CHAPTER 20
DNA TECHNOLOGY AND GENOMICS

Section B: DNA Analysis and Genomics

1. Restriction fragment analysis detects DNA differences that affect restriction sites
2. Entire genomes can be mapped at the DNA level
3. Genomic sequences provide clues to important biological questions

Introduction

- Once we have prepared homogeneous samples of DNA, each containing a large number of identical segments, we can begin to ask some far-ranging questions.
- These include:
  - Are there differences in a gene in different people?
  - Where and when is a gene expressed?
  - What is the the location of a gene in the genome?
  - How has a gene evolved as revealed in interspecific comparisons?

- To answer these questions, we will eventually need to know the nucleotide sequence of the gene and ultimately the sequences of entire genomes.
- Comparisons among whole sets of genes and their interactions is the field of **genomics**.
- One indirect method of rapidly analyzing and comparing genomes is **gel electrophoresis**.
  - Gel electrophoresis separates macromolecules - nucleic acids or proteins - on the basis of their rate of movement through a gel in an electrical field.
    - Rate of movement depends on size, electrical charge, and other physical properties of the macromolecules.

- For linear DNA molecules, separation depends mainly on size (length of fragment) with longer fragments migrating less along the gel.

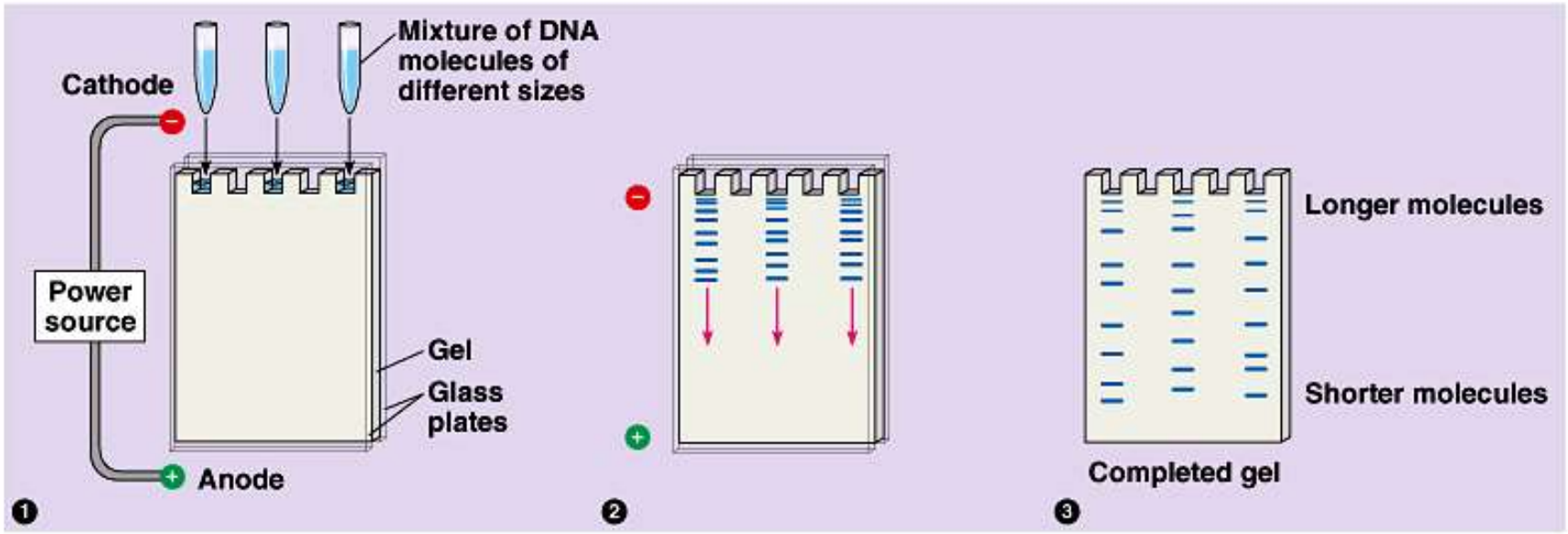

Fig. 20.8

1. Restriction fragment analysis detects DNA differences that affect restriction sites

- Restriction fragment analysis indirectly detects certain differences in DNA nucleotide sequences.
  - After treating long DNA molecules with a restriction enzyme, the fragments can be separated by size via gel electrophoresis.
  - This produces a series of bands that are characteristic of the starting molecule and that restriction enzyme.
  - The separated fragments can be recovered undamaged from gels, providing pure samples of individual fragments.

- We can use restriction fragment analysis to compare two different DNA molecules representing, for example, different alleles.
  - Because the two alleles must differ slightly in DNA sequence, they may differ in one or more restriction sites.
  - If they do differ in restriction sites, each will produce different-sized fragments when digested by the same restriction enzyme.
  - In gel electrophoresis, the restriction fragments from the two alleles will produce different band patterns, allowing us to distinguish the two alleles.

- Restriction fragment analysis is sensitive enough to distinguish between two alleles of a gene that differ by only base pair in a restriction site.

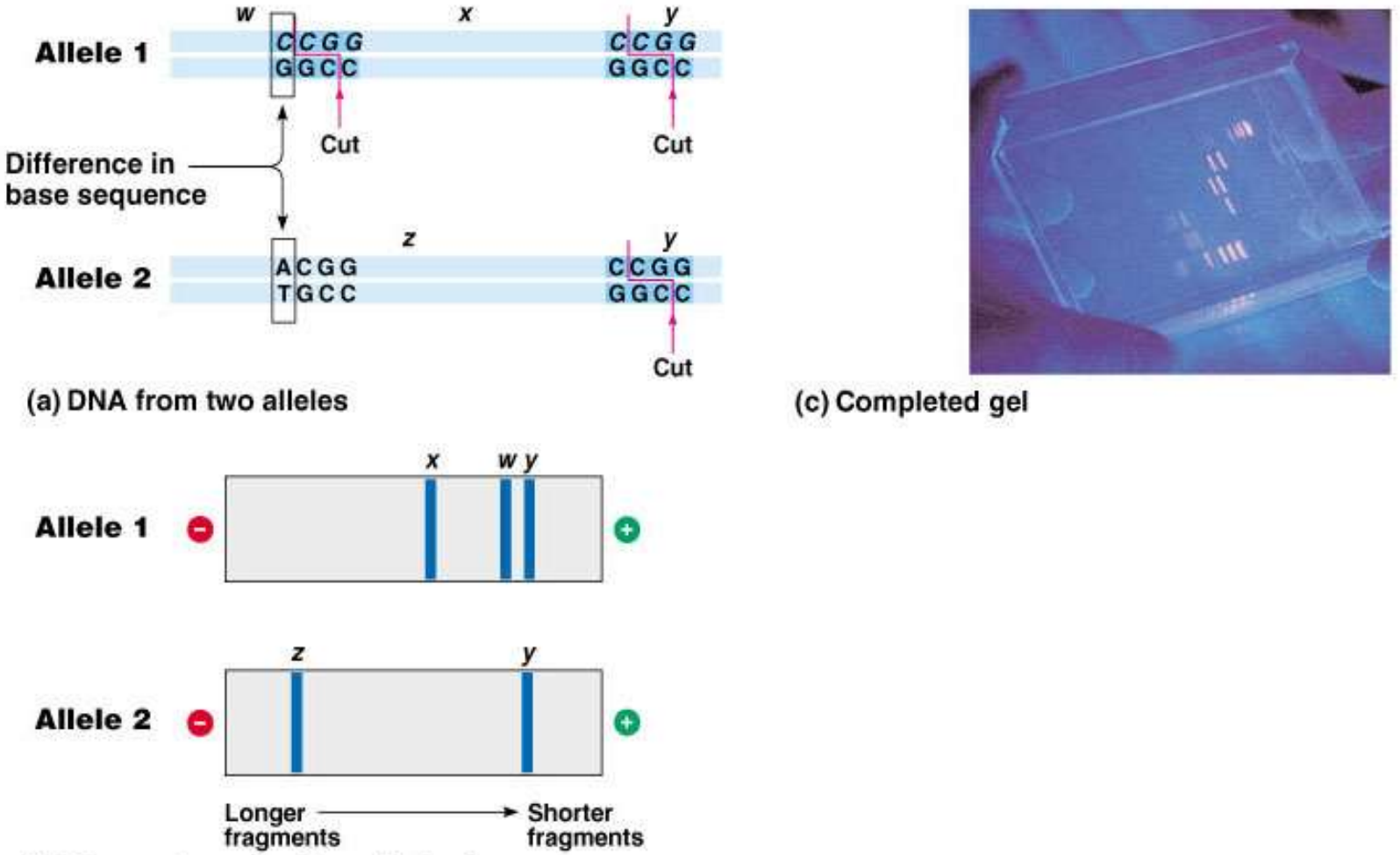

Fig. 20.9

- Gel electrophoresis combined with nucleic acid hybridization allows analyses to be conducted on the whole genome, not just cloned and purified genes.
- Although electrophoresis will yield too many bands to distinguish individually, we can use nucleic acid hybridization with a specific probe to label discrete bands that derive from our gene of interest.
- The radioactive label on the single-stranded probe can be detected by autoradiography, identifying the fragments that we are interested in.

- We can tie together several molecular techniques to compare DNA samples from three individuals.
  - We start by adding the restriction enzyme to each of the three samples to produce restriction fragments.
  - We then separate the fragments by gel electrophoresis.
  - **Southern blotting** (Southern hybridization) allows us to transfer the DNA fragments from the gel to a sheet of nitrocellulose paper, still separated by size.
    - This also denatures the DNA fragments.
  - Bathing this sheet in a solution containing our probe allows the probe to attach by base-pairing (hybridize) to the DNA sequence of interest and we can visualize bands containing the label with autoradiography.

- For our three individuals, the results of these steps show that individual III has a different restriction pattern than individuals I or II.

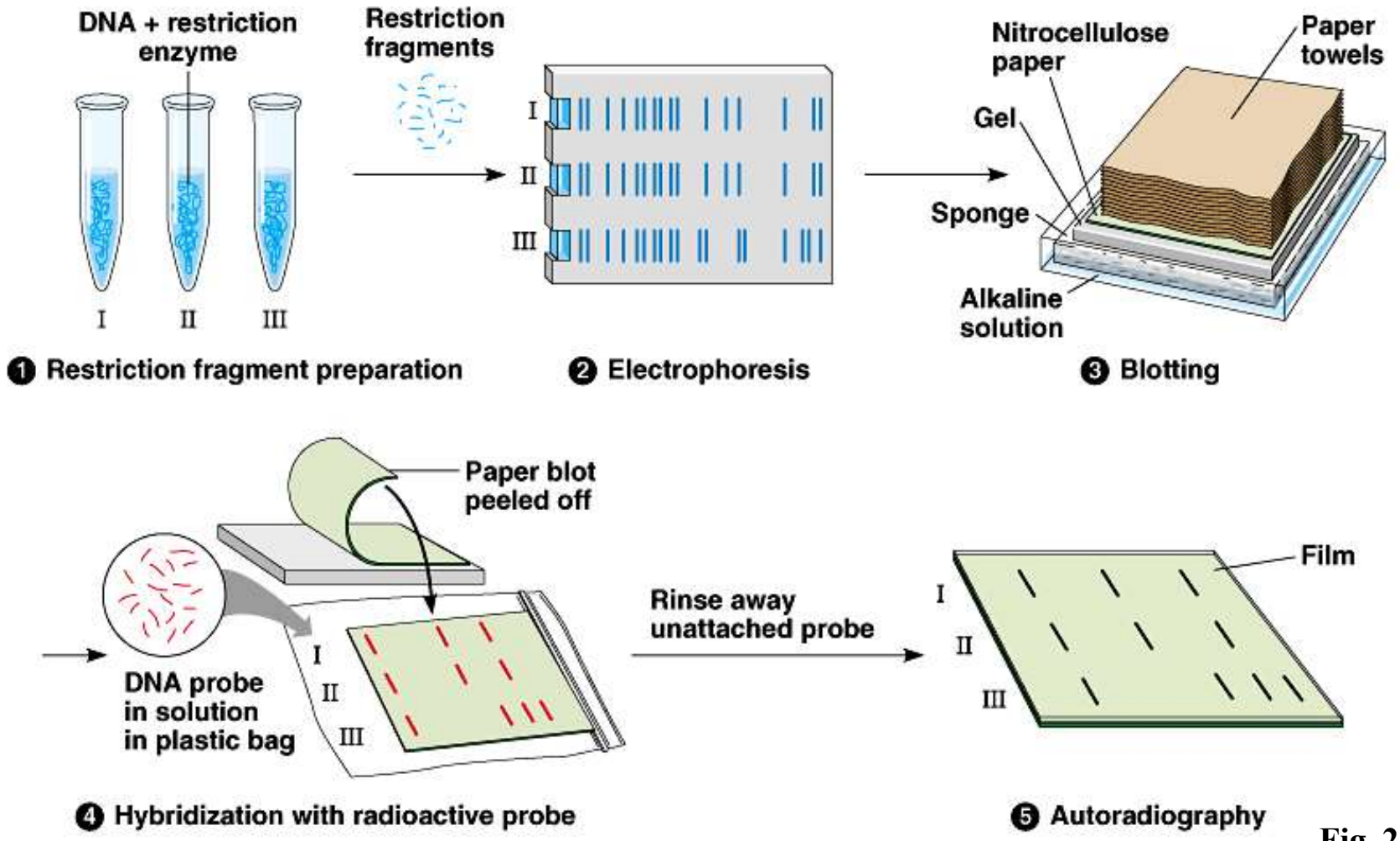

Fig. 20.10

- Southern blotting can be used to examine differences in *noncoding* DNA as well.
- Differences in DNA sequence on homologous chromosomes that produce different restriction fragment patterns are scattered abundantly throughout genomes, including the human genome.
- These **restriction fragment length polymorphisms** (**RFLPs**) can serve as a genetic marker for a particular location (locus) in the genome.
  - A given RFLP marker frequently occurs in numerous variants in a population.

- RFLPs are detected and analyzed by Southern blotting, frequently using the entire genome as the DNA starting material.
  - These techniques will detect RFLPs in noncoding or coding DNA.
- Because RFLP markers are inherited in a Mendelian fashion, they can serve as genetic markers for making linkage maps.
  - The frequency with which two RFPL markers - or a RFLP marker and a certain allele for a gene - are inherited together is a measure of the closeness of the two loci on a chromosome.

2. Entire genomes can be mapped at the DNA level

- As early as 1980, Daniel Botstein and colleagues proposed that the DNA variations reflected in RFLPs could serve as the basis of an extremely detailed map of the entire human genome.
- For some organisms, researchers have succeeded in bringing genome maps to the ultimate level of detail: the entire sequence of nucleotides in the DNA.
  - They have taken advantage of all the tools and techniques already discussed - restriction enzymes, DNA cloning, gel electrophoresis, labeled probes, and so forth.

- One ambitious research project made possible by DNA technology has been the **Human Genome Project,** begun in 1990.
  - This is an effort to map the entire human genome, ultimately by determining the complete nucleotide sequence of each human chromosome.
  - An international, publicly funded consortium has proceeded in three phases: genetic (linkage) mapping, physical mapping, and DNA sequencing.
- In addition to mapping human DNA, the genomes of other organisms important to biological research are also being mapped.
  - These include *E. coli*, yeast, fruit fly, and mouse.

- In mapping a large genome, the first stage is to construct a linkage map of several thousand markers spaced throughout the chromosomes.
  - The order of the markers and the relative distances between them on such a map are based on recombination frequencies.
  - The markers can be genes or any other identifiable sequences in DNA, such as RFLPs or microsatellites.
- The human map with 5000 genetic markers enabled researchers to locate other markers, including genes, by testing for genetic linkage with the known markers.

- The next step was converting the relative distances to some physical measure, usually the number of nucleotides along the DNA.
- For whole-genome mapping, a physical map is made by cutting the DNA of each chromosome into identifiable restriction fragments and then determining the original order of the fragments.
  - The key is to make fragments that overlap and then use probes or automated nucleotide sequencing of the ends to find the overlaps.

- In **chromosome walking,** the researcher starts with a known DNA segment (cloned, mapped, and sequenced) and "walks" along the DNA from that locus, producing a map of overlapping fragments.

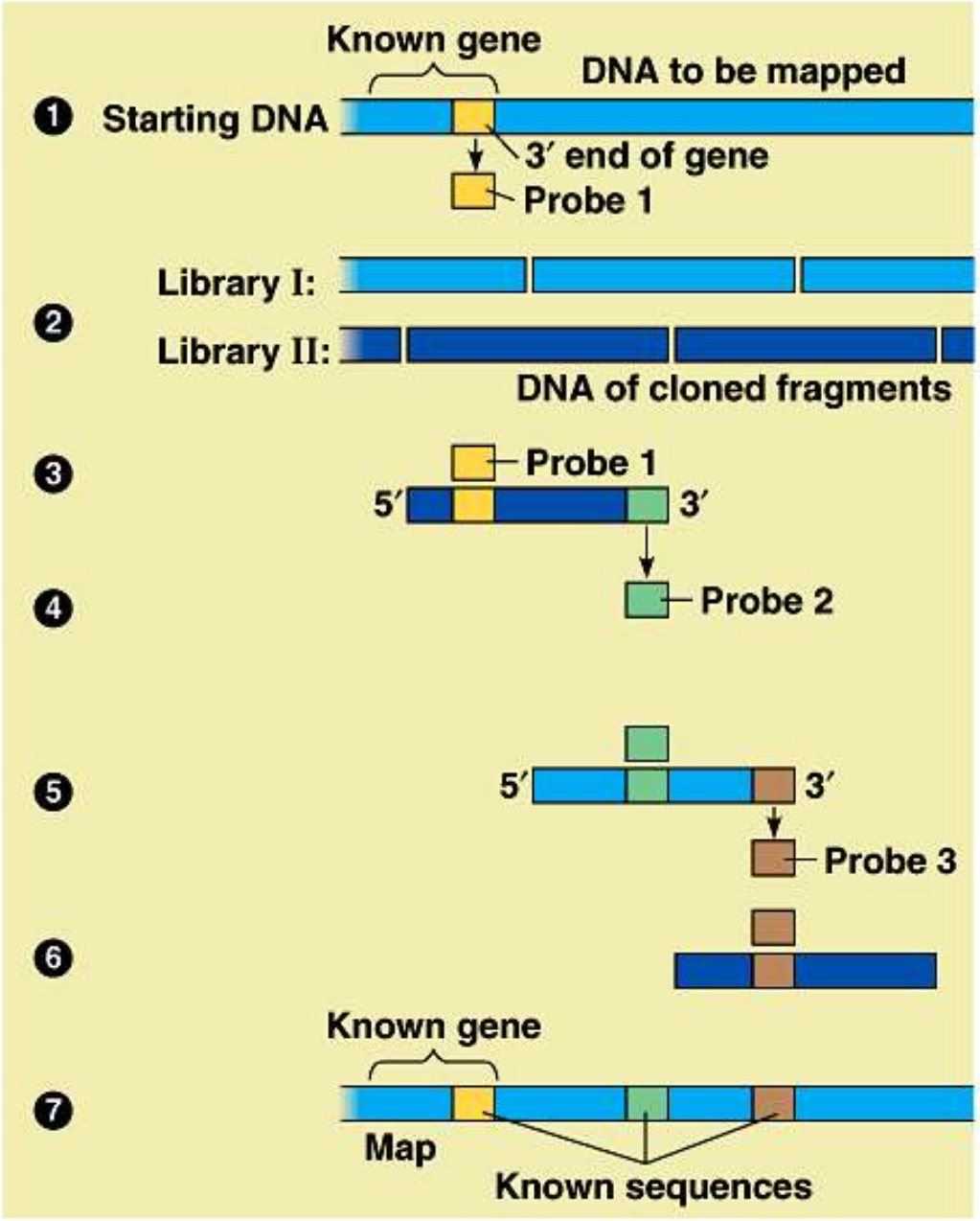

Fig. 20.11

- When working with large genomes, researchers carry out several rounds of DNA cutting, cloning, and physical mapping.
  - The first cloning vector is often a yeast artificial chromosome (YAC), which can carry inserted fragments up to a million base pairs long, or a **bacterial artificial chromosome** (**BAC**), which can carry inserts of 100,000 to 500,000 base pairs.
  - After the order of these long fragments has been determined (perhaps by chromosome walking), each fragment is cut into pieces, which are cloned and ordered in turn.
  - The final sets of fragments, about 1,000 base pairs long, are cloned in plasmids or phage and then sequenced.

- The complete nucleotide sequence of a genome is the ultimate map.
  - Starting with a pure preparation of many copies of a relatively short DNA fragment, the nucleotide sequence of the fragment can be determined by a sequencing machine.
  - The usual sequencing technique combines DNA labeling, DNA synthesis with special chain-terminating nucleotides, and high resolution gel electrophoresis.
  - A major thrust of the Human Genome Project have been the development of technology for faster sequencing and more sophisticated software for analyzing and assembling the partial sequences.

- One common method of sequencing DNA, the Sanger method, is similar to PCR.
- However, inclusion of special dideoxynucleotides in the reaction mix ensures that rather than copying the whole template, fragments of various lengths will be synthesized.
  - These dideoxynucleotides, marked radioactively or fluorescently, terminate elongation when they are incorporated randomly into the growing strand because they lack a 3'-OH to attach the next nucleotide.
- The order of these fragments via gel electrophoresis can be interpreted as the nucleotide sequence.

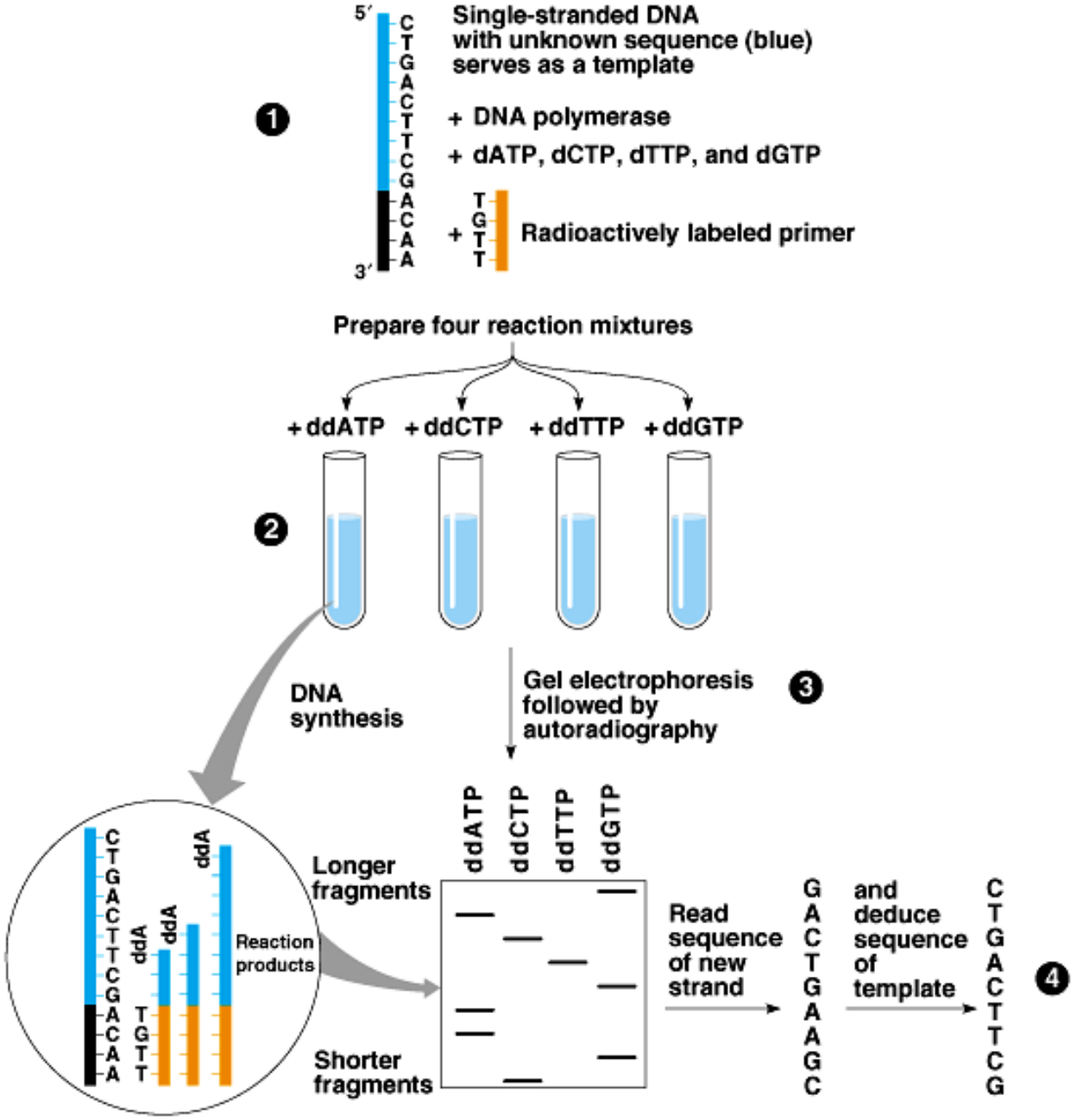

Fig. 20.12

- While the public consortium has followed a hierarchical, three-stage approach for sequencing an entire genome, J. Craig Venter decided in 1992 to try a whole-genome shotgun approach.
  - This uses powerful computers to assemble sequences from random fragments, skipping the first two steps.

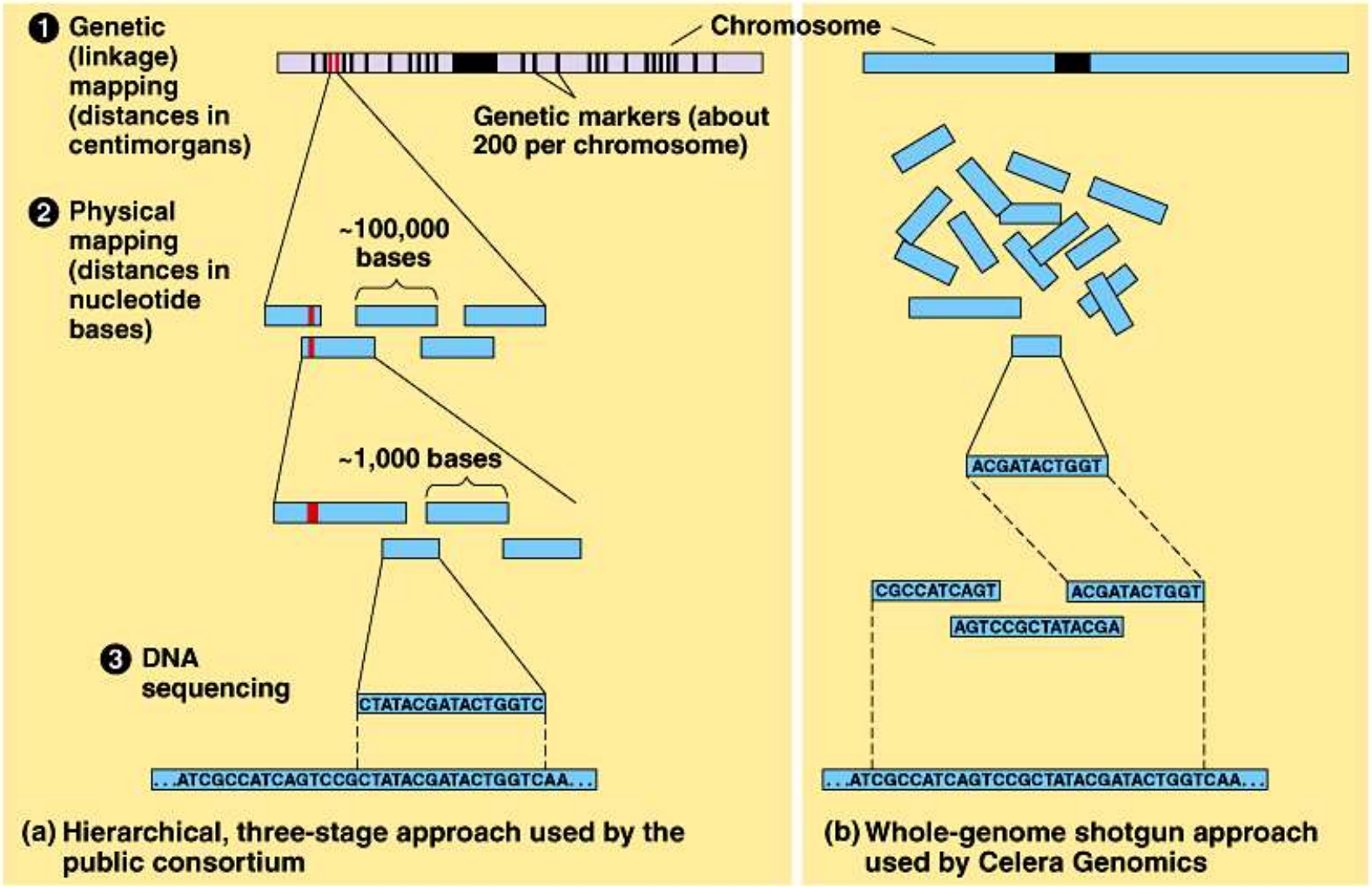

Fig. 20.13

- The worth of his approach was demonstrated in 1995 when he and colleagues reported the complete sequence of a bacterium.
- His private company, Celera Genomics, finished the sequence of *Drosophila melanogaster* in 2000.
- In February, 2001, Celera and the public consortium separately announced sequencing over 90% of the human genome.
  - Competition and an exchange of information and approaches between the two groups has hastened progress.

- By mid-2001, the genomes of about 50 species had been completely (or almost completely) sequenced.
  - They include *E. coli* and a number of other bacteria and several archaea.
  - Sequenced eukaryotes include a yeast, a nematode, and a plant *Arabidopsis thaliana*.
- There are still many gaps in the human sequence.
  - Areas with repetitive DNA and certain parts of the chromosomes of multicellular organisms resist detailed mapping by the usual methods.
  - On the other hand, the sequencing of the mouse genome (about 85% identical to the human genome) is being greatly aided by knowledge of the human sequence.

3. Genome sequences provide clues to important biological questions

- Genomics, the study of genomes based on their DNA sequences, is yielding new insights into fundamental questions about genome organization, the control of gene expression, growth and development, and evolution.
- Rather than inferring genotype from phenotype like classical geneticists, molecular geneticists try to determine the impact on the phenotype of details of the genotype.

- DNA sequences, long lists of A's, T's, G's,and C's, are being collected in computer data banks that are available to researchers everywhere via the Internet.
- Special software can scan the sequences for the telltale signs of protein-coding genes, such as start and stop signals for transcription and translation, and those for RNA-splicing sites.
- From these expressed sequence tags (*EST*s), researchers can collect a list of gene candidates.

- The surprising - and humbling - result to date from the Human Genome Project is the small number of putative genes, 30,000 to 40,000.
  - This is far less than expected and only two to three times the number of genes in the fruit fly or nematodes.
  - Humans have enormous amounts of noncoding DNA, including repetitive DNA and unusually long introns.

Table 20.1 Genome Sizes and Numbers of Genes

Organism	Genome Size	Estimated Number of Genes	Genes per Mb*
H. influenzae (bacterium)	1.8 Mb*	1,700	950
S. cerevisiae (yeast)	12 Mb	6,000	500
C. elegans (nematode)	97 Mb	19,000	200
A. thaliana (plant)	100 Mb	25,000	200
D. melanogaster (fruit fly)	180 Mb	13,000	100
H. sapiens (human)	3,200 Mb	30,000–40,000	10

*Mb = million base pairs

- By doing more mixing and matching of modular elements, humans - and vertebrates in general - reach more complexity than flies or worms.
  - The typical human gene probably specifies at least two or three different polypeptides by using different combinations of exons.
    - Along with this is additional polypeptide diversity via post-translational processing.
  - The human sequence suggests that our polypeptides tend to be more complicated than those of invertebrates.
    - While humans do not seem to have more types of domains, the domains are put together in many more combinations.

- About half of the human genes were already known before the Human Genome Project.
- To determine what the others are and what they may do, scientists compare the sequences of new gene candidates with those of known genes.
  - In some cases, the sequence of a new gene candidate will be similar in part with that of known gene, suggesting similar function.
  - In other cases, the new sequences will be similar to a sequence encountered before, but of unknown function.
  - In still other cases, the sequence is entirely unlike anything ever seen before.
    - About 30% of the *E. coli* genes are new to us.

- Comparisons of genome sequences confirm very strongly the evolutionary connections between even distantly related organisms and the relevance of research on simpler organisms to our understanding of human biology.
  - For example, yeast has a number of genes close enough to the human versions that they can substitute for them in a human cell.
  - Researchers may determine what a human disease gene does by studying its normal counterpart in yeast.
  - Bacterial sequences reveal unsuspected metabolic pathways that may have industrial or medical uses.

- Studies of genomes have also revealed how genes act together to produce a functioning organism through an unusually complex network of interactions among genes and their products.
- To determine which genes are transcribed under different situations, researchers isolate mRNA from particular cells and use the mRNA as templates to build a cDNA library.
- This cDNA can be compared to other collections of DNA by hybridization.
  - This will reveal which genes are active at different developmental stages, in different tissues, or in tissues in different states of health.

- Automation has allowed scientists to detect and measure the expression of thousands of genes at one time using **DNA microarray assays**.
  - Tiny amounts of a large number of single-stranded DNA fragments representing different genes are fixed on a glass slide in a tightly spaced array (grid).
  - The fragments are tested for hybridization with various samples of fluorescently-labeled cDNA molecules.

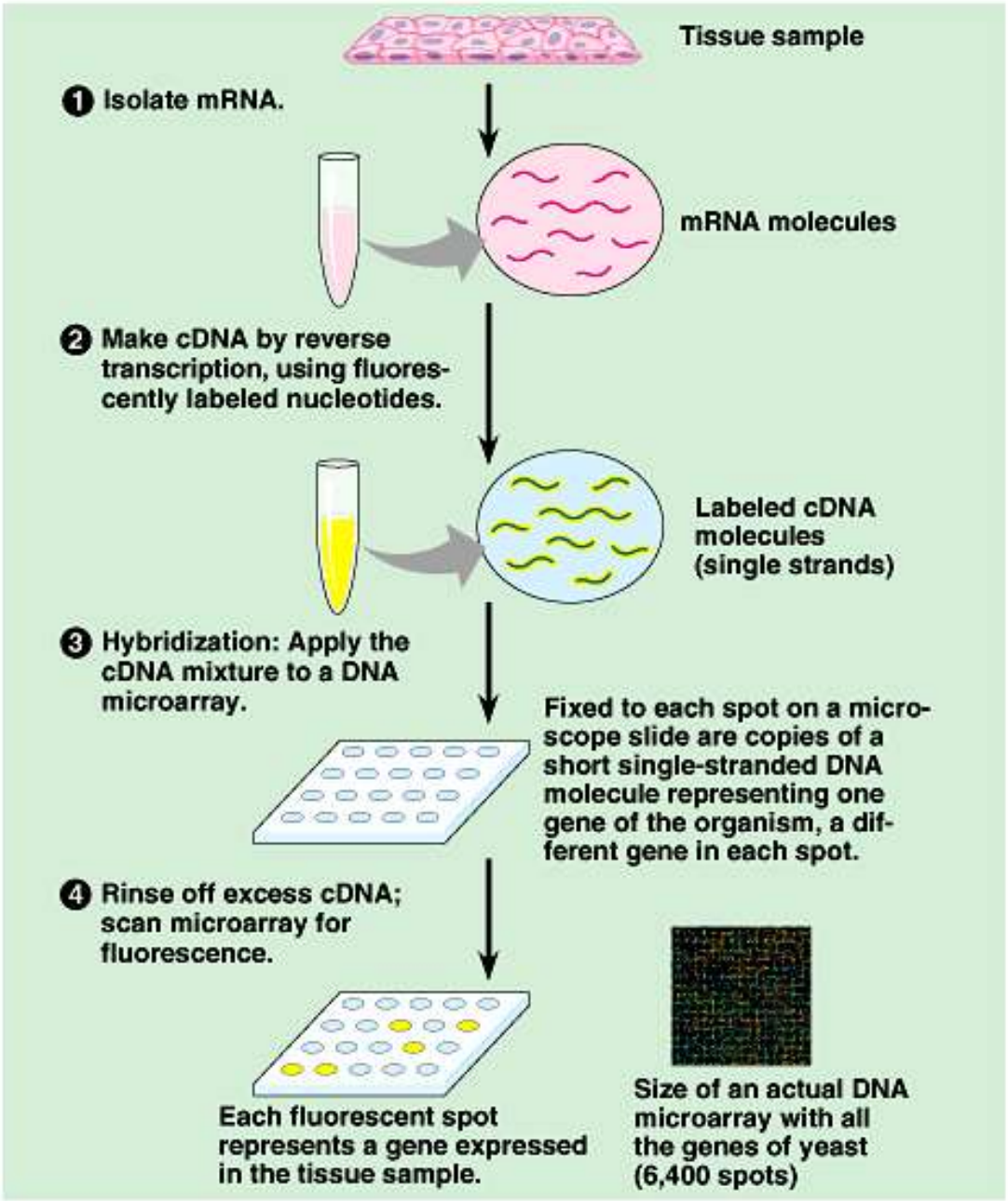

Fig. 20.14a (a) Procedure using labeled cDNA prepared from a tissue sample

- Spots where any of the cDNA hybridizes fluoresce with an intensity indicating the relative amount of the mRNA that was in the tissue.

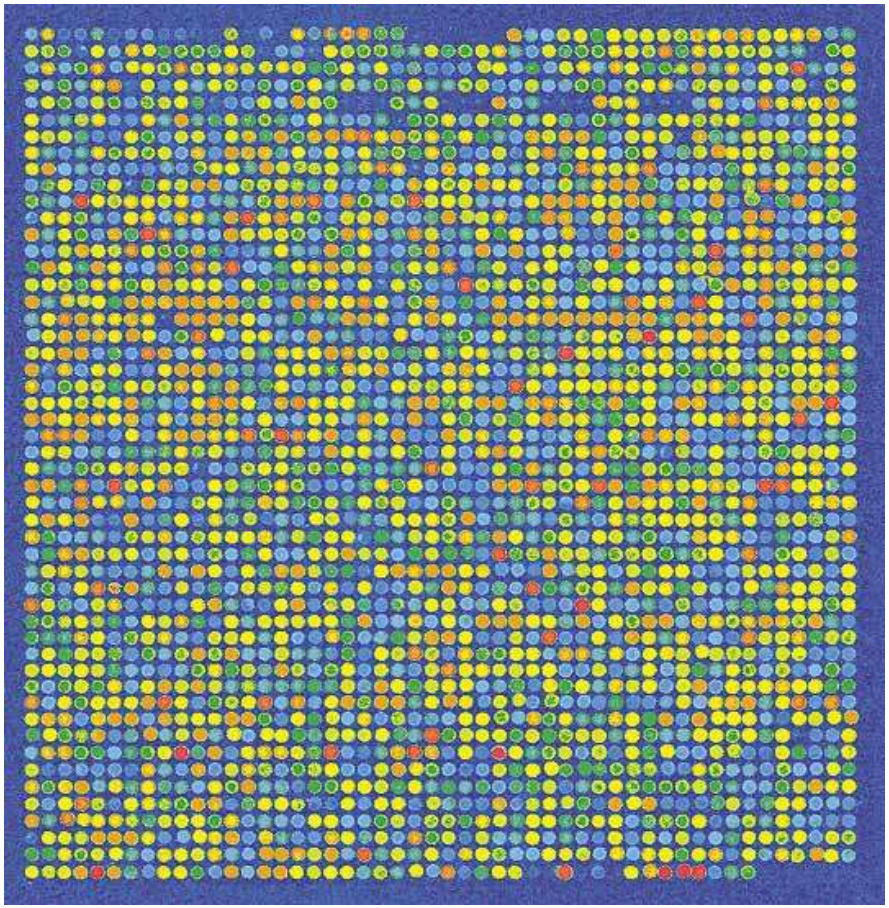

Fig. 20.14b

- Ultimately, information from microarray assays should provide us a grander view: how ensembles of genes interact to form a living organism.
  - It already has confirmed the relationship between expression of genes for photosynthetic enzymes and tissue function in leaves versus roots of the plant *Arabidopsis*.
  - In other cases, DNA microarray assays are being used to compare cancerous versus noncancerous tissues.
    - This may lead to new diagnostic techniques and biochemically targeted treatments, as well as a fuller understanding of cancer.

- Perhaps the most interesting genes discovered in genome sequencing and expression studies are those whose function is completely mysterious.
- One way to determine their function is to disable the gene and hope that the consequences provide clues to the gene's normal function.
  - Using ***in vitro*** **mutagenesis**, specific changes are introduced into a cloned gene, altering or destroying its function.
  - When the mutated gene is returned to the cell, it may be possible to determine the function of the normal gene by examining the phenotype of the mutant.

- In nonmammalian organisms, a simpler and faster method, **RNA interference** (**RNAi**), has been applied to silence the expression of selected genes.
  - This method uses synthetic double-stranded RNA molecules matching the sequences of a particular gene to trigger breakdown of the gene's mRNA.
  - The mechanism underlying RNAi is still unknown.
  - Scientists have only recently achieved some success in using the method to silence genes in mammalian cells.

- The next step after mapping and sequencing genomes is **proteomics**, the systematic study of full protein sets (*proteomes*) encoded by genomes.
  - One challenge is the sheer number of proteins in humans and our close relatives because of alternative RNA splicing and post-translational modifications.
  - Collecting all the proteins will be difficult because a cell's proteins differ with cell type and its state.
  - In addition, unlike DNA, proteins are extremely varied in structure and chemical and physical properties.
  - Because proteins are the molecules that actually carry out cell activities, we must study them to learn how cells and organisms function.

- Genomic and proteomics are giving biologists an increasingly global perspective on the study of life.
- Eric Lander and Robert Weinberg predict that complete catalogs of genes and proteins will change the discipline of biology dramatically.
  - "For the first time in a century, reductionists [are yielding] ground to those trying to gain a holistic view of cells and tissues."
- Advances in **bioinformatics**, the application of computer science and mathematics to genetic and other biological information, will play a crucial role in dealing with the enormous mass of data.

- These analyses will provide understanding of the spectrum of genetic variation in humans.
  - Because we are all probably descended from a small population living in Africa 150,000 to 200,000 years ago, the amount of DNA variation in humans is small.
  - Most of our diversity is in the form of **single nucleotide polymorphisms** (**SNPs**), single base-pair variations.
    - In humans, SNPs occur about once in 1,000 bases, meaning that any two humans are 99.9% identical.
  - The locations of the human SNP sites will provide useful markers for studying human evolution and for identifying disease genes and genes that influence our susceptibility to diseases, toxins or drugs.