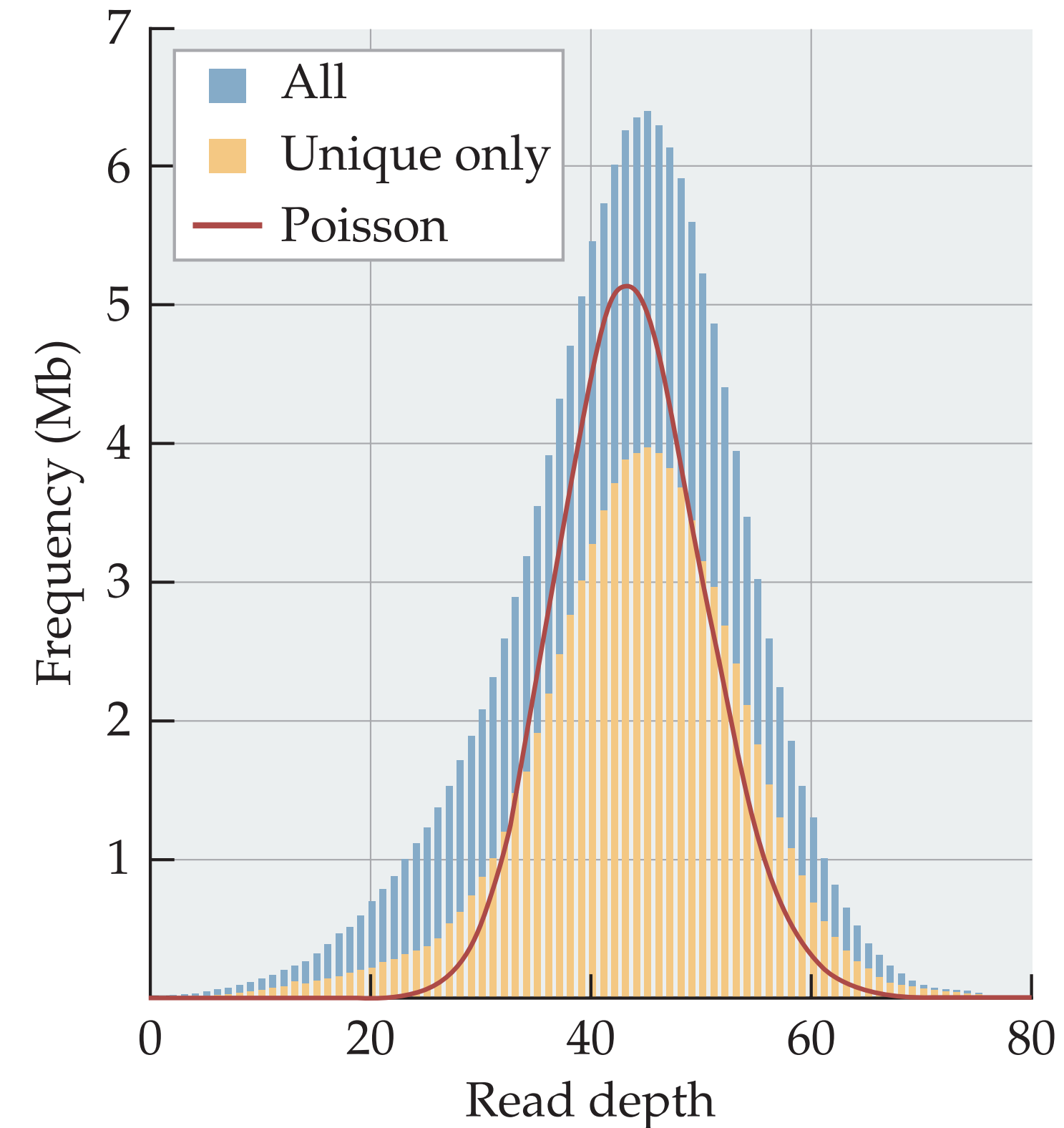


FIGURE 2.3 Read depth for the human X chromosome using Illumina sequencing technology. The read depth sampled at every 50th position along the chromosome is plotted for uniquely mapped reads and all mapped reads. Average read depth across the chromosome was 40.6X, meaning that most 50-bp windows were covered by 40.6 sequence reads. The Poisson expectation for a distribution with the same mean as the uniquely mapping reads is also plotted. (After Bentley et al. 2008.)

Next-generation DNA sequencing

Reads are max 150 bp — short! This is hard for genome assembly and limits things like structural variant detection

Sometimes we don't want the whole genome! Or sometimes it's still too expensive.
What then?

Next-generation DNA sequencing

“Reduced representation” sequencing methods — began to emerge ~2009-2010

Utilize next-generation sequencing to sequence a reduced portion of the genome — thousands to hundreds of thousands of sites

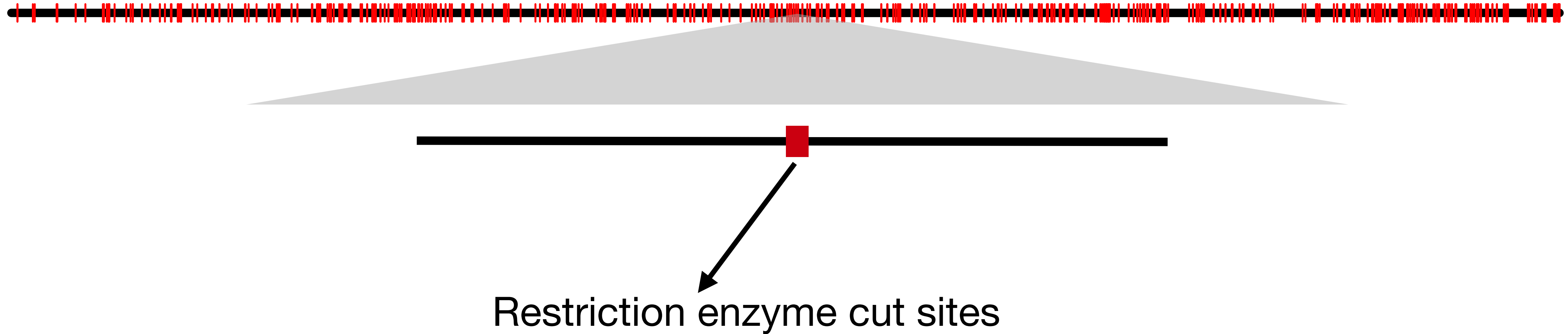
GBS - Genotyping by Sequencing

RADseq - Restriction digest associated sequencing

ddRADseq - double digest RADseq

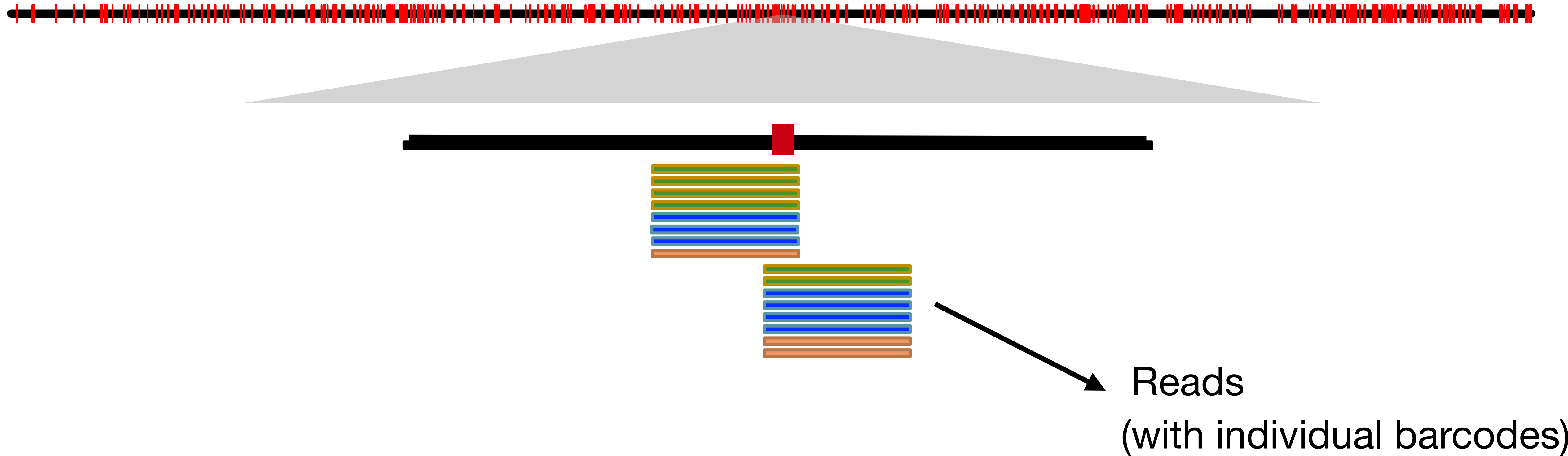
Next-generation DNA sequencing

“Reduced representation” sequencing methods



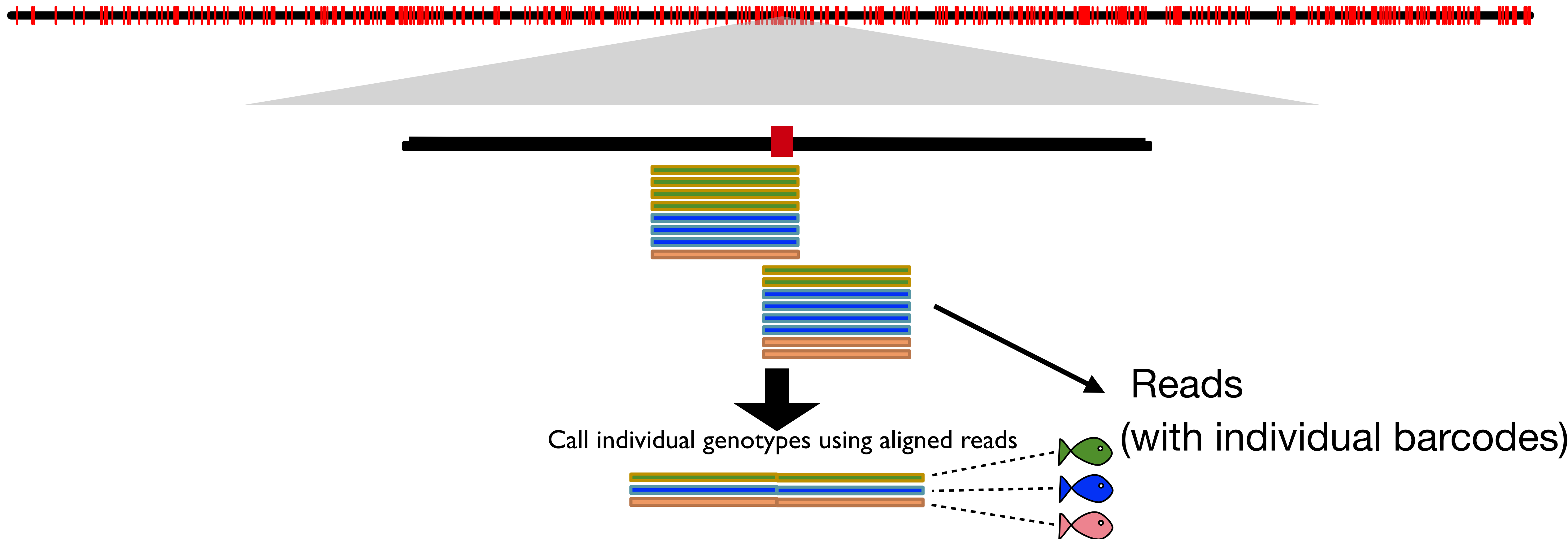
Next-generation DNA sequencing

“Reduced representation” sequencing methods

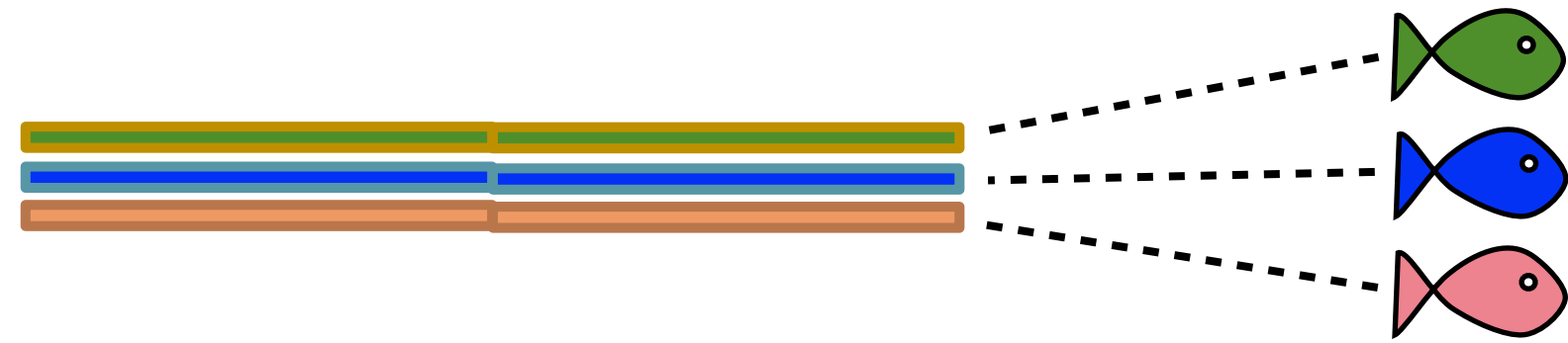


Next-generation DNA sequencing

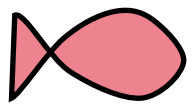
“Reduced representation” sequencing methods



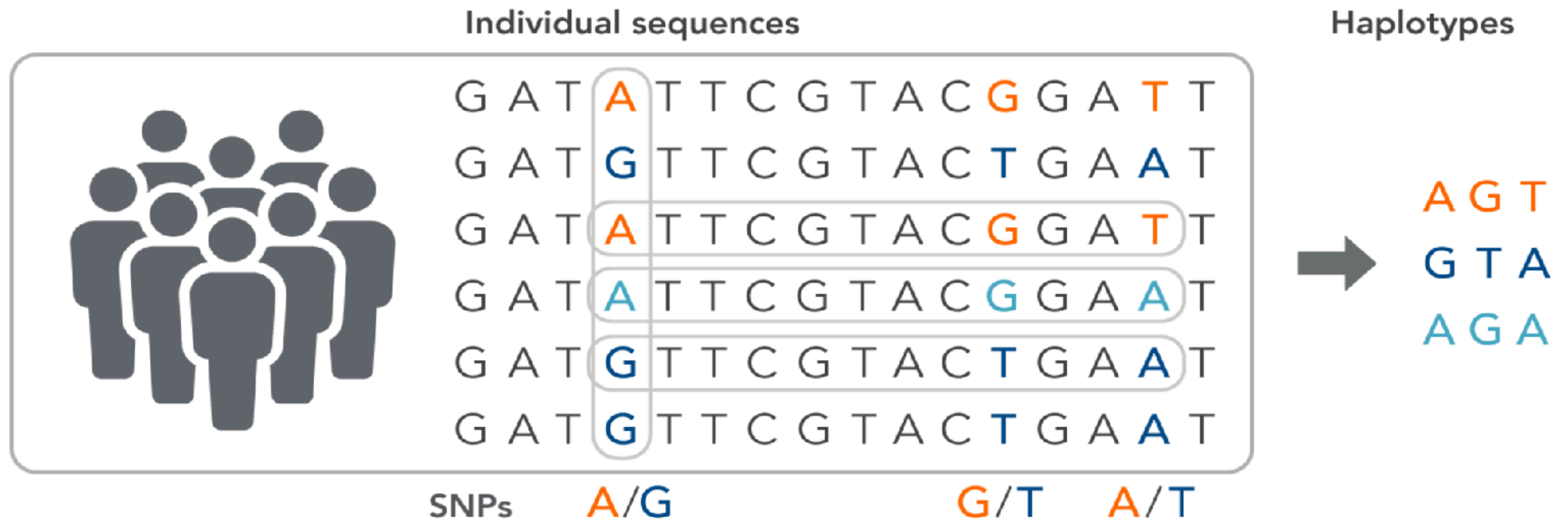
Next-generation DNA sequencing

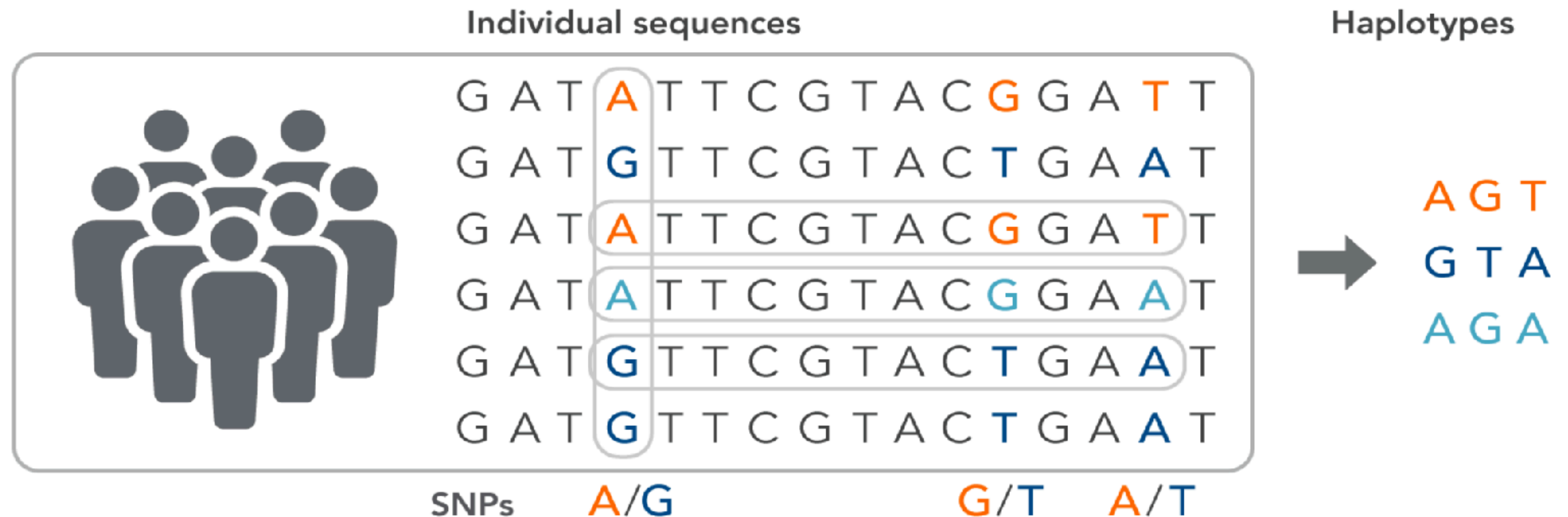


Calling genotypes for diploid organisms

	A	G	T	C	A	A	A	G	G	G	A	A	A	G	G	A	A	G	A
	A	G	T	C	T	A	A	G	G	G	A	A	A	G	G	A	T	G	A
	A	G	T	C	T	A	A	G	G	C	A	A	A	G	G	A	A	G	A
	A	G	T	C	A	A	A	G	G	G	A	A	A	G	G	A	A	G	A
	A	G	T	C	T/A	A	A	G	G	G/C	A	A	A	G	G	A	A/T	G	A

← Called
genotype





Phasing is knowing which variants go together on the same haplotype

Next-generation DNA sequencing

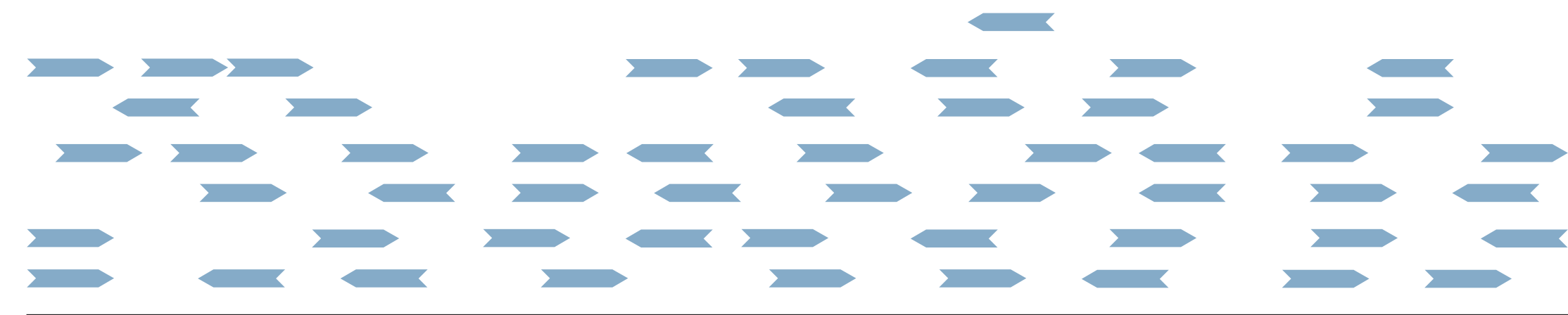
Other forms of “reduced representation” sequencing:

Sequence capture approaches (RAPTURE, ultraconserved elements, exome capture, Hyb-Seq): target particular regions of the genome for sequencing by making use of “baits” (known DNA sequences pulled out using a target sequence) to focus sequencing effort and coverage on pre-selected regions.

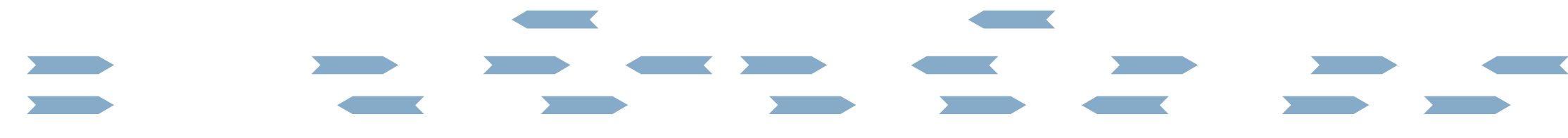
RNA-seq:

captures sequence data from expressed genes. It can be used to generate whole transcriptome sequences, which can be used as a reduced representation genomic sequencing method

(A) High-coverage whole-genome sequencing



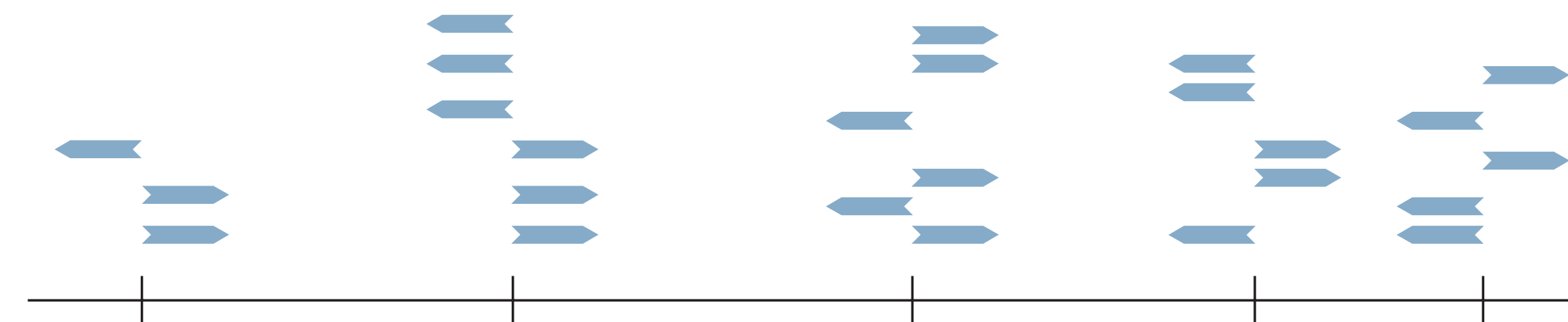
(B) Low-coverage whole-genome sequencing



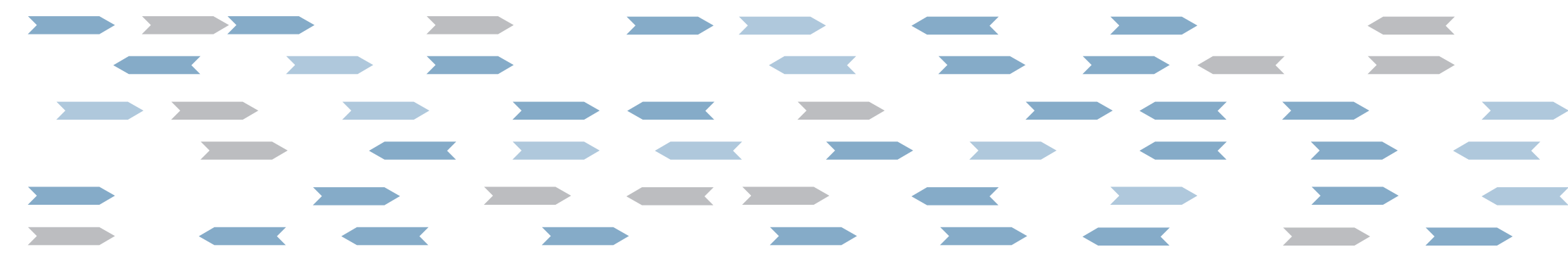
(C) Exome sequencing



(D) RAD sequencing



(E) Pooled sequencing



SNP genotyping

Some methods allow lower-cost genotyping of specific SNPs with known alternate genotypes

Snip-SNP, PCR-RFLP, or cleavable amplified polymorphic sequences (CAPS), GT-seq — usually 10's - 100's of SNPs

SNP-chips — 500 - 900,000 SNPs

Ascertainment bias needs to be carefully considered

Next-generation DNA sequencing

Hahn page 33: “The random sequencing of the input DNA (“shotgun sequencing”) results in a wide distribution of read depths for each position in the genome (FIGURE 2.3). The read depth at each position—in other words, the number of sequence reads originating from a specific portion of the genome—is expected to be Poisson-distributed under simplifying assumptions (Lander and Waterman 1988).”

.Poisson Distribution Properties:

- The Poisson distribution describes the probability of a given number of events (e.g., reads) occurring in a fixed interval (e.g., a specific genomic position), assuming:
 1. Events occur independently.
 2. The average rate (mean coverage depth) is constant across the region.
 3. Two events cannot occur at exactly the same place (for reads, this maps to uniform distribution).