CHAPTER 10

THE NATURE OF THE GENE AND THE GENOME

OBJECTIVES

- Review early discoveries in genetics that put us on the path to gain the knowledge we now have.
- ~~Review the significance of crossing over in the recombination of genes~~.
- Describe the structure of DNA and the discoveries that led to its determination.
- Summarize the importance of the Watson-Crick proposal of DNA structure.
- ~~Describe DNA supercoiling and the topoisomerases that are used to control it in living organisms~~.
- ~~Introduce students to the concept of genome complexity~~.
- ~~Describe the process of DNA denaturation and the factors that influence DNA melting temperature~~.
- ~~Describe the ability of DNA to renature and emphasize the usefulness of nucleic acid hybridization in a variety of biological studies~~.
- Discuss the types of DNA sequences found in prokaryotic and eukaryotic cells and outline what is known about their function.
- Elaborate on the mechanisms by which genes may be duplicated giving rise to gene families and superfamilies.
- ~~Describe transposons and the mechanisms by which they displace genes throughout the genome.~~
- ~~Explain the medical impact of transposons and describe their role in evolution~~.
- Discuss the function and operation of restriction endonucleases.
- ~~Explain RFLPs and the role they have played in the production of molecular maps of the genome and DNA fingerprinting technology.~~

LECTURE OUTLINE

The Concept of a Gene as a Unit of Inheritance

I. Science of genetics began in 1860s with the work of Gregor Mendel (friar at the abbey of St. Thomas located in today's Czech Republic; his lab was a small garden plot on the grounds of the abbey
   A. Mendel crossed pea plants & found the inheritance pattern of defined traits; when he started, he evidently had a clear experimental plan in mind
      1. His goal was to mate or cross pea plants having different inheritable characteristics & to determine the pattern by which these characteristics were transmitted to the offspring
      2. He chose the garden pea for a number of practical reasons, including that he could buy a variety of seeds that would give rise to plants with distinct characteristics

B. Mendel focused on 7 clearly definable traits, like plant height & flower color, each of which occurred in 2 alternate & clearly identifiable forms
C. He crossbred plants through several generations & counted the number of individuals with various traits
D. After several years of research, he published his findings in 1865 in journal of Natural History Society of Brünn, Austria; lost until rediscovered independently in 1900 by 3 botanists who gave him credit
    1. The minutes of the meeting record no discussion of his presentation
    2. They generated no interest whatsoever until 1900, 16 years after his death

II. Mendel's conclusions – expressed in modern genetic terminology
A. Characteristics of the plants were governed by distinct factors or units of inheritance (**genes**)
    1. Each organism has 2 copies of gene that controls development for each trait, one from each parent
    2. The two genes may be identical to one another or nonidentical (may have alternate forms or **alleles**)
    3. For each of the 7 traits he studied, one of the two alleles was dominant over the other; dominant alleles mask recessive alleles when they are together in same organism
B. Gametes (reproductive cells) from each plant have only 1 copy of the gene for each trait (dominant or recessive, not both); plants arise from union of male & female gametes
    1. One of the alleles governing each trait in a plant was inherited from the female parent; the other one was inherited from the male parent
C. Law of Segregation - an organism's 2 alleles separate (segregate) from one another during gamete formation, even though they remained together throughout the individual plant's life
D. Law of Independent Assortment - segregation of allelic pair for one trait has no effect on segregation of alleles for another trait
    1. A particular gamete can get paternal gene for one trait & maternal gene for another

III. Despite the evidence that inherited traits were governed by discrete factors (genes), Mendel's studies were totally unconcerned with the physical nature of these elements or their location in the organism
A. Between the time of Mendel's work & its rediscovery, many investigators became concerned with the physical basis of heredity within the cell

Chromosomes: the Physical Carriers of Genes

I. The discovery of chromosomes (colored bodies) - got their name in 1888
A. By 1880s, many European biologists realized that inherited characteristics must be passed from cell to cell & from generation to generation, even though they knew nothing of Mendel's results
    1. They were observing cell activities with rapidly improving light microscopes
    2. They were able to observe newly discernible cell structures
    3. They realized that all the genetic information needed to build & maintain a complex plant or animal had to fit within the boundaries of a single cell
B. Walther Flemming, early 1880s – observed dividing cells; during mitosis, noted that cytoplasmic contents were divided randomly & nuclear material divided equally between the daughter cells
    1. During cell division, nuclear material became organized into visible threads called **chromosomes** (colored bodies)

2. Chromosomes appeared as doubled structures, split to single structures & doubled at next division

II. Fertilization and meiosis: Changing chromosome numbers
- A. About the time of Flemming's work, fertilization was observed; the 2 gametes' roles were described
- B. Sperm & egg are very different cells - egg very large, sperm tiny
  1. Sperm & egg equally important genetically - nuclei & chromosomes were apparently identical
  2. Importance of chromosomes contributed by male demonstrated by Boveri
- B. Theodore Boveri (German biologist) - studied sea urchin eggs fertilized by 2 sperm (**polyspermy**) instead of the normal one single sperm
  1. Results in disruptive cell divisions & early death of embryo
  2. Second sperm donates extra chromosome set & extra centriole, causing abnormal cell divisions
  3. Daughter cells receive variable numbers of chromosomes
  4. Concluded that the orderly process of normal development depends upon a particular combination of chromosomes & that each chromosome possesses different qualities
  5. This was the first evidence for a qualitative difference among chromosomes
- C. Events occurring after fertilization were followed most closely in the roundworm *Ascaris*; its few chromosomes were large & readily observed in the 19th century
  1. Edouard van Beneden (Belgian biologist, 1883) found that *Ascaris* egg & sperm nuclei in egg just after fertilization had 2 chromosomes each before fusion, but that somatic cells had 4 chromosomes
  2. About this time, the process of meiosis was described
- D. August Weismann (German biologist, 1887) - proposed that meiosis included reduction division during which chromosome number was reduced by half before gamete formation
  1. If no reduction division, the gametes would contain same number of chromosomes as adult cells & union of 2 gametes would double chromosome number in cells of progeny
  2. This would double chromosome number with every succeeding generation which cannot occur

III. Chromosomes as the carriers of genetic information – after the rediscovery of Mendel's work, researchers realized that genetic material must behave in a manner consistent with Mendelian principles
- A. Walter Sutton (Columbia U., 1903) – published a paper that pointed directly to chromosomes as the physical carriers of Mendel's genetic factors
  1. Studied grasshopper sperm formation (big, easily observable chromosomes); saw 23 chromosomes (11 homologous chromosome pairs [look alike] & extra accessory [sex or X] chromosome (no mate)
  2. Saw 2 different kinds of spermatogonia (cells giving rise to sperm) division: mitosis (spermatogonia make more spermatogonia) & meiosis (spermatogonia make cells that differentiate into sperm)
  3. The homologous pairs he noticed correlated perfectly with Mendel's inheritable pairs of factors
  4. Hypothesized that chromosomes are physical carriers of Mendel's genetic factors
- B. In cells just starting meiosis, he saw the members of each pair associated with one another; forming a complex called a **bivalent**; 11 were visible, each with cleft along its length where they were associated
  1. First meiotic division separated pair members (Weismann's reduction division proposed 15 yr before on theoretical grounds)

2. Explained Mendel's proposal that hereditary factors exist in pairs that remain together through organism's life until they separate with the production of gametes

C. His observations explained a number of other Mendelian findings:
1. How gametes could only contain 1 allele of each gene
2. How the number of gametes containing 1 allele was equal to the number containing the other allele
3. 2 gametes that united at fertilization would produce an individual with 2 alleles for each trait (reconstitution of allelic pairs)

IV. Chromosome as linkage group – Sutton's discovery suggested that many of Mendel's factors are grouped on a single chromosome & should act as if they are linked & not independently

A. How did Mendel's factors assort independently? - the 7 traits were on different linkage groups (pairs of homologous chromosome) or were far enough apart on the same chromosome to act independently

1. Mendel owed his results to good luck or lack of interest in traits that did not fit his predictions

B. Traits soon seen on same chromosome in many organisms, like flower color & pollen shape in sweet peas

V. Genetic analysis in *Drosophila* - Thomas Hunt Morgan (Columbia U., 1909) and others; began what was to be the start of a new era in genetic research

A. Fruit flies were ideal for genetic studies & became the focus of them

1. Short generation time (14 days) from egg to maturity & produce up to 1000 eggs in a lifetime
2. Very small so one can keep large number on hand; easy to maintain & breed; also very inexpensive
3. Only one disadvantage; there was only strain of fly available, the **wild type**

B. Needed variants from wild type so he bred flies & looked for mutants thinking that if he bred enough flies mutants would appear; finally found one with white eyes, not the normal red eyes

C. Soon (by 1915) he & his students found 85 mutants with a wide variety of affected structures
1. It became obvious that on rare occasions, a spontaneous change or **mutation** occurred within a gene altering it permanently so that the change could be passed from generation to generation
2. The demonstration of spontaneous, inheritable alteration in a gene suggested a mechanism for the origin of variation that exists in populations & evidence for a vital link in the theory of evolution
3. Variants of genes could arise spontaneously —> natural selection can act on these mutations —> then new species slowly emerge

D. The 85 mutations did not all assort independently; they belonged to 4 different linkage groups, one of which contained very few mutant genes (only 2 in 1915)

1. This correlated to the number of *Drosophila* homologous chromosome pairs (4), one very small

VI. Crossing over & recombination - linkage was found to be incomplete; alleles of 2 different genes originally present on the same chromosome do not always stay together during gamete production

A. Can reshuffle genes; could have alleles originally derived from both parents on same chromosome or alleles from 1 parent could end up on both members of a homologous pair & in different gametes

B. F. A. Janssens (1909) – observed that homologous chromosomes of bivalents wrap around each other during early stage in meiosis; he hypothesized breakage & exchange of pieces of chromosomes

C. Morgan (1911) - thought that this crossing over (genetic recombination) could account for appearance of offspring (recombinants) having unexpected genetic trait combinations & breakdown in linkage

1. Said Janssens was right; homologous chromosomes break/rejoin, exchange pieces & separate alleles

D. Analyses of offspring from a large number of crosses between adults carrying a variety of alleles on the same chromosome indicated that:

1. Recombination percentage between a given pair of genes, such as eye color & wing length, was constant from cross to cross
2. Recombination percentages between different pairs of genes, such as between eye color & wing length vs. eye color & body length could be very different

E. These results suggest that the position (**locus**) of genes on chromosomes is fixed & does not vary fly to fly; therefore, recombination percentage (frequency) is a measure of distance between 2 genes

1. A bigger distance between genes leads to a higher probability of a break & thus more crossovers between genes & the greater the recombination frequency

F. Alfred Sturtevant (Columbia U. undergraduate working for Morgan, 1911) realized recombination frequencies could be used to map relative positions of individual genes along given chromosome

1. Sturtevant constructed detailed maps of the 4 chromosomes of the fruit fly
2. This principle has been used to map genes in many organisms (viruses, bacteria, eukaryotes)
3. Alleles at opposite ends of chromosome have so many crossovers between them that they assort independently (essentially like they are on separate chromosomes)
4. Genes very close to one another are less likely to become unlinked (since less recombination)

VII. Mutagenesis & giant chromosomes – finding mutants was tedious at first since investigators had to wait for their spontaneous appearance, but then X-rays were found to increase the rate of mutant production

A. Herman Muller (Indiana U., 1927) – showed genetic material can be damaged by X-rays; used special strain of fruit flies designed to reveal the presence of recessive alleles

1. Sublethal X-ray dose raises spontaneous mutation frequency >100 times that of nonirradiated controls
2. Later, other mutagenic agents, like UV irradiation, were demonstrated; they increased the mutants available for research & pointed out hazards of radiation usage in fields of industry & medicine
3. Revealed that the genetic material, whatever its chemical nature, had to be sensitive to electromagnetic radiation
4. Today, mutations in *Drosophila* are most often generated by adding a chemical mutagen (ethyl methane sulfonate) to the animal's feed

B. Theophilus Painter (U. of Texas, 1933) - rediscovered giant chromosomes in certain insect cells

1. Cells from the larval salivary gland of *Drosophila* contain chromosomes that are ~100 times thicker than those found in most other cells of the organism; they grow without division

2. During larval development, these cells stop dividing, but keep growing; DNA replication continues, provides added genetic material needed to support the high level of secretory activity in these cells
3. Duplicated DNA strands stay attached to each other in perfect side-by-side alignment (in register)
4. The cells produce giant chromosomes with as many as 1,024 times the number of DNA strands of normal chromosomes because cell stops dividing
5. But the cells keep growing, allowing these huge cells to maintain their high secretory activity rate
6. These unusual polytene chromosomes are rich in visual detail displaying ~5000 bands when stained & examined microscopically
7. The banding pattern is essentially constant from one individual to the next, but considerable differences are seen between the chromosomes from flies of different species of *Drosophila* genus
8. Soon found that individual bands correlated with specific genes & the relative positions of these genes agreed with those predicted by genetic maps from recombination frequencies
9. Provided visual confirmation of the validity of the entire mapping procedure

C. Giant insect chromosomes have been useful in other ways
1. Banding pattern comparisons of polytene chromosomes of different species allow investigation of evolutionary changes at chromosome level
2. Chromosomes are dynamic structures; certain regions puff out at particular developmental stages; puffs are sites where DNA is transcribed into RNA (good for direct visualization of gene expression)

The Chemical Nature of the Gene

I. The structure of DNA (much of this is discussed in detail in outline for Chapter 2) – structure was solved by James Watson & Francis Crick at Cambridge Univ. in 1953; they started with facts known at the time

A. Base composition - DNA building blocks are nucleotides (5C-sugar deoxyribose, phosphate group esterified to the 5' carbon of the sugar, nitrogenous base [A, G, C, T] at 1'-position of the sugar)
1. 2 types of bases - pyrimidines (cytosine [C], thymine [T]) which contain a single ring & purines (adenine [A], guanine [G]), which contain 2 rings
2. Nucleotides are covalently linked to one another to form a linear polymer (strand) with backbone of alternating sugar & phosphate groups joined by **3'-5'-phosphodiester bonds**
3. Bases attached to each sugar was thought to project from backbone like column of stacked shelves
4. Nucleotide (& polymer) is polarized so that each strand has a direction - **5' end** (phosphate) & **3' end** (3'-OH of sugar)

5. X-ray diffraction analysis revealed dimensions (the distance between nucleotides of the stack - 3.4 Å (0.34 nm) between nucleotides in stack; suggested a large structural repeat every 34 Å (3.4 nm)

B. Some terminology
   1. 1 of 4 nitrogenous bases + a pentose sugar = a nucleoside; if the sugar is deoxyribose, it is a deoxyribonucleoside; 4 of them: deoxyadenosine, deoxyguanosine, deoxythymidine, deoxycytidine
   2. If ≥1 phosphates are attached to nucleoside —> nucleotide; add 1 phosphate —> deoxyadenosine 5'-monophosphate, deoxyadenosine 5'-diphosphate, deoxyadenosine 5'-triphosphate
      a. Phosphate groups are usually attached at 5'-position but alternatively can attach at 3'-position
      b. There are nucleoside 5'-monophosphates, nucleoside 5'-diphosphates & nucleoside 5'-triphosphates, depending on the number of phosphates attached
      c. Examples of these are: deoxyadenosine 5'-monophosphate (dAMP), deoxyguanosine 5'-diphosphate (dGDP), & deoxycytidine 5'-triphosphate (dCTP)
   3. A similar set of nucleosides & nucleotides involved in RNA metabolism contain ribose rather than deoxyribose; these nucleotides contain ribose & are often employed in energy metabolism like ATP

C. Tetranucleotide Theory – for many years, the original concept of DNA was a simple repeating tetranucleotide (—ATGCATGC—); unlikely to carry much information; base ratio should be 1:1:1:1

D. Erwin Chargaff (Columbia, 1950) - determined relative amounts of every base (base composition) in various DNA samples; he believed the sequence of nucleotides in DNA held the key to its importance
   1. He hydrolyzed bases from their attached sugars, separated the bases in the hydrolysate by paper chromatography, & determined the amount of material in each of 4 spots to which bases migrated
   2. Found relative base amounts varied from species to species, it was not always 1:1:1:1 as would be expected if Tetranucleotide Theory were correct, but it was always constant within species
   3. Purines always equaled pyrimidines in given DNA sample: [A] = [T]; [G] = [C]; [A] + [T] ≠ [G] + [C]
   4. Different species had different ratios: A:G ratio of tubercle bacillus = 0.4; human DNA - 1.56; within species, different tissues had the same ratio
   5. Chargaff gave DNA molecule specificity & individuality from one organism to another; the significance of the base equivalencies remained obscure for a time

II. Watson - Crick Proposal - used X-ray diffraction data [from Rosalind Franklin & Maurice Wilkins, King's College, London] & made cutout nucleotide models [Watson & Crick] —> proposed DNA structure

A. Information about 3D-organization of DNA was needed if its biological activity was to be understood

B. Composed of 2 nucleotide chains (Pauling had suggested three strands)

C. The 2 chains spiral around each other & central axis to form a pair of right-handed helices
   1. In right-handed helix, if one looks down central axis, each strand follows a clockwise path moving away from observer
   2. The helical nature of DNA was shown by the pattern of spots in Franklin's X-ray diffraction image, which was shown to Watson during a visit to King's College

D. The 2 chains of double helix run in opposite directions (**antiparallel**); one 5' —> 3', the other 3' —> 5'

E. The —sugar-phosphate-sugar-phosphate— backbone of each strand is on the outside of the molecule; bases project in toward center; phosphate groups give the molecule a large negative charge (an acid)
    1. Pauling's model had backbone in center
F. Planes of the bases are approximately perpendicular to the long axis of the molecule & stack one on top of another like a pile of plates
    1. Hydrophobic interactions & van der Waal forces between the stacked, planar bases provide stability for the whole DNA molecule
    2. Together the helical turns & planar base pairs (bp) make it resemble a spiral staircase
G. Chains held together by H bonds between each base of one strand & the associated base on other strand
    1. Since individual H bonds are weak & easily broken, the DNA strands can become separated during various activities
    2. H bond strength is additive; the large numbers of H bonds holding the strands together make the double helix a stable structure
H. The distance from the backbone phosphorus atom to the axis center is 10 Å (1 nm); DNA double helix width is thus 20 Å (2 nm)
I. Pyrimidine in one chain always pairs with purine in the other chain (results in 20 Å or 2 nm width along its entire length as well); 2 purines would be wider, 2 pyrimidines narrower
J. Only A can bind to T; only G can bind to C (matches Chargaff's Rules); A—T pairs joined by 2 H bonds & G—C pairs joined by 3 H bonds
    1. Nitrogen atoms on cytosine C4 & adenine C6 are mostly in amino (NH_2) rather than imino (NH) form
    2. Oxygens on guanine C6 & thymine C4 are mostly in keto (C=O), rather than enol (C—OH) form
    3. These structural restrictions on base configurations are responsible for pairing restrictions: A was only purine able to bind T & G was only purine that could bind to C
K. From outside of molecule, spaces between adjacent turns of helix form 2 grooves of different width
    1. Wider major groove; more narrow minor groove; grooves spiral around double helix outer surface
    2. DNA binding proteins often contain domains that fit into grooves; sometimes such a protein can read the sequence of nucleotides along the DNA without having to separate the strands
L. Double helix makes 1 full turn every 10 residues (3.4 nm); 150 turns/million daltons molecular mass
M. No restriction on given chain base sequence, but sequence of 1 chain specifies the other's sequence (**complementarity**); their nucleotide sequences are always fixed relative to each other
    1. Complementarity is of overriding importance in nearly all of the activities & mechanisms in which nucleic acids are involved

III. Importance of the Watson-Crick Proposal
A. Primary functions of genetic material – from the time it was first thought to be the genetic material, it was expected to fulfill 3 primary functions
    1. Storage of genetic information - determines all of the inheritable characteristics of organisms; amino acid sequence of all proteins in organism must be specified within the DNA structure
    2. Self-duplication & inheritance – DNA must contain information for its own replication; allows genetic instructions to be transmitted from one cell to its daughters & to organism's offspring

3. Expression of genetic message – DNA is director of cellular activity; genes usually encode proteins; DNA must direct protein synthesis & the order of amino acid incorporation into polypeptide

B. Watson - Crick model explains how the first 2 functions might be accomplished; model confirmed that:

1. Information resides in DNA base sequence; a given DNA segment would correspond to each gene; change DNA sequence —> get inheritable change (mutation) in protein coded for
2. Chains separate as H bonds break (like zipper) & serve as template for synthesis of new strand —> at the end of replication, get 2 strands identical to each other & the original DNA molecule
   a. According to the proposal, each new DNA double helix would contain one strand from original DNA molecule & one newly synthesized strand
3. Only the mechanism by which DNA governs assembly of specific protein stayed a total mystery

C. Once structure was accepted, any theory of genetic code, DNA synthesis or information transfer had to be consistent with that structure

IV. DNA supercoiling

A. Jerome Vinograd et al. (Caltech, 1963) found that 2 closed DNA circles of identical molecular mass could exhibit very different sedimentation rates after density gradient centrifugation

1. Fast sedimenting DNA had more compact shape because the molecule was twisted (supercoiled) upon itself, much like rubber band with ends twisted in opposite directions or tangled phone cord
2. DNA in this state is said to be supercoiled; it is more compact than its relaxed counterpart, occupies less volume & moves more rapidly in response to centrifugal force or an electrical field

B. Picture DNA lying free on flat surface - it has right number of 10 base pairs/turn & is said to be relaxed

1. Fuse two ends of strands to form circle without twisting & it would still be relaxed
2. Twist molecule in direction opposite to that in which duplex was wound before fusion into circle —> molecule unwinds (greater number of bp/turn); DNA is underwound (**negatively supercoiled**)
3. Since molecule is most stable with 10 bp/turn, it tends to resist the strain of becoming underwound by becoming twisted upon itself into a supercoiled conformation
4. Circular DNA in nature (mitochondrial, viral, bacterial, etc.) is invariably negatively supercoiled
5. **Positively supercoiled** DNA is overwound

C. Supercoiling is not restricted to small circular DNAs, but also occurs in linear, eukaryotic DNA

1. Negative supercoiling plays a role in allowing chromosomal DNA to be compacted so as to fit into microscopic cell nucleus
2. Since negatively supercoiled DNA is underwound, it exerts a force that helps separate 2 strands of helix, which is required during both replication (DNA synthesis) & transcription (RNA synthesis)
3. Linear DNA wrapped around protein cores (**nucleosomes**) also supercoils

V. Topoisomerases alter DNA duplex supercoiled state (topology), they further supercoil or relax it; seen in both prokaryotic & eukaryotic cells; cells have variety of topoisomerases - 2 classes; different mechanisms

A. Type I topoisomerases - create transient break in one duplex strand, & then allows the intact, complementary strand to undergo a controlled rotation, which relaxes the supercoiled molecule
   1. Essential for processes like DNA replication & transcription to prevent excessive supercoiling from developing as the complementary strands of the DNA duplex separate & unwind

B. Type II topoisomerases - make transient break in both DNA duplex strands, then transport another segment of DNA molecule (or separate molecule entirely) through the break & reseal severed strands
   1. This complex reaction mechanism is accompanied by a series of dramatic conformational changes
   2. They not only relax & supercoil DNA, but also tie DNA molecule into knots, untie knot, interlink independent DNA circles (catenation) or separate interlinked circles into individual components
   3. Topoisomerase II is required to unlink DNA molecules before duplicated chromosomes can be separated during mitosis
   4. Human topoisomerase II is target for numerous drugs (etoposide, doxorubicin) used to kill rapidly dividing cancer cells; the drugs bind to enzyme & keep cleaved DNA strands from being resealed

~~The Structure of the Genome: Denaturation, Renaturation and Complexity~~

I. Genome - unique content of genetic information (humans - all the genetic information present in a single **haploid** set of 23 chromosomes, including 22 different autosomes & both the X & Y sex chromosomes
   A. DNA structure can be probed by X-ray crystallography; genome structure is tougher to get at
   B. The sum of the genetic information in an individual is the sum of all the DNA segments in the fertilized egg at the start of life
   C. All members of species population share the same set of genes, even though each individual possesses slightly different versions (**alleles**) of many genes

II. DNA denaturation (melting) – heating & separation of the 2 DNA chains held together by weak, noncovalent bonds; very important property of DNA double helix
   A. Dissolve DNA in saline solution & slowly warm it —> strands start to separate at specific temperature
      1. Within a few degrees, the process is complete & the solution contains single-stranded molecules that are completely separate from their original partners
      2. Usually follow denaturation progress by monitoring the increase in absorbance of dissolved DNA
      3. Nucleic acid nitrogenous bases absorb ultraviolet (UV) radiation with an absorbance maximum near 260 nm
      4. Once 2 DNA strands have separated, hydrophobic interactions resulting from base stacking are greatly decreased, changing the electronic nature of bases & increasing their UV light absorbance
   B. Melting temperature (T_m) is the temperature at which the shift in absorbance is half completed
      1. T_m rises if DNA has a high G-C content (%G + %C) due to the extra H bonds holding it together indicating a higher stability as compared with AT-rich DNA regions (with only 2 H bonds/bp)

2. Even within a single DNA molecule, A-T-rich sections melt before G-C-rich segments

III. DNA renaturation – DNA denaturation not a surprise, but reassociation of single strands into stable double-stranded molecules seemed much less likely
   A. Julius Marmur et al. (Harvard, 1960) - after denaturation, separated DNA chains can reassociate into stable double helices in a process called **renaturation** or **reannealing**
      1. Slowly cooled thermally denatured bacterial DNA —> it renatured, regaining properties of double helix (absorbance of UV light dropped & it transformed bacteria acting like genetic material) **or**
      2. Got similar results when DNA was melted at 100°C & then its temperature was rapidly dropped to ~25°C below T_m followed by incubation at the lower temperature for a period of time
      3. These studies showed that complementary single-stranded DNAs can reassociate or reanneal
      4. This is one of the most valuable observations ever made in molecular biology
   B. Reannealing has served as the basis for investigations of DNA complexity & has led to the development of nucleic acid hydridization technology
      1. What is genome complexity? - it is the variety & number of DNA sequence copies in the genome
      2. In hybridization, complementary nucleic acid strands from different sources can be mixed to form double-stranded (hybrid) molecules – can be used to answer many questions
   C. Factors determining the rate of renaturation of a given DNA preparation
      1. Ionic strength of the solution
      2. Temperature
      3. DNA concentration
      4. Period of incubation
      5. Size of the interacting molecules

IV. Complexity of viral & bacterial genomes - compare renaturation rates of entire genomes of small virus (SV40; 5.4 x 10^3 bp), a larger virus (T4; 1.8 x 10^5 bp) & a bacterial cell (*E. coli*; 4.5 x 10^6 bp)
   A. The primary difference between these DNAs is their length – to compare their renaturation, it is important that the reacting molecules have the same length (typically ~1000 bp)
      1. Can fragment DNA into pieces of this length in various ways, one of which is to force all of the DNAs through a tiny orifice under high pressure to make them uniform in size at ~1000 bp
      2. Reanneal all of them at the same DNA concentrations (mg/ml) & same lengths —> they reanneal at distinctly different rates
   B. The smaller the genome is, the faster is the renaturation of that genome; the reason becomes apparent when one considers the concentration of complementary sequences in the three preparations
      1. Since all 3 preps have the same amount of DNA in a given volume of solution, it follows that the smaller the genome size, the greater the number of genomes present in a given weight of DNA
      2. This increases the chance of a collision between complementary fragments of small genomes
   C. Get same reannealing profiles regardless of whether the 3 DNAs reanneal in separate tubes or together in the same solution (rate of reannealing reaction is not affected by presence of unrelated sequences)
      1. Renaturation of viral & bacterial DNAs occurs along single, symmetrical curves since all (but for a very few sequences in bacterial DNA) are present at the same concentration

2. Thus, every nucleotide sequence in the population is as likely to find a partner in a given time as any other sequence

V. Complexity of eukaryotic genome - DNA fragments in same sample reanneal at very different rates

A. Roy Britten & David Kohne (Caltech) - mammalian genome DNA fragments reannealed at markedly different rates; not just one gene after another as seems to be the case in bacteria & viruses

B. The various nucleotide sequences in eukaryotic DNA fragments are present at very different concentrations; the first indication that eukaryotic DNA has a much more complex organization

1. When DNA fragments from plants and animals are allowed to reanneal, curves usually show 3 more-or-less distinct steps, corresponding to the reannealing of 3 broad DNA sequence classes
2. The 3 classes reanneal at different rates since they differ in the number of times their sequence is repeated within the fragment population
3. The greater the number of copies of a particular sequence in genome, the greater is its concentration & the faster is its reannealing rate

C. The classes - highly repeated, moderately repeated & nonrepeated (unique) DNA sequence fractions

~~The Structure of the Genome: Highly Repeated DNA Sequences~~

I. Traits – sequences present in at least 10^5 copies per genome; ~10% of total vertebrate DNA; to follow them, solution must be very dilute (& their concentration very low) since this fraction reanneals so fast

A. Usually short (a few 100 bp at most); seen in clusters, in which the sequence repeats over & over again without interruption

B. Sequences arranged in this end-to-end manner are said to be present in tandem (**tandem repeats**)

II. Highly repeated sequences fall into overlapping categories: satellite, minisatellite & microsatellite DNAs

A. Satellite DNAs – consist of short sequences (~5 – a few 100 bp long) repeated a large number of times in tandem to form very large clusters, each containing up to several million base pairs of DNA

1. In many species, base composition of this DNA is often sufficiently different from bulk of genome that it is easy to separate during density gradient centrifugation (form satellite band hence the name)
2. Can have >1 satellite sequence; *Drosophila virilis* has 3 different satellite sequences (each 7 nucleotides long & all very similar in sequence; indicating a common genetic origin)
3. Longer satellite DNAs don't form satellites since base makeup differs little from rest of genome
4. Satellite DNA has been localized within the centromeres of chromosomes, however, despite decades of research, the precise function of satellite DNA remains a mystery

B. Minisatellite DNAs – such sequences range from ~12 - 100 bp in length; found in clusters containing as many as 3000 repeats; occupy considerably shorter stretches of genome than satellite sequences

1. For an unknown reason, minisatellites tend to be unstable & the number of copies of a particular sequence often increases or decreases from one generation to the next
2. Thus, the length of a particular minisatellite locus is highly variable in the population, even among members of the same family
3. Since they are so variable (**polymorphic**), minisatellite sequences are used to identify individuals in criminal or paternity cases through the technique of DNA fingerprinting
4. Minisatellite sequence changes can alter expression (transcription) of nearby genes, which, in turn, can have serious outcomes; changes at certain minisatellite loci implicated in causing cancer, diabetes

C. Microsatellite DNAs – shortest sequences, 1 - 5 bp long; present in small clusters of ~10-40 bp in length

1. Scattered quite evenly throughout DNA (>100,000 different microsatellite loci in human genome)
2. DNA replicating enzymes have trouble copying genome regions that contain these small, repetitive sequences, which causes these stretches of DNA to change in length through the generations
3. Due to their variable lengths within population, they have been used to analyze relationships between different ethnic human populations
   a. Many anthropologists argue that modern human species arose in Africa
   b. If this is true, members of different African populations should have greater DNA sequence variation than humans on other continents, since African populations have had longer to diverge
   c. Argument for African genesis has been supported by studies on human DNA sequences
   d. One study analyzed 60 different microsatellite loci & showed that members of African populations had a significantly greater genetic divergence than Asian or European populations
4. Changes in the numbers of copies of certain microsatellite sequences are responsible for several debilitating inherited diseases

III. Where are these satellite sequences located? - *in situ* (in place) hybridization used to answer the question

A. Mary Lou Pardue & Joseph Gall (Yale) - developed *in situ* hybridization initially to localize satellite DNA
   1. DNA of the chromosomes is kept in place while it is allowed to react with a particular labeled DNA preparation; used to locate satellites on chromosomes
   2. Prepare mitotic cell chromosomes & spread them on slide; treat them with hot salt solution to separate the strands & keep them apart since the interacting DNAs must be single-stranded
   3. Do hybridization, incubate chromosomes with labeled, single-stranded satellite DNA (probe DNA; early labels radioactive) —> probe hybridizes to complementary sequences immobilized on slide
   4. After incubation, wash away soluble, unhybridized probe or digest it enzymatically, then detect with autoradiography —> satellite DNAs localized in chromosome centromeric region, but role unclear

B. **F**luorescent ***i****n* ***s****itu* **h**ybridization (FISH) – gives better resolution than autoradiography with radiolabel
   1. Attach biotin to probe (DNA or RNA); detect with fluorescent avidin (binds biotin with high affinity)
   2. Can be used to map the locations of specific sequences along single DNA fibers

~~The Structure of the Genome: Moderately Repeated and Nonrepeated DNA Sequences~~

I. Moderately repeated DNA sequences - ~20 - >80% of total DNA, depending on the organism; this fraction includes sequences repeated within genome from a few times to tens of thousands of times
   A. Includes sequences that lack a known coding function & those coding for known gene products (either RNAs or proteins)

   B. Repeated DNA sequences with coding functions - tRNAs, rRNAs, histone mRNAs (important chromosomal proteins); typically identical to one another & located in tandem array
      1. Need multiple copies of these genes, since the RNAs & histones for which they code are needed in large amounts
      2. tRNA & rRNA synthesis does not benefit from the extra amplification step that occurs for protein-coding genes in which each mRNA acts as a template for the repeated synthesis of a polypeptide
      3. Histone synthesis does involve an intermediary mRNA, but so many copies of this protein are needed during early development that several hundred DNA templates must be present

   C. Repeated DNAs lacking coding functions – the bulk of moderately repetitive DNAs codes for nothing
      1. Members of these families are scattered (interspersed) throughout genome as individual elements; most of these sequences can be grouped into 2 classes referred to as SINEs & LINEs
      2. SINEs (**s**hort **in**terspersed **e**lements) - typically <500 bp long; ex. in humans: *Alu*
      3. LINEs (**l**ong **in**terspersed **e**lements) - typically >1000 bp long; ex. in humans: L1
      4. Sequences of both groups of elements vary greatly from species to species

II. Nonrepeated (single-copy) DNA sequences - ~70% of human DNA fragments of 1000 bp length
   A. When denatured eukaryotic DNA is allowed to reanneal, a significant fraction of fragments are very slow to hybridize; they are presumed to be present as single copy/haploid genome (low concentration)
      1. Include Mendelian genes (they confirm Mendel's concept of one gene per haploid genome); always seen at particular site on particular chromosome
      2. Contain code for virtually all proteins but histones & the greatest amount of genetic information

   B. Genes coding for polypeptides are usually members of families or superfamilies of related genes, but the individual members of the families can differ a lot
      1. Globins, actins, myosins, collagens, tubulins, integrins, most other eukaryotic proteins are examples
      2. Each member of a multigene family is encoded by a different, but related, nonrepeated sequence

The Size of the Genome

I. Common sense dictates that the size of the genome (the amount of DNA in a haploid set of chromosomes) should increase in a roughly proportional manner with the complexity of the organism
   A. The size of the genome of an amoeba should be less than that of a salamander, which, in turn, should be less than that of a human; however, with the above examples just the opposite is true

1. Genomes of certain salamanders are roughly 30 times the size of the human genome
2. The genomes of certain amoebae are 5 times the size of those salamanders

B. For many years, these discrepancies in genome sizes from species to species were very puzzling
   1. The explanation is that different genomes have a tremendously variable number of repeated DNA sequences & most of these repeated sequences do not code for proteins
   2. For some reason, some species have a huge amount of extra, noncoding DNA
   3. Thus, the amount of DNA in a genome is not a measure of the number of genes it contains; in fact, humans are thought to have more genes than both amoebae & salamanders

C. The following quote says as much – "DNA appears **not** to be in proportion to the number of different genes in a cell."; the quote was remarkable for two reasons
   1. The quote was taken from a paper by Alfred Mirsky & Hans Ris of Rockefeller Institute in 1951
   2. This was one year before the Hershey-Chase experiment that nailed down DNA as the genetic material & 2 years before Watson & Crick published the structure of DNA

II. Genome sequence studies have shown that <1.5% of human genome codes for amino acids of our proteins; in 1960, this would have been thought ridiculous – how did remaining 98+% of DNA sequences evolve

The Stability of the Genome

I. The genome's sequence organization is capable of rapid change from generation to generation & within a given individual's lifetime, despite its reputation for slow change & stability over evolutionary time

II. Whole-genome duplication (polyploidization)

A. Most eukaryotes have pairs of homologous chromosomes in each of their cells; they have a **diploid** number of chromosomes
   1. Comparison of the number of chromosomes in closely related organisms, especially higher plants, shows that some species have a much greater number of chromosomes than a close relative
   2. The amphibian *Xenopus laevis* has twice the number of chromosomes as its cousin *X. tropicalis*

B. This is caused by **whole-genome duplication** or **polyploidization** which is an event in which offspring are produced that have twice the number of chromosomes in each cell as their diploid parents
   1. The offspring have 4 homologues of each chromosome instead of two

C. Polyploidization is thought to occur in either of two ways:
   1. Two related species mate to form a hybrid organism that contains the combined chromosomes from both parents (this mechanism occurs most often in plants), **or, alternatively**
   2. Single-celled embryo undergoes chromosome duplication but duplicates are not separated into separate cells, but are retained in single cell that develops into viable embryo (most often in animals)

D. A sudden doubling of chromosome number gives an organism remarkable evolutionary potential, assuming that it can survive the increased chromosome number & reproduce
   1. Polyploidization results in new species that has a great deal of "extra" genetic information

E. Several different fates can befall extra copies of a gene:
   1. They can be lost by deletion

2. They can be rendered inactive by deleterious mutations
3. Most importantly, they can evolve into new genes that possess new functions

F. Thus, extra genetic information is the raw material for evolutionary diversification – Susumu Ohno (City of Hope Cancer Center in Los Angeles, 1971)
1. Proposed in book that evolution of vertebrates from much simpler invertebrate ancestor was made possible by 2 separate rounds of whole-genome duplication during an early evolutionary period
2. He suggested that the thousands of extra genes generated by genetic duplication could be molded over time into new genes required to encode the more complex vertebrate body
3. Ohno's idea has been hotly debated for 30 years as people search for evidence to support or refute it

G. Mutation erodes ancestral genome face so data supporting Ohno's idea & human gene origin are hard to find – polyploidization evidence during early vertebrate evolution comes from *Hox* gene cluster analysis
1. *Hox* genes play a key role in the development of an animal's basic body plan
2. Invertebrates that have been studied have a single *Hox* gene cluster, whereas vertebrates have 4 such clusters, supporting Ohno's hypothesis of 2 rounds of ancestral genome duplication
3. Most other regions of the genome generally fail to show such evidence of large-scale duplication

H. Failure to show evidence of large-scale duplication in most of genome may be result of two processes (remains a subject of debate):
1. The failure may be the result of the erosion of nucleotide sequences over time, erasing evidence of whole-genome duplications **or**
2. A reflection of the fact that no such duplication has ever occurred

III. Duplication and modification of DNA sequences – more common than polyploidization, an extreme case of gene duplication that occurs only rarely during evolution

A. Three kinds of duplication can be distinguished
1. Whole-genome duplication
2. Gene duplication
3. Segmental duplication – duplication of a large block of chromosomal material (from a few kilobases to hundreds of kilobases in length); also has a significant impact in genome evolution
a. ~5% of present genome consists of segmental duplications that have arisen during the past 35 million years)

B. Gene duplication refers to the duplication of a small portion of a single chromosome; it happens with surprisingly high frequency & its occurrence is readily documented by genome analysis
1. A recent estimate - each gene in genome has ~1% chance of being duplicated every million years

C. Gene duplication can probably occur by several different mechanisms but is most often thought to be produced by a process of unequal crossing over
1. Unequal crossing over occurs when a pair of homologous chromosomes comes together during meiosis in such a way that they are not perfectly aligned
2. Due to misalignment, genetic exchange between homologues causes one chromosome to acquire an extra DNA segment (**duplication**) & the other to lose a DNA segment (**deletion**)

D. Most duplicates are either lost during evolution via deletion or made nonfunctional by unfavorable mutation; a small percentage acquire favorable mutations that allow extra copy to find a new function

E. Usually, the 2 genes have closely related sequences & encode similar polypeptides; they encode different isoforms of a particular protein, like α- & β-tubulin
1. Later duplication of one of the genes can lead to formation of more isoforms (γ-tubulin)
2. Gene duplication is thought to be responsible for the creation of gene families that encode polypeptides with related amino acid sequences

IV. Globin gene evolution (& a few others) - hemoglobin is tetramer of 4 globin polypeptides (2 pairs: 1 pair always in α–family, 1 in β-family); combinations differ with developmental stage (embryonic, fetal, adult)

A. Globin genes from different organisms (mammals, fish, etc.) have characteristic organization

1. Each gene is built of 3 **exons** (coding sequences) & 2 **introns** (noncoding intervening sequences)

2. Genes for certain globin-like proteins (leghemoglobin from plants; the muscle protein myoglobin) have 4 exons & 3 introns; it is proposed to represent ancestral form of globin gene

B. Proposed events in modern globin gene & polypeptide evolution
1. Modern globin protein may have arisen from the above ancestral forms by fusion of 2 of the globin exons (~800 million years ago)
2. Some primitive fish have only one globin gene, suggesting that these fish diverged from other vertebrates before the first duplication of the globin gene
3. After this duplication ~500 million yrs ago; the 2 copies of the gene diverged by mutation & formed 2 distinct globin types, an α-type & a β-type chain
4. α & β globin genes are thought to have been separated from one another by a process of rearrangement that moved them to separate chromosomes
5. Each gene then underwent later duplications & divergence generating today's organization of human globin genes; α-globin genes cluster on chromosome 16, β-globin gene cluster on chromosome 11

C. Analysis of globin gene cluster DNA sequences revealed "genes" homologous to those of functional globin genes with severe accumulated mutations that rendered them nonfunctional
1. Genes of this type are evolutionary relics (**pseudogenes**); their occurrence in genomes is widespread
2. Pseudogenes are found in both the human α- & β-gene clusters

D. Globin gene analysis also shows that much of the DNA in genes is noncoding DNA (introns within genes or spacers between genes), most of the noncoding regions have no known function

1. <15% of DNA located in these regions actually codes for globin polypeptide amino acid sequences

E. Sometimes, the polypeptides encoded by different members of a family have evolved divergent functions, even though their amino acid sequences still show their ancestral relationship
1. Growth hormone & prolactin are pituitary hormones with clearly related amino acid sequences, but they evoke completely different responses in target cells
2. Even if there is very little change in amino acid sequence of a certain protein over long periods of evolutionary time (e.g., actins), there may be substantial changes in nucleotide sequences
3. For example, substantial differences in nucleotide sequence are seen between corresponding genes of distantly related organisms, like the actins
4. Nucleotide substitutions that have been tolerated are those that do not alter the sequence of amino acids in the polypeptide chain

5. The same amino acid may be encoded by a number of different nucleotide triplet (codons); thus, the nucleotide sequence can change without changing the encoded amino acid sequence

"Jumping Genes" and the Dynamic Nature of the Genome

I. Repeated sequences sometimes present in tandem arrays, sometimes present on 2 or a few chromosomes & sometimes dispersed through genome - genome once thought to be stable information repository; not so

II. How do members of a gene family get dispersed to different chromosomes? - Barbara McClintock (Cold Spring Harbor geneticist, late 1940s; Nobel - 1983)
   A. She found certain mutations in traits shown by patterns & markings in maize leaf & kernel coloration (corn); her work was at first ignored, since the papers are hard to read & corn genetics is complex
      1. She found that some mutations were very unstable, appearing & disappearing from one generation to the next or even during the lifetime of an individual plant – late 1940s
      2. After several years, she concluded that certain genetic elements had moved from one place in a chromosome to an entirely different site, affecting gene expression
      3. Called the genetic rearrangement **transposition** & the mobile genetic elements **transposable elements**
   B. Molecular biologists working with bacteria saw no evidence of "jumping genes"
      1. Genes seemed to them to be stable elements situated in linear array on chromosome; the genes seemed to stay constant from one individual to another & one generation to next; she was ignored
   C. Eventually (late 1960s), bacteria were found to have DNA sequences that moved on rare occasions from one place in the genome to another & called these transposable elements **transposons**
      1. Transposons were found to encode a protein (**transposase**) that single-handedly catalyzes the excision of transposon from donor DNA site & its subsequent insertion at a target DNA site

III. Studies on bacterial transposition indicate that this "cut & paste" mechanism is mediated by 2 separate transposase subunits that bind to specific sequences at the 2 ends of the transposon
   A. The 2 subunits then come together to form an active dimer that catalyzes a series of reactions leading to the excision of the transposon
   B. The transposase-transposon complex then binds to a target DNA where the transposase catalyzes the reactions required to integrate the transposon into its new residence

IV. The sequence at one end of element is repeated at the other end in opposite orientation (**inverted repeat**)
   A. The terminal repeats are recognized by transposases & needed for transposition into target DNA
   B. These repeating sequences are found at ends of all transposable element types in bacteria & eukaryotes
      1. Integration of element creates small duplication in target DNA that flanks the transposed element at the insertion site
      2. These target site duplications can serve as footprints to identify genome sites that are or have been occupied by transposable elements

V. Eukaryotic genomes contain large numbers of transposable elements – at least 45% of DNA in human cell nucleus has been derived from transposable elements
   A. The vast majority (>99%) of these transposable elements cannot move from place to place; they have either been crippled by mutation or their movement is suppressed by the cell
   B. However, when transposons do change position, they insert almost randomly within target DNA
      1. Many transposable elements can insert themselves within the center of a protein-coding gene
      2. Several documented in humans – a number of cases of hemophilia result from a mobile genetic element that jumped into the center of one of the key blood-clotting genes
      3. ~1 in 500 human disease-causing mutations is the result of transposable element insertion

VI. Two major types of eukaryotic transposable elements & their differing mechanisms of transposition — DNA transposons & retroposons
   A. DNA transposons are excised from DNA at donor site & inserted into a distant target site
      1. This "cut-&-paste" mechanism is utilized by *mariner* family of transposons, which are found throughout plant & animal kingdoms
      2. Some are replicated & DNA copy is inserted into target site, leaving donor site unchanged (bacteria)
   B. **Retrotransposons**, in contrast, operate by means of "copy-&-paste" mechanism that involves an RNA intermediate; transposons requiring reverse transcriptase for movement are called retrotransposons
      1. DNA of transposable element is transcribed, producing an RNA, which is then reverse transcribed to DNA by reverse transcriptase, producing a complementary DNA
      2. The DNA copy is made double stranded & then integrated into a target DNA site

VII. In many cases, the retrotransposon itself contains the sequence coding for a reverse transcriptase
   A. Retroviruses, like the AIDS virus, use a very similar mechanism to replicate their genome & integrate a DNA copy into a host chromosome
   B. They may have evolved from retrotransposons by acquiring genes that allowed them to leave the cell & become infectious (e.g., genes encoding envelope proteins)

~~The Role of Mobile Genetic Elements in Evolution~~

I. Most of the moderately repeated sequences of the genome constitute a significant portion of the human genome; they are interspersed & arise by transposition of mobile genetic elements
   A. In contrast, the sequences of the highly repeated fraction of the genome (satellite, minisatellite, microsatellite DNA) reside in tandem & arise by gene duplication
   B. In fact, the 2 most common families of moderately repeated sequences in human DNA are transposable elements — the *Alu* & L1 families; both of which transpose by means of RNA intermediates

II. L1 – human genome has ~500,000 copies of L1; ~15% of total nuclear DNA, but the vast majority of these are incomplete elements & not capable of transposition
   A. Seen in all types of eukaryotes & may have been present in earliest eukaryotic cells
   B. In humans, a full length, transposable L1 sequence (at least 6000 bp long); encodes a unique protein with 2 catalytic activities:

1. Reverse transcriptase activity that makes a DNA copy of the RNA that encoded it
2. Endonuclease activity that cleaves the target DNA prior to insertion

III. *Alu* sequences - more abundant than L1 sequences; interspersed at more than 1 million different sites throughout the human genome
   A. Family of short, related sequences ~300 bp in length
   B. *Alu* sequence closely resembles that of the 7S RNA present in signal recognition particles found in conjunction with membrane-bound ribosomes
      1. It is presumed that during evolution, this cytoplasmic RNA was copied into DNA repeatedly by reverse transcriptase & the DNA copies were integrated into the human genome over generations
      2. The tremendous amplification of *Alu* sequence is thought to have occurred by retrotransposition using reverse transcriptase & endonuclease encoded by L1 sequences
      3. Due to its prevalence, one would expect it to be spread throughout the rest of the animal kingdom, but this is not the case
      4. Studies of various mammal genomes indicate that the *Alu* sequence first appeared as a transposable element in higher primate genomes ~60 million years ago (increasing in copy number ever since)
      5. The rate of *Alu* transposition has slowed, over the course of primate evolution, to its current estimated rate in humans of about once in every 200 births

IV. The function of transposable elements
   A. Many believe that transposable elements are primarily "junk" with no function
      1. A transposable element is a type of genetic parasite that can invade a host genome from the outside world, spread within that genome & be transmitted to offspring
      2. The above happens if it has no serious adverse effects on the host's ability to survive & reproduce
   B. Regardless of its origin, once a DNA sequence is present in the genome, it has the potential to be used in some beneficial manner during the course of evolution

V. The potential for their use as a key element of evolution has been realized in a number of ways:
   A. Transposable elements can carry adjacent parts of host genome with them as they move from one site to another, so 2 unlinked segments of host genome can be joined to form new, composite segment
      1. May be a primary mechanism in the evolution of proteins that are composed of domains derived from different ancestral genes
   B. DNA sequences that were originally derived from transposable elements are found as essential parts of eukaryotic genes
      1. In a recent study, the coding regions of ~1.3% of human genes contain inserts that are derived from *Alu* elements
      2. *Alu* sequences are only found in primate genomes, which indicates that these 400 or so *Alu*-containing genes have been modified substantially by transposition in the past 60 million years
      3. These findings underscore the importance of transposition in evolution & species divergence
   C. In some cases, transposable elements appear to have given rise to genes, themselves
      1. Telomerase (plays key role in replicating DNA at ends of chromosomes) is thought to be derived from a reverse transcriptase encoded by an ancient retrotransposon
      2. Enzymes involved in antibody gene rearrangement may be derived from a transposase encoded by an ancient DNA transposon

3. If this is the case, our ability to ward off infectious diseases is a direct consequence of transposition

VI. It is clear that transposition has had a profound impact on altering the genetic composition of organisms
   A. The transposable element of *Drosophila melanogaster* (P element) is an example of this
      1. Examination of laboratory fruit flies descended from individuals trapped by T. H. Morgan & his colleagues at the start of the 20th century are devoid of the P element
      2. In contrast, every member of the species caught in wild today has multiple copies of P element
      3. The P element is thought to have been introduced into the genome of a single *D. melanogaster* within past 80 years, probably by transmission from an individual of another *Drosophila* species
      4. Then it spread rapidly through the entire species population
   B. Transmission of genetic material from one species to another, whether between different fruit flies or different types of vertebrates, is likely mediated by parasites
      1. They pick up DNA fragments from one host & transfer it to a subsequent host

Sequencing Genomes: The Genetic Basis of Being Human

I. Determining the sequence of all of the DNA in a genome is a formidable task
   A. During 1980s & 1990s, technology to do this gradually improved
      1. Researchers developed new vectors to clone large DNA segments
      2. Increasingly automated procedures helped to sequence these large fragments
   B. The first complete sequence of a prokaryotic organism was reported in 1995; the first complete eukaryote sequence, the budding yeast *S. cerevisiae* was published in 1996
      1. Since then >100 different prokaryotes have been sequenced, along with those of a number of eukaryotes (fruit fly, a nematode, a fish, & several plants)
      2. These genomes of these organisms are considerably simpler & smaller than that of the human genome, which contains ~ 3 billion base pairs
      3. If each base pair in the DNA were equivalent to a single letter on this page, the information contained in the human genome would produce a book ~1 million pages long
   C. In 2000, the rough draft of the nucleotide sequence of the entire human genome was published
      1. The sequence was rough since each segment was sequenced an average of ~4 times, which is not enough to ensure complete accuracy
      2. The rough draft covered ~90% of the genome, excluding certain regions that proved difficult to sequence
      3. In 2001, came the first attempts to annotate the human genomic sequence (to interpret the sequence in terms of the numbers & types of genes it encoded)
   D. The most striking fact to emerge from the initial study concerned gene number; researchers concluded that the human genome probably contained around 30,000 protein-coding genes
      1. This is not much greater than the number in a fruit fly (~14,000 genes) or a nematode (~19,000 genes) & roughly equivalent to that of a pufferfish (~30,000 genes)
      2. Until sequencing had been completed, it had been widely assumed that the human genome contained at least 50,000 and maybe as many as 150,000 different genes
   E. The number of genes may be less than expected since a single gene can encode a number of related proteins as the result of alternate splicing
      1. >50% of human genes are thought to engage in alternate splicing

2. Thus, the actual number of proteins encoded by the human genome is several times greater than the number of genes it contains
3. It is likely that greater differences between organisms will emerge when this mechanism is explored
4. The difference in complexity between humans & other animals (particularly other vertebrates) is less a matter of the amount of genetic information inherited than what we do with it as we develop
5. For example, the gene expression control mechanisms may be more complex in humans than other animals, particularly as it pertains to brain development

II. 2003 – the nucleotide sequence of the entire human genome was published in its "finished" form, which means that:
   A. Each site had been sequenced on average ~10 times to ensure a very high degree of accuracy and
   B. The sequence contained a minimal number of small gaps (roughly 10 or so per chromosome, none larger than 150,000 base pairs
      1. The gaps represent chromosome regions consisting largely of long stretches of highly repeated DNA
      2. Despite exhaustive efforts, these regions have proven impossible to clone or their sequences have proven impossible to correctly order using current technology

III. Despite the achievement of sequencing the human genome, there is still uncertainty about the actual number of protein-coding genes in the human genome
   A. The true number is probably between 30,000 & 35,000, but that could still change – why is it so hard to come up with a firm number?
      1. It is difficult to determine whether or not a particular stretch of nucleotides in a DNA molecule contains a gene since it is often difficult to identify a gene
      2. This is related to the fact that genes consist of alternating coding regions (exons) & noncoding regions (introns)
      3. The coding regions are typically small (~150 bp in length), whereas the noncoding regions tend to be much larger, often several thousand base pairs
   B. Cells have no problem utilizing certain clues hidden in the DNA sequence to recognize where an exon begins & ends
      1. These clues are not always obvious enough for a gene-hunting computer program to pick out the exons & introns of a human gene simply by analyzing its nucleotide sequence
   C. The surest way to find a gene in a given DNA stretch is to know something about the encoded product
      1. Researchers know the amino acid sequence of thousands of different proteins
      2. Thus, the genes encoding these (& related) proteins are readily identified
      3. Even if the protein encoded by a gene has not been isolated & studied, it is likely that something is known about the mRNA transcribed from that gene
      4. Researchers can extract the entire mRNA population from a given tissue or organ & make a cDNA library
      5. It is a relatively straightforward process to use nucleotide sequences of cDNAs to identify the genes from which their mRNAs were initially transcribed
      6. Scientists are presently trying to obtain as many different full-length cDNAs as possible
   D. The genes for which researchers have a corresponding protein or cDNA are the easiest targets in the human genome to identify
      1. Genes that encode yet-to-be characterized proteins or those that are only transcribed during a brief embryonic development period for which cDNA libraries have not been obtained are hard targets

2. This makes estimates of the total number of human genes no more than educated guesses
3. Discovery of hard-to-find genes will probably depend heavily on comparison of human genome sequences with those of other mammalian species, like the mouse

Comparative Genomics

I. The protein-coding portion of the genome represents a remarkably small percentage of total DNA (~1.1%)
   A. The great majority of the genome consists of DNA that resides between the genes & thus represents intergenic DNA
   B. Each of the 30,000 or more protein-coding genes consists largely of noncoding portions (intronic DNA)

II. Most of the intergenic & intronic DNA of genome does not contribute to the reproductive fitness of an individual & thus is not subject to any significant degree of natural selection
   A. Thus, intergenic & intronic sequences tend to change rapidly as organisms evolve; these sequences tend not to be conserved
   B. Those portions of the genome that code for protein sequences & regulatory sequences that control gene expression are subject to natural selection & tend to be conserved among related species

III. Despite the fact that human & mouse species have not shared a common ancestor for ~75 million years, the two species share similar genes, which tend to be clustered in a similar pattern
   A. Distribution of human globin genes is basically similar to that found on a comparable segment of DNA in the mouse genome
      1. It is relatively simple to align corresponding regions of the mouse & human genomes
      2. Preliminary studies using this comparative approach suggest that ~5% of DNA sequences are highly conserved between mouse & human genomes
      3. This is considerably higher than would have been expected based solely on the predicted number of coding sequences & gene regulatory regions
      4. If it is true that conserved sequences must be important, then the studies say that parts of the genome presumed to consist of useless noncoding sequences actually have important, unidentified functions
   B. Some of these regions undoubtedly encode recently discovered, small, noncoding RNAs whose function remains to be determined
      1. Other regions may have chromosomal functions instead of genetic functions
      2. These conserved sequences could be important for chromosome pairing prior to cell division
      3. The presence of these extra conserved regions will make it harder to identify the coding sequences & regulatory regions that comparative genetics was expected to reveal

   C. Primary goal of geneticists centers on gene function
      1. Once a human gene is identified by its sequence similarity to mouse gene, some idea of the function of that gene may be ascertained by studying phenotypes of mice in which gene is incapacitated
      2. For example, if elimination of a particular mouse gene leads to birth of deaf animals, then it is likely that the human counterpart to this gene is involved in some way in the process of hearing

IV. Plans have been made to sequence the chimpanzee genome which is ~98.5% similar in overall nucleotide sequence to that of human (although recent evidence indicates the differences may be greater)

A. Chimpanzees are thought to be our closest living relative having shared a common ancestor as recently as 4.5 – 6.2 million years ago

1. Detailed study of relatively small DNA sequence & gene organization differences between chimps & humans may say much about genetic basis for recently evolved features that make us uniquely human
2. The human brain has a volume of ~1300 cm^3 (nearly 4 times that of a chimpanzee)
3. Common human epithelial cancers (like breast & colon cancer) are only rarely seen in chimps

B. The *FOXP2* gene – comparison of *FOXP2* in humans & chimps shows 2 amino acid differences that have appeared in the human lineage since the time of separation from our last common ancestor

1. Persons with mutations in *FOXP2* gene suffer from a severe speech & language disorder
2. Among other deficits, persons with this disorder are unable to perform the fine muscular movements of lips & tongue that are required to engage in vocal communication
3. Calculations suggest changes in this "speech gene" that distinguish it from the chimp version were fixed in human genome in past 120,000 - 200,000 years when modern humans may have emerged
4. These findings suggest that changes in the *FOXP* gene may have played an important role in human evolution

~~The Human Perspective: Diseases That Result from Expansion of Trinucleotide Repeats~~

Background on Trinucleotide Repeat Expansion Diseases

I. For decades, biologists thought that genes were transmitted from generation to generation as stable entities

A. On rare occasions, a change occurs in the nucleotide sequence of a gene in the germ line, creating a mutation that is subsequently inherited

B. 1991 – several labs reported a new type of dynamic mutation in which the nucleotide sequence of particular genes changed dramatically between parents & offspring

1. These mutations affected genes that contained a repeating trinucleotide unit (e.g., CCG or CAG) as part of their sequence
2. Usually, the genes contain a relatively small, but variable, number of repeated trinucleotides; they are transmitted from one generation to the next without a change in the number of repeats
3. A small fraction of the population possesses a mutant version of the gene that contains a larger number of these repeating units
4. Unlike the normal genes, the mutant alleles are highly unstable & the number of repeating units tends to increase as the gene passes from parents to offspring
5. When the number of repeats increases beyond a critical number, the individual inheriting the mutant allele develops a serious disease

II. >12 distinct diseases have been attributed to the expansion of trinucleotide repeats; they fall into two basic categories

A. Type I diseases – all are neurodegenerative disorders resulting from expansion of the number of repeats of CAG trinucleotide within the mutant gene's coding portion – Huntington's disease (HD) is example
B. Type II diseases – trinucleotide repeat diseases that differ from Type I diseases in a number of ways; best studied is fragile X syndrome

Type I Trinucleotide Repeat Expansion Diseases – Huntington's Disease

I. HD is characterized by involuntary, uncoordinated movements, changes in personality (including depression & irritability) & gradual intellectual decline
   A. Symptoms usually begin in the third to fifth decade of life & increase in severity until death
II. Normal *HD* gene, which is stably transmitted, contains between 6 & 35 copies of the CAG trinucleotide, the triplet that codes for the amino acid glutamine
   A. The protein encoded by the *HD* gene is called **huntingtin**; its precise function remains unclear
   B. The normal huntingtin polypeptide contains a stretch of 6 – 35 glutamine residues, a polyglutamine tract, as part of its primary structure
      1. While most polypeptides have a highly defined primary structure, huntingtin is normally polymorphic with respect to the length of its polyglutamine tract
      2. The protein seems to function normally as long as tract length stays below ~35 glutamine residues
      3. If this number is exceeded, protein has new properties & person is predisposed to developing HD

III. HD (& the other trinucleotide degenerative diseases) exhibits a number of unusual characteristics
   A. Unlike most inherited diseases, HD is a dominant genetic disorder; a person with the mutant allele will develop the disease regardless of whether or not he or she has a normal *HD* allele
      1. Persons who are homozygous for the *HD* allele are no more seriously affected than heterozygotes
      2. Suggests that mutant huntingtin polypeptide causes the disease, not because it fails to carry out a particular function, but because it acquires some type of toxic property (**gain-of-function mutation**)
      3. Gain-of-function mutation interpretation supported by mouse studies - mice engineered to carry mutant *HD* allele (along with normal alleles) develop neurodegenerative disease like HD in humans
      4. The presence of the one abnormal allele is sufficient to cause disease
   B. Another unusual characteristic of HD & other CAG disorders is phenomenon called **genetic anticipation**
      1. As the disease is passed from generation to generation, its severity increases & it strikes at an increasingly early age
      2. Once puzzling, it is now explained by the fact that the number of CAG repeats in a mutant allele (& the resulting consequences) often increases dramatically from one generation to the next

IV. Molecular basis of HD remains unclear
   A. One thing is sure: when polyglutamine tract of huntingtin exceeds 35 residues, the protein (or a fragment cleaved from it) undergoes abnormal folding to produce a misfolded molecule that:
      1. Binds to other mutant huntingtin molecules (or fragments) to form insoluble aggregates, not unlike those seen in the brains of Alzheimer's victims, **and**

2. Binds to a number of unrelated proteins that do not interact with normal, wild-type huntingtin molecules

B. Among the proteins bound by mutant huntingtin are several transcription factors (proteins involved in the regulation of gene expression)
   1. Several of the most important transcription factors found in cells, including TBP & CBP, contain polyglutamine stretches themselves
   2. This makes them particularly susceptible to aggregation by proteins with mutant, expanded polyglutamine tracts
   3. The protein aggregates present in degenerating neurons of HD patients contain both of these key transcription factors
   4. This suggests that mutant huntingtin sequesters these transcription factors, thus removing them from the remainder of the nucleus
   5. This disrupts the transcription of genes that are needed for health & survival of affected neurons

C. This hypothesis has received support from a study in which mice were genetically engineered so that their brain cells lacked the ability to produce CREB (& a related transcription factor CREM)
   1. The mice showed the same type of neurodegeneration seen in animals carrying a mutant *HD* gene
   2. The part of the brain affected by HD (the striatum) is particularly dependent for survival upon CREB-dependent transcription

D. An alternate hypothesis postulates that it is not protein aggregates that are toxic, but the soluble mutant protein itself; some believe that the aggregates protect the cell by sequestering harmful molecules

V. 2003 – encouraging report on a strain of mice carrying a mutant *HD* allele that led to neurodegeneration & premature death

A. Congo red is a molecule that has been used for decades as a dye to stain tissues, but it also has the capability to block interactions between misfolded proteins

B. Treatment of mutant *HD* mice with Congo red led to marked improvement in their health
   1. Treated them with Congo red after they demonstrated unmistakable symptoms of neurodegeneration
   2. They showed improvements in motor function, dispersal of existing protein aggregates in affected regions of brain & prolonged survival
   3. Provides hope that polymerization of proteins containing expanded polyglutamine repeats may be inhibited & that progression of these diseases may be slowed

Type II Trinucleotide Repeat Expansion Diseases

I. Type II trinucleotide repeat expansion diseases differ from the Type I diseases in a number of ways

A. They arise from the expansion of a variety of trinucleotides, not only CAG

B. The trinucleotides involved are present in a part of the gene that does not encode amino acids

C. The trinucleotides are subject to massive expansion into thousands of repeats

D. The diseases affect numerous parts of the body, not only the brain

II. Best studied Type II disease is fragile X syndrome, so-named because the mutant X chromosome is especially susceptible to damage

A. Fragile X syndrome is characterized by mental retardation & a number of physical abnormalities

B. It is caused by a dynamic mutation in a gene called *FMR1* that encodes a protein thought to bind certain mRNAs involved in neuronal development and/or synaptic function
   1. Normal allele of gene contains anywhere from ~5 to 55 copies of a specific trinucleotide (CGG) that is repeated in a part of the gene that corresponds to the 5' noncoding portion of the mRNA
   2. A person can carry up to ~200 copies of this triplet without showing adverse effects
   3. However, once the number of copies rises above ~60, the locus becomes very unstable & the copy number tends to increase rapidly into the thousands

C. Persons with an *FMR1* gene containing 60 – 200 copies of the triplet generally exhibit a normal phenotype, but they are carriers for the transmission of a highly unstable chromosome to offspring
   1. If the repeat number in offspring rises above ~200, the individual is almost always mentally retarded

D. Unlike abnormal *HD* gene (causes disease as result of gain-of-function), an abnormal *FMR1* allele causes disease as a result of a loss of function
   1. *FMR1* alleles containing an expanded CGG number are selectively inactivated so that the gene is not transcribed or translated

E. Although there is no effective treatment for any diseases caused by trinucleotide expansion, the risk of transmitting or possessing a mutant allele can be assessed through genetic screening

LECTURE HINTS

Early Studies in Genetics

If time permits, it is useful to describe briefly the early studies that put us on the road to understanding inheritance, a road along which we are still traveling. Outline the studies of Mendel (Laws of Segregation and Independent Assortment), Boveri, van Beneden, Weismann, Sutton and Morgan. Talk about the discovery of chromosomes and their importance, linkage, recombination, crossing over, mutagenesis and polytene chromosomes. You may decide to minimize this part of the lecture, leaving it for the Genetics course that often follows a Cell Biology course.

The Chemical Nature of the Gene

This material has been discussed in Chapter 2. I will not reiterate it here. You may wish to recommend outside reading to give your students an idea of the climate and politics in the scientific community when these discoveries were being made. The Double Helix by James Watson comes to mind as does The Eighth Day of Creation by Horace Freeland Judson. Both are excellent presentations of the science and personalities involved. Watson has recently published a new book called The DNA Story, which I recommend highly.

The Structure of the Genome

Describe the process of DNA denaturation. Emphasize the effect that base composition can have on the melting temperature. Describe hypothetical DNAs to the students - one with high GC content, the other with high AT content. Ask them which sample melts at the highest temperature and ask them to explain their answer.

Describe renaturation and outline for your students the process of renaturation as well. Point out the factors that influence renaturation. Ask questions (leading ones, if necessary) about these factors. Get your students to explain those effects.

Define the meaning of complexity with respect to the genome. Differentiate between the three categories of DNA sequences found in eukaryotes. Highly repetitive DNA sequences are usually short (a few hundred base pairs at most) and found in clusters called tandem repeats, where they repeat over and over in an uninterrupted fashion. They make up about 10% of total vertebrate DNA. Some of these are known as satellite DNAs because their sequences are short and different in base composition from the bulk of the organism's DNA. When the organism's DNA is analyzed by density gradient centrifugation, satellite DNAs form bands separate from most of the DNA in the sample. Ask the students why this DNA is called satellite DNA. Describe the technique of *in situ* hybridization used to localize such DNAs on chromosomes. Ask leading questions to get your students to "figure out" how the technique works rather than simply telling them. Also, differentiate between satellite, minisatellite and microsatellite DNAs. Describe the moderately repetitive sequences that can compose ~20 - 80% of the genome depending on the organism. Point out the types of DNA sequences included, such as those that code for tRNAs, rRNAs and histone mRNAs and those that lack any known coding function like the SINEs and LINEs. Finally, point out the nonrepeated (unique) sequences that can make up around 70% of the genome in humans. Their function is to code for virtually all proteins other than histones. Ask students which of the three types of eukaryotic DNA sequences is most like the type found in bacteria and viruses.

The Stability of the Genome

Describe the mechanism by which gene duplication can occur and how that has played a role in the construction of gene families and superfamilies. Once copies of genes have been made, they can experience different base substitutions and will, therefore, exhibit divergence in their sequences. This would appear to be a mechanism of evolution and would explain the existence of genes with related sequences but different functions. If there is time, discuss the globin gene family. Point out the frequent dispersion of members of particular gene families throughout the genome, although duplication by unequal crossover would not seem to explain such dispersion.

Mobile Genetic Elements

The question of dispersion of duplicated genes was raised above. The discussion of mobile genetic elements answers that question. Barbara McClintock was a woman ahead of her time. She made a discovery that mainstream investigators did not understand or believe. It demonstrates how accepted scientific ideas can sometimes be difficult to alter even in the face of significant information that contradicts them. As in the case of McClintock's "jumping genes", years may go by before an investigator running counter to established opinion is vindicated. Her story is an important one. A recent biography tells it well.

Outline the characteristics of transposable elements (transposons) and their effects on the genome. Point out the differences between eukaryotic and prokaryotic transposons. Discuss the mechanisms by which transposons duplicate and move DNA sequences to other locations in the genome. Emphasize the medical significance of transposons in terms of growing antibiotic resistance in microorganisms and the development of some cancers. For a discussion of this problem and other related aspects of infectious disease, I recommend The Coming Plague by

Laurie Garrett. It is a compelling work outlining the societal, medical and economic impacts that microorganisms and viruses will have on humans in the coming years. Also, point out the role that transposons are thought to have played (and may continue to play) in evolution.

Having said all this, the topic of transposons may be one topic that you may decide to skip so that it can be covered in a Genetics course.