Chapter 14: The Eukaryotic Genome and Its Expression

I Eukaryotic Genome

See *Table 14.1* for a list of differences between prokaryotic and eukaryotic genomes.

A. The eukaryotic genome is larger and more complex than the prokaryotic genome

- Viral genomes are smaller than bacterial genomes.
- Bacterial genomes are smaller than eukaryotic organisms' genomes.
- Among eukaryotes, there is not always a direct relationship between complexity and genome size.
  - Humans have 6 billion base pairs in each cell's nucleus.
  - The lily, which produces fewer different proteins than a human, has 18 times more DNA.
  - *See Figure 14.1.*
- Most eukaryotic DNA codes for nothing.
  - Interspersed throughout the genome are various repeated sequences that are not transcribed.
  - Even within genes, are sequences that are not translated.
  - Some of these nontranscribed regions are structural, such as the telomeres; some regulate gene expression, and some have no known use.
- Eukaryotes have several, separate, linear chromosomes.
  - Each eukaryotic chromosome has a single, continuous, double helix of DNA.
  - Each is from 20 million to 100 million base pairs in length.
  - Each eukaryotic chromosome must have origins of replication, a centromeric sequence and telomeric sequences on the ends.
- Chromosomes are housed in the nucleus.
  - RNA synthesis occurs in the nucleus.
  - RNA translation occurs exclusively *outside* the nucleus.
- The DNA is packaged into nucleosomes.
- Genes have regulatory sequences that are not transcribed.
  - Promoter regions are where RNA polymerase binds.
  - A second set of regulatory DNA sequences, the enhancers and silencers are common in eukaryotes, but rare in prokaryotes.
  - They can be located quite far away from the promoter and structural gene.
- Noncoding DNA, scattered within protein coding portion of genes, exists. These get transcribed, but are spliced out of the pre-mRNA prior to transport out of the nucleus. *See Figure 14.2.*
- The entire genomes of yeast, roundworms and the fruit fly have been sequenced. In 2000, the whole human genome had been sequenced (roughed in).

B. The yeast genome adds some eukaryotic functions onto a prokaryotic model

  - *Saccharomyces cerevisiae* has 16 chromosomes and more than 12,068,000 base pairs.
  - The sequencing has found 6,200 genes. Around 70% now have probable roles assigned.
  - The other 30% are being studied by gene inactivation techniques.

- About 11% of yeast proteins are for general metabolism; 3% for energy production and storage. Three percent are for DNA replication and repair; 12% for protein synthesis; and, 6% for protein targeting and secretion.
    - One difference between the *E. coli* and yeast is the number of genes for protein targeting.
    - This is because of the compartmentalization found in yeast and other eukaryotic organisms. *See Table 14.2.*

C. The nematode genome adds developmental complexity

- *Caenorhabditis elegans* is a 1-mm long nematode (round worm) that lives in soil.
- In 3 days it develops from an egg to an adult.
- Its transparent body makes it possible to observe its development.
- The adult has 1000 cells. It has a nervous system, digestive system, reproductive system, and it ages.
- *C. elegans* has been intensely studied, especially in regards to its development. Its genome has now been completely sequenced.
    - It consists of around 97 million base pairs, and has about 19,000 protein coding genes.
    - Before sequencing, scientists guessed it would have about 6000 genes.
- About 3,000 genes in the worm have counterparts in common yeast.
    - These are considered to be the ones essential to all eukaryotes.
    - *See Table 14.3.*

D. The fruit fly genome has surprisingly few genes

- *Drosophila melanogaster* is much larger than *C. elegans*.
- Its genome contains 180,000,000 base pairs.
- Surprising was that *Drosophila melanogaster* has fewer genes than *C. elegans*, 13,600 genes.
- One reason is that *C. elegans* has more copies of related genes than Drosophila.
- Many of the genes found in the worm have homologous genes to those of the fly.
- About half of the fly genes have mammalian homologs.
    - The fly genome contains 177 genes, whose sequences are known to be directly involved in human disease.
    - Because of the simpler genomes, the role of these genes are often easier studied in the fly.

E. Gene sequences for other organisms are rapidly becoming known

- A rough human genome sequence has recently been completed.
- A weedy plant *Arabidopsis thaliana* (130 million base pairs) is progressing.
- Over the next decade tremendous progress will occur.

II Repetitive Sequence in the Eukaryotic Genome

A. Highly repetitive sequences are present in large numbers of copies

- Three types of highly repetitive sequences are found in eukaryotes:

    - Satellites are 5-50 base pairs long, repeated side by side up to a million times.
    - Minisatellites are 12-100 base pairs long and repeated several thousand times. They are variable in the number of copies between individuals in a population.
    - Microsatellites are very short (1-5 base pairs) present in 10-50 copies per cluster. They are scattered all over the genome.
- The useful purpose for repetitive sequences, if any exists, is currently unknown.

B. Telomeres are repetitive sequences at the ends of chromosomes

- Telomeres are sequences at the end of the chromosomes.
- They are moderately repetitive sequences.
- Telomeres are specialized sequences, generated by an enzyme called telomerase.
- A problem exists for the replication of linear DNA molecules, like those of eukaryotic chromosomes.
    - There is nothing beyond the primer in the 5' direction to replace the RNA that makes up the primer. *(DNA polymerase can only add to an existing molecule.)*
    - Their new chromosomes lack a bit of double-stranded DNA at each end.
    - These ends get trimmed, so chromosomes get shorter each cell division.
- The human telomere repeated sequence is TTAGGG
    - It is repeated around 2,500 times at each end.
    - These repeats bind special proteins that protect the ends.
    - Telomeres can be elongated by an enzyme called telomerase. *See Figure 14.3.*
    - Telomeres shorten from 50 to 200 base pairs per cell division.
    - Telomerase activity occurs only during early embryonic periods of development in many human cell types.
    - As cells divide, their telomeres get shorter each time. After 20-30 divisions, the cells senesce.
    - Aging cells in culture have been demonstrated to have finite life spans in terms of the number of divisions.
    - Cells in culture have been experimentally altered to produce telomerase, and such cells become immortal.
    - Human cancer cells are also immortal, and have been found to express telomerase.

C. Some moderately repetitive sequences are transcribed

- Some moderately repetitive DNA sequences code for tRNA's and rRNA's.
- Having copies of coding regions makes it possible to produce these RNA's more, quickly.
- In mammals there are four different rRNA molecules that make up the ribosome – 18S, 5.8S, 28S, and 5S rRNA's.
- The 18S, 5.8S and 28S rRNA's are transcribed as a single precursor RNA, which is twice the size of all three ultimate products.
- *See Figure 14.4.*

- There are 280 copies of sequences coding for the transcript located in clusters on five different chromosomes.
- Another moderately repeated sequence, which is not known to be transcribed, is the *Alu* family.
  - These are scattered throughout the genome.
  - They are 300 bp long.
  - There are 300,000 copies of the *Alu* family.
  - They may act as multiple origins of replication.

D. Transposable elements move about the genome

- Some moderately repetitive DNA are transposons.
- There are four main types:
  - *SINEs* are short interspersed elements and are up to 500 base pairs long. They are transcribed but not translated.
  - *LINEs* are long interspersed elements. They are 7,000 base pairs long, and some are transcribed and translated into proteins. They constitute about 15% of the human genome.
  - Both of these elements are present in more than 100,000 copies.
    - They move about the genome by making an RNA copy, which acts as a template for the new DNA.
    - The new DNA then inserts itself at a new location.
  - Retrotransposons make an RNA copy also.
    - These are rare in mammals but common in other animals and in yeast.
    - They resemble HIV, but code for no protein coat.
  - DNA transposons do not use an RNA intermediate, but actually move to a new spot without replicating.
- Any beneficial role for these sequences is unknown.
- Detrimental effects have been found.
- Insertion into a functional gene can disable it. In some cases, insertions near a gene can alter its transcription rate. *See Figure 14.5.*
- When an insertion occurs in germ cell line, a gamete carrying the new mutation might form.
- If it occurs in a somatic cell, cancer might result.
- Transposition certainly increases genetic variation.

III The Structures of Protein-Coding Genes

- Eukayrotes have genes with non-coding internal sequences.
- Eukaryotes form gene families with structurally and functionally related cousins in the genome.

A. Protein-coding genes contain non coding internal and flanking sequences

- *See Figure 14.6.*
- Within the initially produced transcript are noncoding sequences, called introns.
- They are interspersed in the coding regions called exons.
  - First the transcript is created, which is called primary transcript, as a pre-mRNA.
  - This is processed into an mRNA.

- Nucleic acid hybridization was useful for the initial discovery of introns.
    - Hybridization between cDNA and the genomic gene sequence that coded globin showed a bubble of DNA not present on the cDNA.
    - *See Figure14.7 and 14.8.*
- Most but not all genes of vertebrates contain introns. Other eukaryotes also have introns.

B. Many eukaryotic genes are members of gene families

- About half of all eukaryotic protein-coding genes are single-copy.
- The rest have multiple copies per haploid genome.
- Pseudogenes are non-functional copies, often found near the functional copy.
- Sometimes copies of genes are functional, but slightly different.
- A set of duplicated or related genes is called a gene family.
    - Immunoglobins have hundreds of members.
    - β-globins have a few members.
- Copies change over time, which provide an opportunity to make useful new versions.
- *See Figure 14.9.*
    - During development, different members of the β-globin gene family are expressed at four different times.
    - See Figure 14.10.
    - Each is best suited for the time when they are expressed.
    - There are also pseudogenes, which are part of the β-globin gene family.

IV RNA Processing

See Figure 14.11

A. The primary transcript of a protein coding gene is modified at both ends

- The pre-mRNA gets modified at the 5' and 3' end.
    - The G cap, a modified guanosine triphosphate is added to the 5'end.
    - A poly A tail is added to the 3' end. It is 100-300 residues in length.
    - *See Figure 14.11*

B. Splicing removes introns from the primary transcript

- This process is illustrated in *Figure 14.12*.
- At the boundaries between introns and exons are consensus sequences. This helps recognize the introns.
- The spliceosomes, composed of RNA and protein, require energy acquired from ATP to cut and rejoin the RNA.
    - There is a human disease that is caused by a mutation at the consensus sequence of the β-globin gene.
    - It causes an inadequate supply of red bloodcells.

V Transcriptional Control

- Regulation of gene expression can occur at many points.
- *See Figure 14.14.*

A. Specific genes can be selectively transcribed

- Different cells in multicellular organisms produce some of the same proteins, and also some that are unique to each type.
- Both cell types have the same DNA sequences, but express genes differently.
  - For example, the genes that both brain and liver cells transcribe are called housekeeping genes.
  - Housekeeping genes code for enzymes and proteins essential to the survival of all cells, enzymes like those needed for glycolysis and ion transporters.
  - Muscle cells have characteristic proteins as do other cell types.
  - The difference in the presence of proteins is due to differential transcription.
- Unlike prokaryotes, eukaryotes tend to have solitary genes, not operons.
- Coordinated regulation of gene expression for activities such as biochemical pathways requires common control elements in each of the genes. These allow the genes to respond to the same signal.
- Eukaryotes have three different RNA polymerases.
  - RNA polymerase II transcribes protein-coding genes.
  - RNA polymerase I transcribes rRNA coding sequences.
  - RNA polymerase III transcribes tRNA and small nuclear RNA's
- Most eukaryotic genes have other DNA sequences that regulate transcription: regulators, enhancers and silencers.
- Many different proteins are involved in initiating transcription.
- *Transcription factors:*
    - Prokaryotes have a recognition sequence about 40 base pairs from the 5' initiation point and a TATAAT box just upstream from the intersection point.
    - In contrast, eukaryotes, the TATAAT box is about 25 base pairs away from the initiation site, and one or two recognition sequences, which are 50-70 base pairs upstream of TATAAT.
  - Transcription in eukaryotes requires transcription factors.
    - *See Figure 14.14.*
    - RNA polymerase II does not bind until several other proteins have already bound the protein–DNA complex.
    - Some sequences are common to all promoters; others unique.
- *Regulators, Enhancers, and Silencers:*
    - Regulator regions affect transcription in addition to those associated with the promoter.
    - Regulator proteins bind these regions.
  - Much farther away, up to 20,000 base pairs away, are the enhancer regions.
    - Enhancers bind activator proteins.
    - These strongly stimulate the transcription complex.
  - There also are negative regulatory regions of DNA called silencers.
    - They have the opposite effect of enhancers.
    - Silencers bind repressors.
  - *See Figure 14.15.*
  - All genes in most tissues transcribe a small amount of RNA.

    - The correct combination of factors determines the maximum rate of transcription.
    - Rates can be modulated by varying concentrations of factors.
- *Coordinating the expression of genes:*
    - Regulation of various genes can be coordinated if all have the same regulatory sequences near them.
    - One example is stress response elements, which coordinate expression of *heat shock* genes.
    - *See Figure 14.16.*
- *The binding of proteins to DNA:*
    - Key to transcription regulation is the binding of protein to DNA.
    - Proteins need to recognize and bind appropriate sites.
    - There are four different styles or motifs for protein-DNA interactions.
    - The helix-turn motif involves several α-helices, one of which makes contact with DNA.

<<Can the figure on 273 “Helix-turn-helix motif be inserted here? There is no figure number to be referenced.>>

    - The motif is common in genes involved in embryonic development.
    - The zinc finger motif has a loop that form when a zinc ion is held by the amino acids cysteine and histidine.
    - It is common for steroid hormone receptors.

    << Can the figure on 273 zinc finger motif be inserted here? >>

    - Leucine zipper motif has hydrophobic leucine residues on one side of a polypeptide.
    - Two polypeptide chains interact (zipper) hydrophobically.
    - Positively charged residues just past the zipper interact with the DNA.

    << Can the figure on 273 Leucine zipper motif be inserted here? >>

    - This is a common DNA-binding motif.
    - Overactivity of one, AP-1, has been linked to several types of cancer.
    - The helix-loop-helix motif is two helices separated by a loop. Two adjacent regions associate with the DNA.

    << Can the figure on 273 helix-loop-helix motif be inserted here? >>

    - This motif occurs in the activator protein that binds to enhancers for immunoglobulin genes and factors involved in muscle protein synthesis.

B. Genes can be inactivated by chromatin structure

- Packaging of DNA by nuclear proteins can make DNA physically inaccessible to RNA polymerase and associated components.
- Local effects:
    - Nucleosomes inhibit both initiation and elongation of transcription.
    - These effects are countered by two protein complexes.
    - One binds upstream of the initiation site, disaggregating the nucleosomes, allowing the large initiation complex to form.
    - The other bind after transcription begins, allowing the transcription complex to move through these nucleosomes.
    - *See Figure 14.17.*
- Global Effects:
    - Euchromatin is diffuse and stains lightly.
    - Heterochromatin stains densely and is generally not transcribed.

- X – chromosome inactivation is an example.
- In mammals just one of the X – chromosomes in each cell of a female is activated.
- This creates a gene dosage equivalence between the cells of XX females and XY males.
- Early during development one of the X – chromosomes in each of the cells of a female embryo is inactivated.
- Which one is inactivated is simply chance. However, once inactivated, all progenitors of the cell have the same X inactivated.
- This was discovered in 1961 by three scientists; Mary Lyon, Liane Russell and Ernest Beutler.
- Interphase cells of XX (normal) females have a single, stainable nuclear body called a Barr body.
- *See Figure 14.18.*
- This is the inactive X.
- An abnormal XXXX female will have 3 Barr bodies. An abnormal XXY male has one. Addition of a methyl group ($-CH_3$) to the 5 carbon of cytosine on DNA is involved with inactivation.

<<Page 274, first line of text on the right columns says 5' position. This is certainly a mistake. It is the 5 position of the pyrimidine, cytosine, which is numbered without a ' mark. The ' mark applied to the deoxyribose carbons'.>>

- The inactive X has one less methylated gene, *XIST*, while the rest are heavily methylated, and, except for the XIST gene, are not transcribed.
- The other X is active.
- The transcript from *XISTt* is an RNA that remains in the nucleus. It binds the X chromosome and triggers inactivation.

C. A DNA sequence can move to a new location to activate transcription

- *Saccharomyces cerevisiae*, a yeast species, has two mating types, a and α.
- Physically separate sites exist: the a, α and MAT regions.
- Mating types depends on which copy a or α exists at the MAT site.
- It is the one at MAT that is expressed.
- Transportation can occur from a or α to the MAT site, which switches the mating type.

D. Selective gene amplification results in more templates for transcription

- The process of increasing the number of copies of a gene is called gene amplification.
- Mature frog and fish eggs have up to a trillion ribosomes.
- To make this number, ribosomal rRNA gene clusters are copied until in just one cell there are a million copies.
- These sequences go from being 0.2% to 68% of the total genomic DNA.
- Later, after cell division begins, the number of copies returns to normal.

VI Posttranslational Control

Pre-mRNA must be processed in the nucleus before it can be translated in the cytoplasm.

A. Different mRNA's can be made from the same gene by alternate splicing

- Alternate splicing of a specific pre-mRNA can generate different proteins.
- For example, a single pre-mRNA for the structural protein tropomyosin, is alternatively spliced into five different mRNA's and five different forms of tropomyosin found in five different types of tissues.
- Depending on the tissue: skeletal, muscle, smooth muscle, fibroblast, liver or brain; tissue specific splicing occurs to generate the appropriate form of tropomyosin.
- *See Figure 14.20.*

B. The stability of mRNA can be regulated

- After entering the cytoplasm, mRNA is subject to digestion by ribonucleases.
- Different mRNA species have different half lives.
- An interesting example of cellular modulation of mRNA half-life is with tubulin.
- When the tubulin concentration is high, some tubulin molecules bind tubulin mRNA, accelerating tubulin mRNA breakdown.

VII Translational and Postranslational Control

- Only a third of the mRNA species in yeast have a positive correlation between the number of mRNA molecules coding for a protein and the amount of that protein synthesized.
- Two-thirds showed no relationship.
- Proteins can affect translation.

A. The translation of mRNA can be controlled

- Before mRNA can be transcribed in eukaryotes, it must have a modified 5' guanosine molecule.
- Examples of mRNA that are stored by delaying modification of the G cap exist in the oocytes of the tobacco hornworm moth.
- When the mRNA is needed, post-fertilization, the cap gets modified and the mRNA's are transcribed.
- Ferritin, an iron storage protein, increases with elevation in Fe^{2+}.
- The amount of ferritin mRNA remains constant.
- When iron is low, a translational repressor protein binds to ferritin mRNA.
- When iron levels rise, excess iron binds the repressor and alters its three-dimensional structure, causing it to detach from the mRNA.
- Hemoglobin consists of four polypeptide chains and a nonprotein pigment, heme.
- Excess heme in the cell increases the rate of translation of globin mRNA by removing a block to initiation of translation at the ribosome.

B. The proteasome controls the longevity of proteins after translation

- Proteins involved in cell division are hydrolyzed at the correct time to control events.

- Proteins identified for breakdown are often linked to a 76-amino acid protein, ubiquitin.
- The protein-ubiquitin complex then binds a complex called a proteasome *(see Figure 14.21)*.
- The protein is cleaved from the ubiquitin and then enters a hollow cylinder nicknamed the molecular chamber of doom.
- Three different proteases digest it there.
- Overall, concentrations of proteins depend on rates of synthesis and rates of digestion.