Basic Methods in Cellular and Molecular Biology

(Sequencing)

Both DNA and RNA Can Be Rapidly Sequenced

- Most current methods of manipulating DNA, RNA, and proteins rely on prior knowledge of the nucleotide sequence of the genome of interest.
- But how were these sequences determined in the first place? And how are new DNA and RNA molecules sequenced today?

- In the late 1970s, researchers developed several strategies for determining, simply and quickly, the nucleotide sequence of any purified DNA fragment.

- The one that became the most widely used is called **dideoxy sequencing or Sanger sequencing**.

- This method was used to determine the nucleotide sequence of many genomes, including those of E. coli, fruit flies, nematode worms, mice, and humans.

- Today, cheaper and faster methods are routinely used to sequence DNA, and even more efficient strategies are being developed.

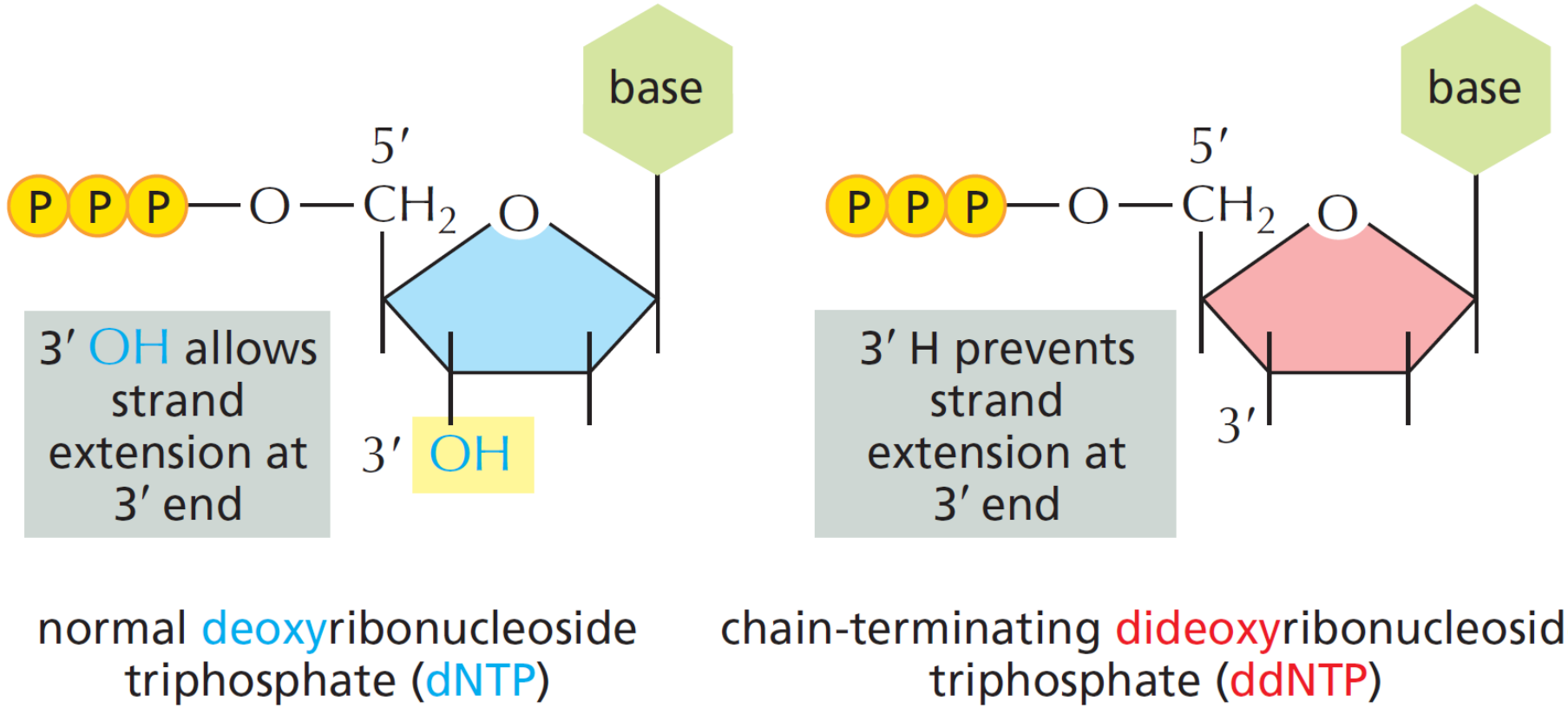
- The original “reference” sequence of the human genome, completed in 2003, cost over $1 billion and required many scientists from around the world working together for 13 years.

- The enormous progress made in the past decade makes it possible for a single person to complete the sequence of an individual human genome in less than a day.

- Dideoxy sequencing, or Sanger sequencing (named after the scientist who invented it), uses DNA polymerase, along with special chain-terminating nucleotides called **dideoxyribonucleoside triphosphates**, to make partial copies of the DNA fragment to be sequenced.

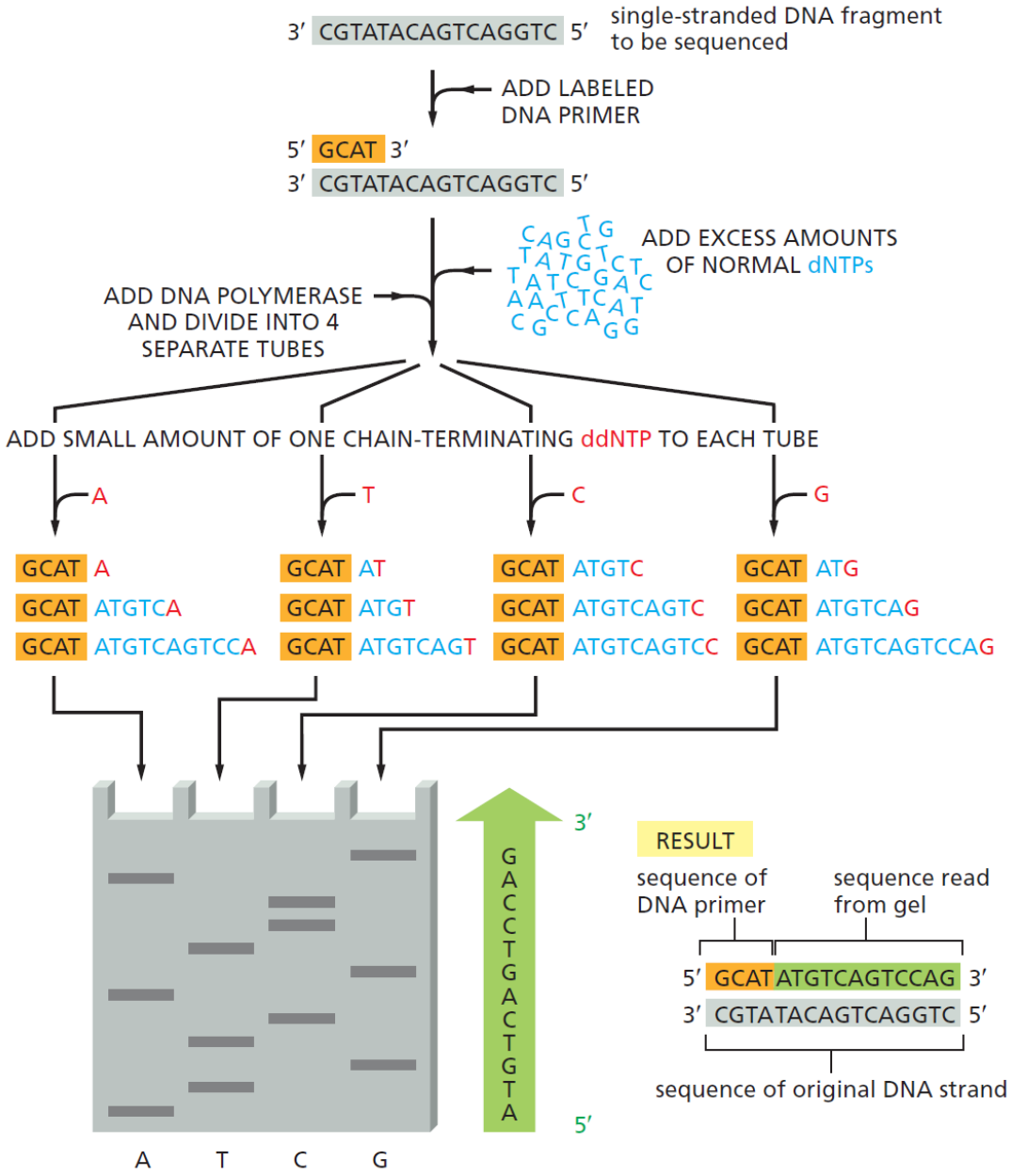
- These ddNTPs are derivatives of the normal deoxyribonucleoside triphosphates that lack the **3′ hydroxyl group**.

- When incorporated into a growing DNA strand, they block further elongation of that strand.



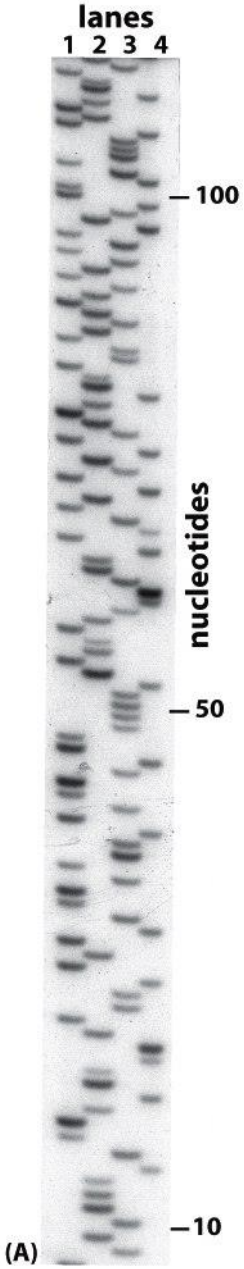
MANUAL DIDEOXY SEQUENCING

- To determine the complete sequence of a single-stranded fragment of DNA (gray), the DNA is first hybridized with a short DNA primer (orange) that is labeled with a fluorescent dye or radioisotope.
- DNA polymerase and an excess of all four normal deoxyribonucleoside triphosphates (blue A, C, G, or T) are added to the primed DNA, which is then divided into four reaction tubes.
- Each of these tubes receives a **small amount of a single chain-terminating dideoxyribonucleoside triphosphate** (red A, C, G, or T).
- Because these will be incorporated only occasionally, each reaction produces a set of DNA copies that terminate at different points in the sequence.
- The products of these four reactions are separated by electrophoresis in four parallel lanes of a polyacrylamide gel (labeled here A, T, C, and G).
- **In each lane, the bands represent fragments that have terminated at a given nucleotide but at different positions in the DNA.**
- By reading off the bands in order, starting at the bottom of the gel and reading across all lanes, the DNA sequence of the newly synthesized strand can be determined.
- The sequence, which is given in the green arrow to the right of the gel, is complementary to the sequence of the original gray single-stranded DNA.

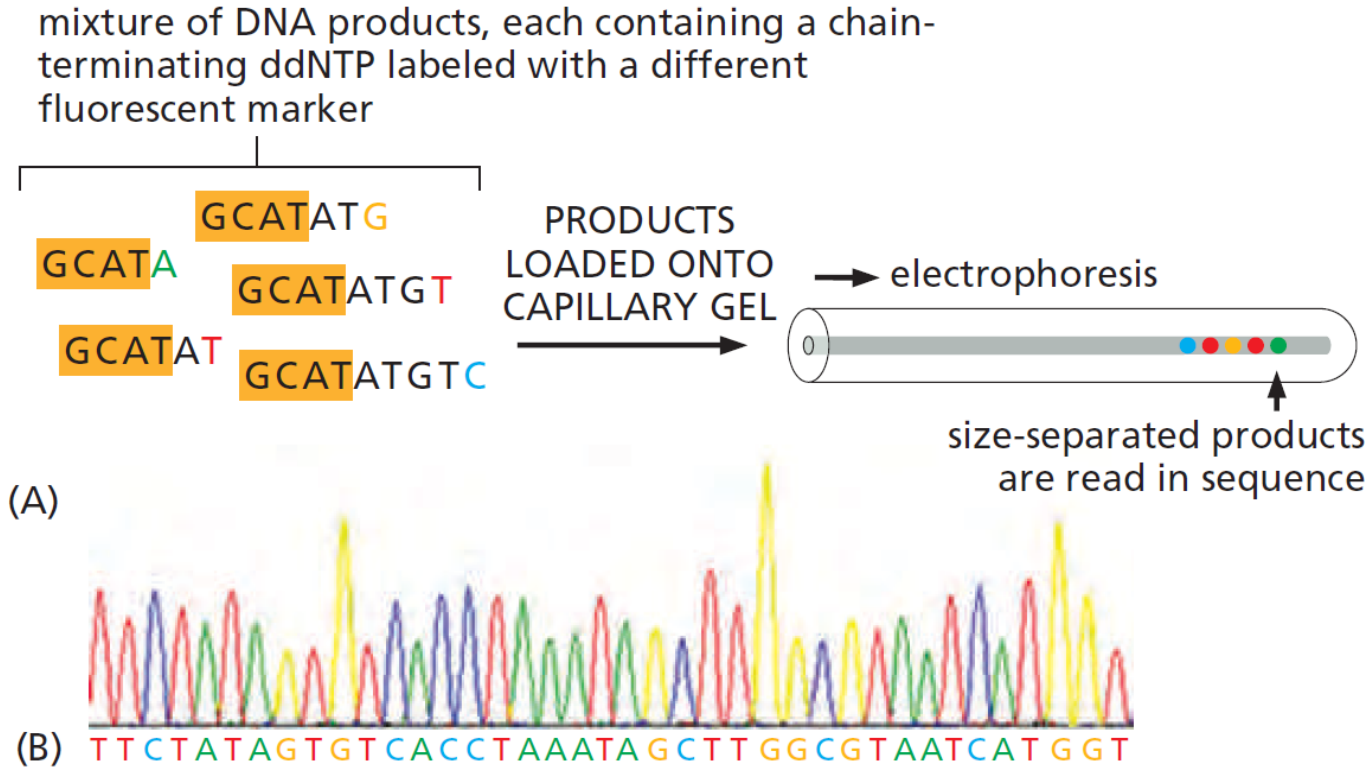


MANUAL DIDEOXY SEQUENCING

- To determine the complete sequence of a single-stranded fragment of DNA (gray), the DNA is first hybridized with a short DNA primer (orange) that is labeled with a fluorescent dye or radioisotope.
- DNA polymerase and an excess of all four normal deoxyribonucleoside triphosphates (blue A, C, G, or T) are added to the primed DNA, which is then divided into four reaction tubes.
- Each of these tubes receives a **small amount of a single chain-terminating dideoxyribonucleoside triphosphate** (red A, C, G, or T).
- Because these will be incorporated only occasionally, each reaction produces a set of DNA copies that terminate at different points in the sequence.
- The products of these four reactions are separated by electrophoresis in four parallel lanes of a polyacrylamide gel (labeled here A, T, C, and G).
- **In each lane, the bands represent fragments that have terminated at a given nucleotide but at different positions in the DNA.**
- By reading off the bands in order, starting at the bottom of the gel and reading across all lanes, the DNA sequence of the newly synthesized strand can be determined.
- The sequence, which is given in the green arrow to the right of the gel, is complementary to the sequence of the original gray single-stranded DNA.



AUTOMATED DIDEOXY SEQUENCING



Fully automated machines can run dideoxy sequencing reactions.

(A) The automated method uses an excess amount of normal dNTPs plus a mixture of four different chain-terminating ddNTPs, each of which is labeled with a fluorescent tag of a different color. The reaction products are loaded onto a **long**, **thin capillary gel** and separated by **electrophoresis**. A camera (not shown) reads the color of each band as it moves through the gel and feeds the data to a computer that assembles the sequence.

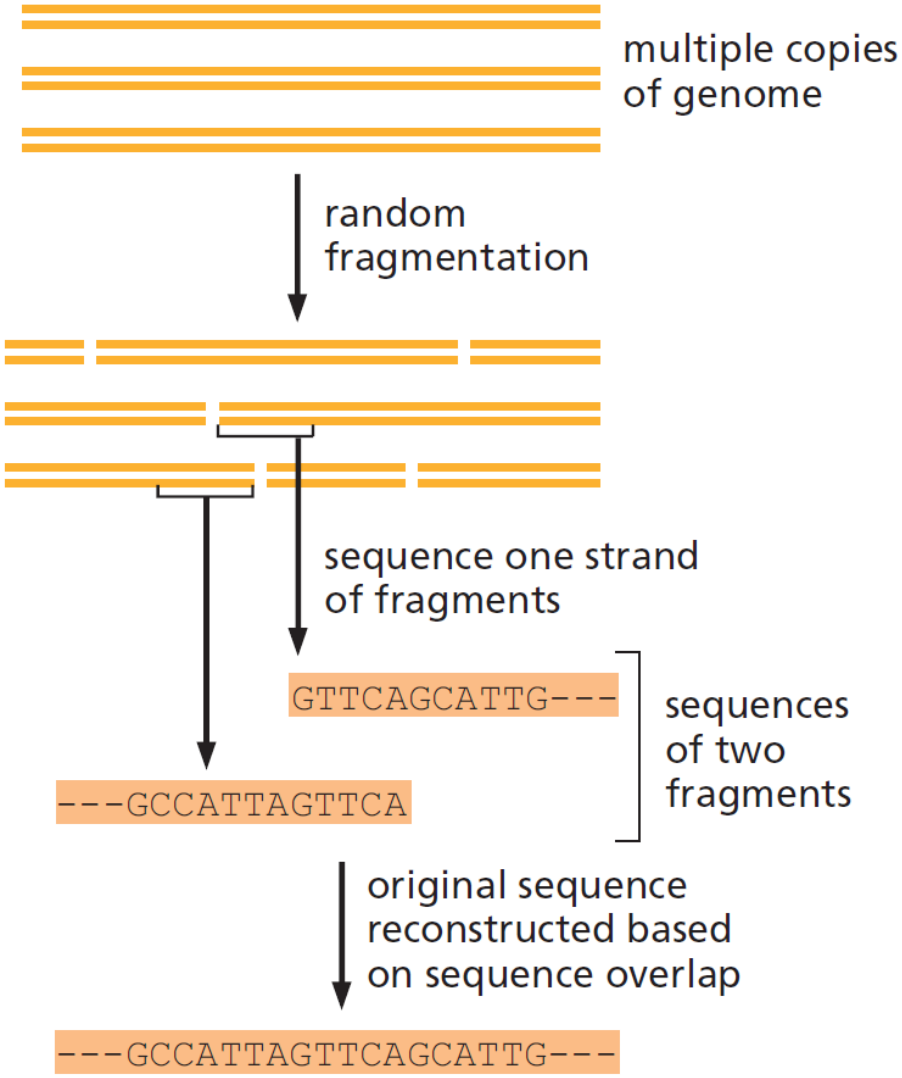
(B) A tiny part of the data from such an automated sequencing run. Each colored peak represents a nucleotide in the DNA sequence.

SEQUENCING WHOLE GENOMES

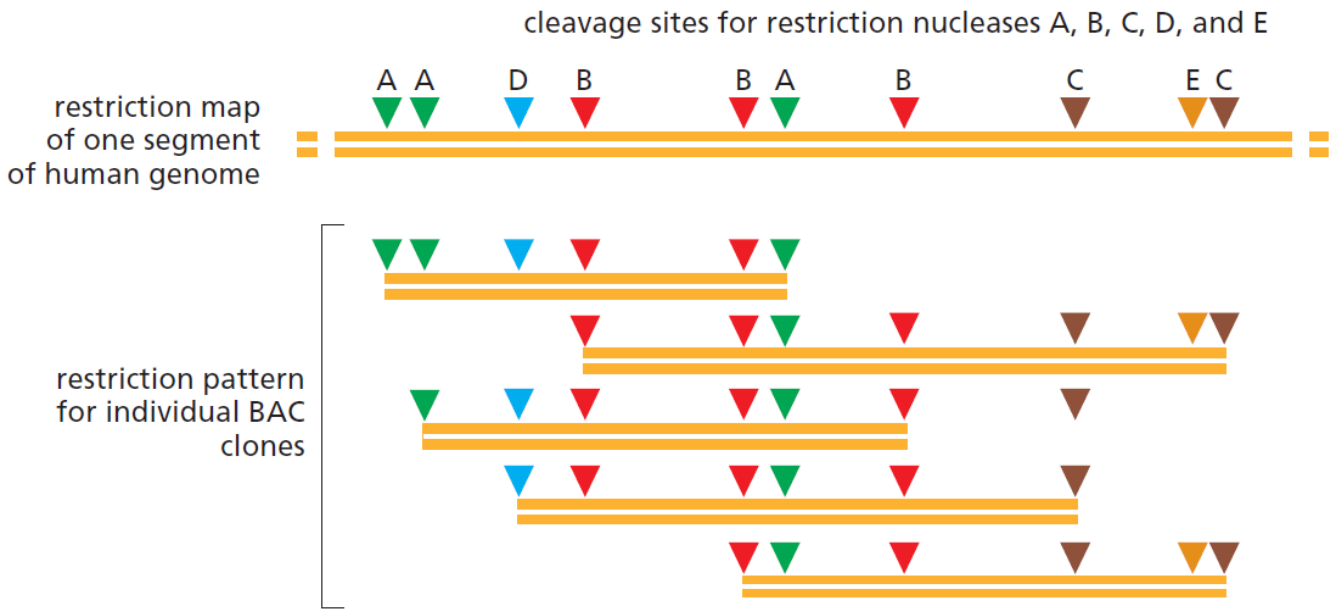
- **Shotgun sequencing**: To determine the nucleotide sequence of a whole genome, the genomic DNA is first fragmented into small pieces and a **genomic library** is constructed, typically using plasmids and bacteria.

- In shotgun sequencing, the nucleotide sequence of tens of thousands of individual clones is determined;
- the full genome sequence is then **reconstructed** by stitching together (in silico) the nucleotide sequence of each clone, using the overlaps between clones as a guide.

- The shotgun method works well for small genomes (such as those of viruses and bacteria) that lack repetitive DNA.



- **BAC clones**: Most plant and animal genomes are large (often over 10^9 nucleotide pairs) and contain extensive amounts of repetitive DNA spread throughout the genome.
- Because a nucleotide sequence of a fragment of repetitive DNA will "**overlap**" every instance of the repeated DNA, it is difficult, if not impossible, to assemble the fragments into a unique order solely by the shotgun method.
- To circumvent this problem, the human genome was first broken down into very large DNA fragments (each approximately 100,000 nucleotide pairs) and cloned into **BACs (Bactarial Artificial Chromosome)**.
- The order of the BACs along a chromosome was determined by comparing the pattern of restriction enzyme cleavage sites in a given BAC clone with that of the whole genome.
- In this way, a given BAC clone can be mapped, say, to the left arm of human Chromosome 3.
- Once a collection of BAC clones was obtained that spanned the entire genome, each individual BAC was sequenced by the shotgun method.
- At the end, the sequences of all the BAC inserts were stitched together using the knowledge of the position of each BAC insert in the human genome.
- In all, approximately 30,000 BAC clones were sequenced to complete the human genome.



Thousands of genomes from individual humans have now been sequenced and it is not necessary to painstakingly reconstruct the order of DNA sequence "reads" each time;
they are simply assembled using the order determined from the original human genome sequencing project.
For this reason, **resequencing**, the term applied when the genome of a species is sequenced again (even though it may be from a different individual), is far easier than the **original sequencing**.

SECOND-GENERATION SEQUENCING TECHNOLOGIES

- The dideoxy method made it possible to sequence the genomes of humans and most of the other organisms.
- But newer methods, developed since 2005, have made genome sequencing even more rapid—and very much cheaper.
- With these so-called **second-generation sequencing** methods, the cost of sequencing DNA has decreased dramatically.
- Not surprisingly, the number of genomes that have been sequenced has increased enormously.
- These rapid methods allow multiple genomes to be sequenced in parallel in a matter of weeks, enabling investigators to examine thousands of individual human genomes, **catalog the variation in nucleotide sequences from people around the world**, and **uncover the mutations that increase the risk of various diseases**, from cancer to autism.
- These methods have also made it possible to determine the genome sequence of extinct species, including Neanderthal man and the wooly mammoth.
- By sequencing genomes from many closely **related species**, they have also made it possible to understand the molecular basis of key evolutionary events in the tree of life, such as the "inventions" of multicellularity, vision, and language.
- The ability to rapidly sequence DNA has had major impacts on all branches of biology and medicine; it is almost impossible to imagine where we would be without it.

ILLUMINA® SEQUENCING

- Several second-generation sequencing methods are now in wide use.

- Two of the most common rely on the construction of libraries of DNA fragments that represent—in toto—the DNA of the genome.

- Instead of using bacterial cells to generate these libraries, they are made using PCR amplification of billions of DNA fragments, each attached to a solid support.

- The amplification is carried out so that the PCR-generated copies, instead of floating away in solution, remain bound in proximity to the original DNA fragment.

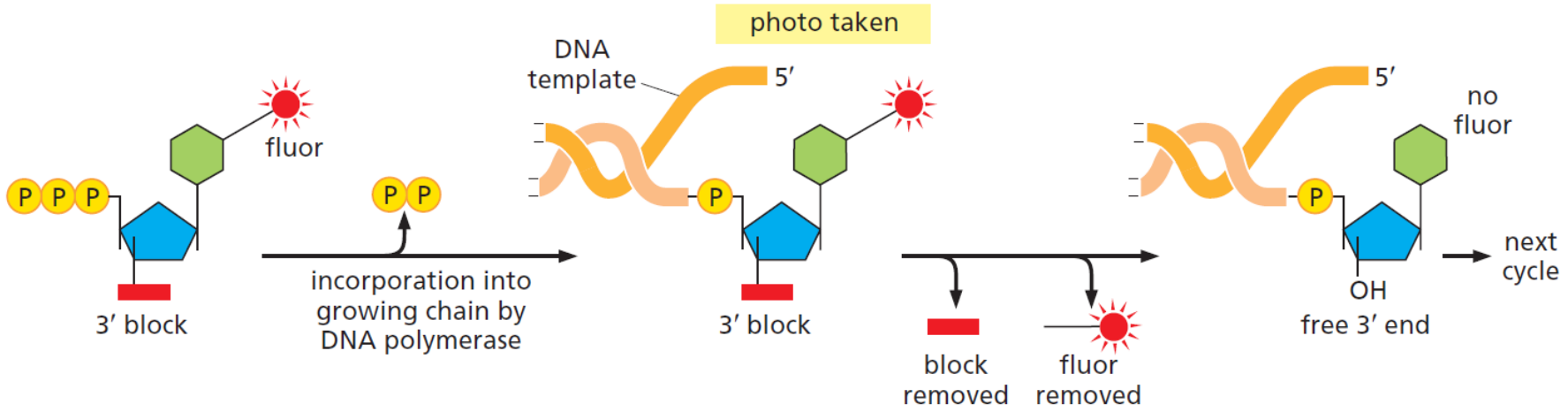
- This process generates **clusters of DNA fragments**, where each cluster contains about 1000 identical copies of a small bit of the genome.
- These clusters—a billion of which can fit in a single slide or plate—are then sequenced at the same time; that is, in parallel.

- One method, known as **Illumina sequencing**, is based on the dideoxy method, but it incorporates several innovations:

-Here, each nucleotide is attached to a **removable fluorescent molecule** (a different color for each of the four bases) as well as a **special chain-terminating chemical adduct**: instead of a 3'-OH group, as in conventional dideoxy sequencing, the nucleotides carry a chemical group that blocks elongation by DNA polymerase but which can be removed chemically.

ILLUMINA® SEQUENCING

-Sequencing is then carried out as follows: the four fluorescently labeled nucleotides along with DNA polymerase are added to billions of DNA clusters immobilized on a slide.
-Only the appropriate nucleotide (that is complementary to the next nucleotide in the template) is covalently incorporated at each cluster; the unincorporated nucleotides are washed away, and a **high-resolution digital camera** takes an image that registers which of the four nucleotides was added to the chain at each cluster.
-The fluorescent label and the 3′-OH blocking group are then removed enzymatically, washed away, and the process is repeated many times.
-In this way, billions of sequencing reactions are carried out simultaneously.
-By keeping track of the color changes occuring at each cluster, the DNA sequence represented by each spot can be read.
-Although each individual sequence read is relatively short (approximately 200 nucleotides), the billions that are carried out simultaneously can produce several human genomes worth of sequence in about a day.



Principle behind Illumina sequencing. This reaction is carried out stepwise, on billions of DNA clusters at once. The method relies on a color digital camera that rapidly scans all the DNA clusters after each round of modified nucleotide incorporation. The DNA sequence of each cluster is then determined by the sequence of color changes it undergoes as the elongation reaction proceeds stepwise.
Each round of modified nucleotide incorporation, image acquisition, and removal of the 3′ block and the fluorescent group takes less than an hour.
Each cluster on the slide contains many copies of different, random bits of a genome; in preparing the clusters, a DNA sequence (specified by the experimenter) is joined to each copy in every cluster, and a primer complementary to this sequence is used to begin the elongation reaction by DNA polymerase.

ION TORRENT™ SEQUENCING

Another widely used strategy for rapid DNA sequencing is called the **ion torrent** method:

- Here, a genome is fragmented, and the individual fragments are attached to **microscopic beads**.
- Using PCR, each DNA fragment is then amplified so that copies of it eventually **coat the bead** to which it was initially attached.
- This process produces a library of billions of individual beads, each covered with identical copies of a particular DNA fragment.
- Like eggs in a carton, the beads are placed into individual wells on an array that can hold a billion beads in a square inch.
- Beginning with a primer, DNA synthesis is then initiated on each bead.
- **A hydrogen ion (H+) is released (along with pyrophosphate)** each time a nucleotide is incorporated into a growing DNA chain, and the ion torrent method is based on this simple fact.
- Each of the four nucleotides is washed in, **one at a time**, over the array of beads; when a nucleotide is incorporated in the DNA of a given bead, the release of an H+ ion **changes the pH**, which is registered by a semiconductor chip placed beneath the array of wells. In this way, the DNA sequence on a given bead can be read from the pattern of pH changes observed as nucleotides are washed over them.
- Like a high-resolution sensor in a digital camera, the ion torrent semiconductor chip can register enormous amounts of information and can thus keep track of billions of parallel sequencing reactions.
- Using this technology it is currently possible, using a single chip, to determine the nucleotide sequences of several human genomes in just a few hours.