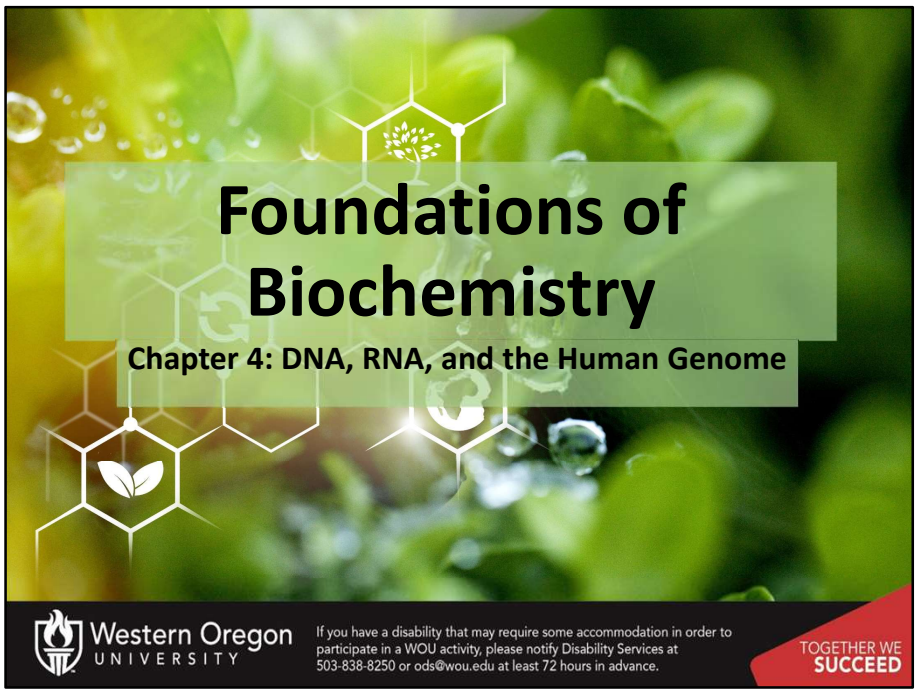
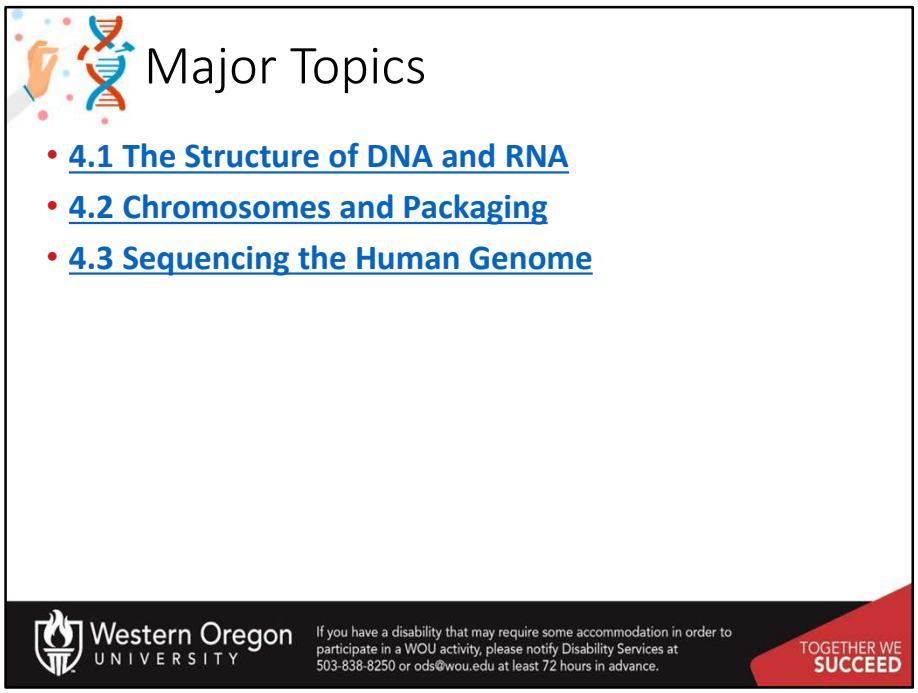
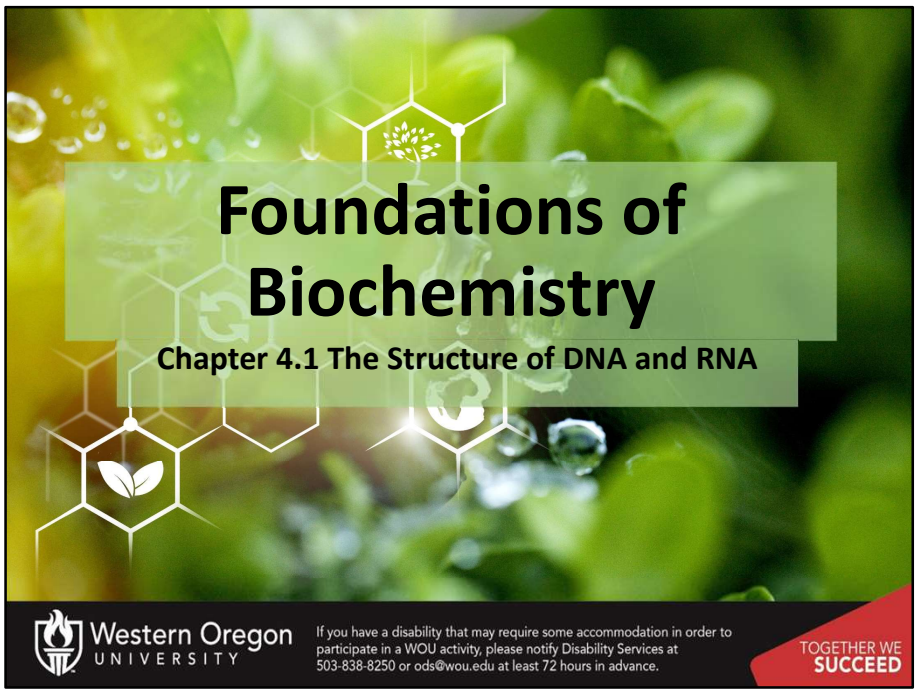


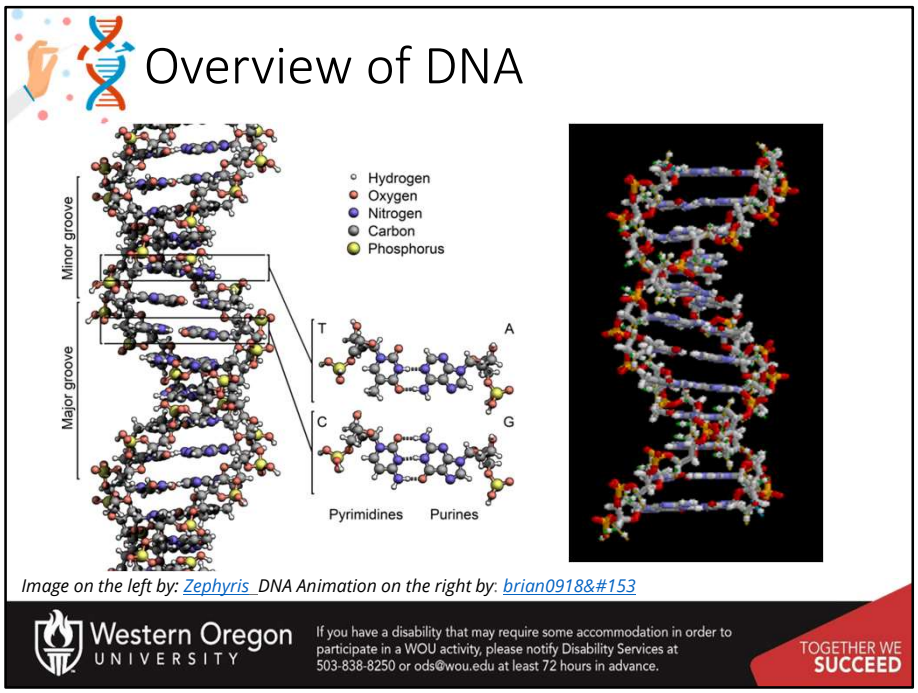
Welcome to our lecture series on DNA, RNA, and the Human Genome.



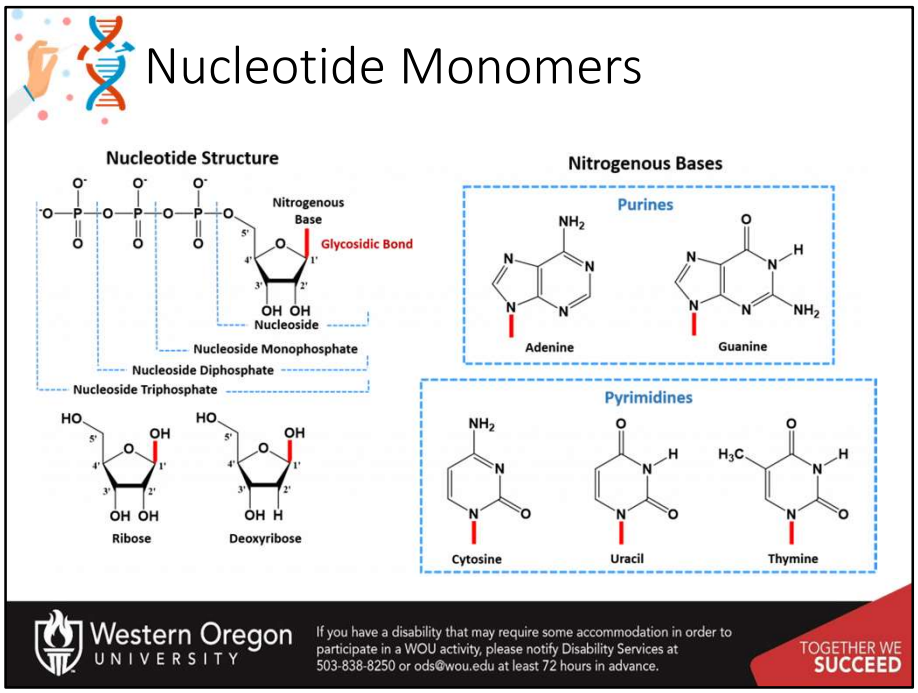
In this chapter, you will become familiar with the structures of the nucleic acids DNA and RNA, discover have chromosomes are packaged within he cell, and learn about the sequencing of the human genome.



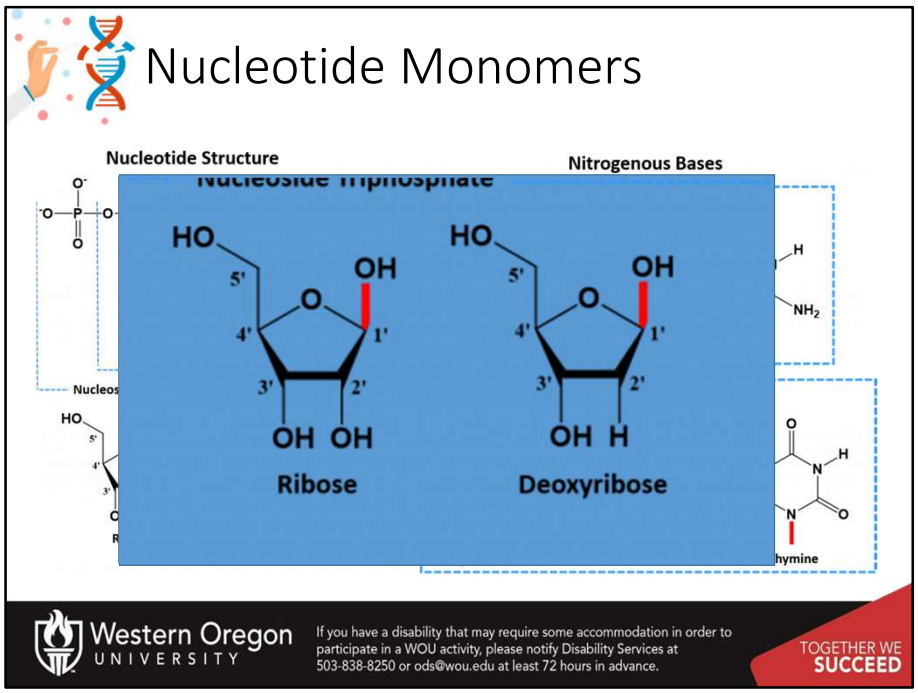
In this first section, we will cover the structure of DNA and RNA.



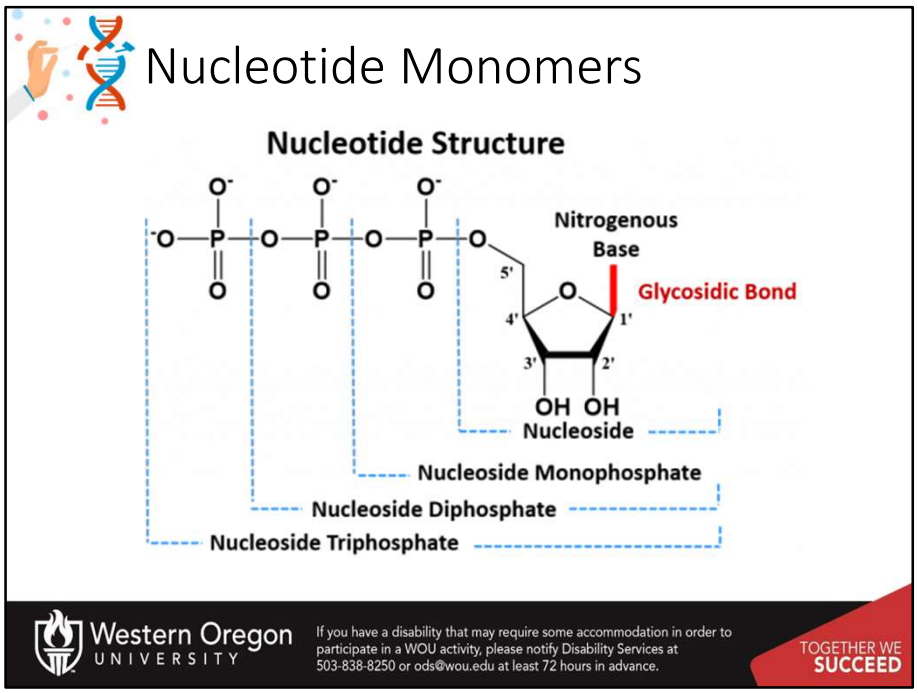
nucleic acids are one of the four major types of macromolecules that are essential for all known forms of life. The nucleic acids consists of two major macromolecules, ***Deoxyribonucleic acid*** *(**DNA**)* and ***ribonucleic acid (RNA)*** that carry the genetic instructions for the development. The DNA macromolecule is composed of two polynucleotide chains that coil around each other to form a double helix.



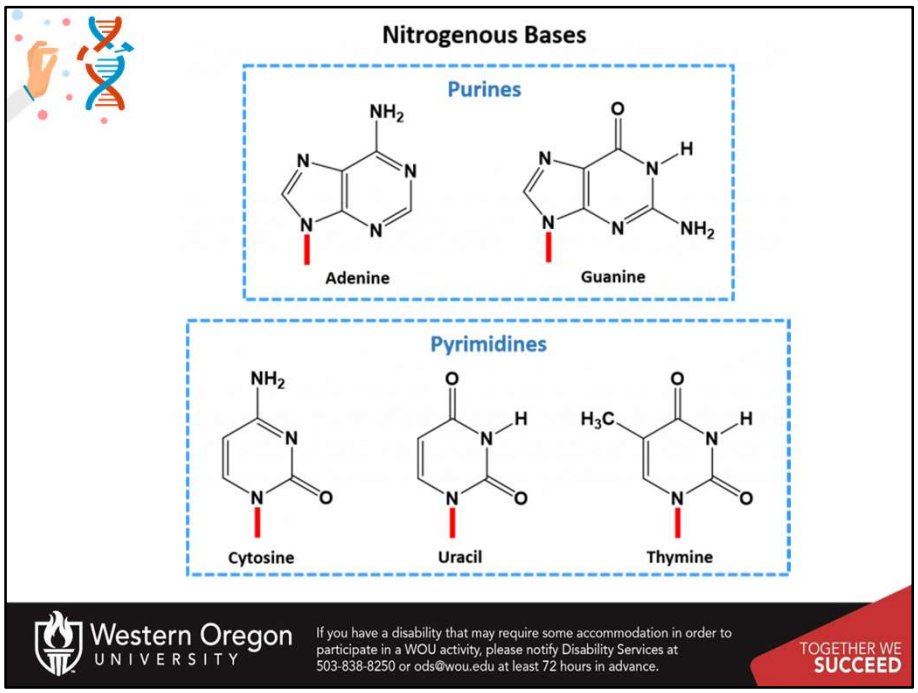
DNA is a polymer that is made up of monomer building blocks called ***nucleotides***. Nucleotides can be broken down into three components: A sugar, a phosphate group, and a nitrogenous base.



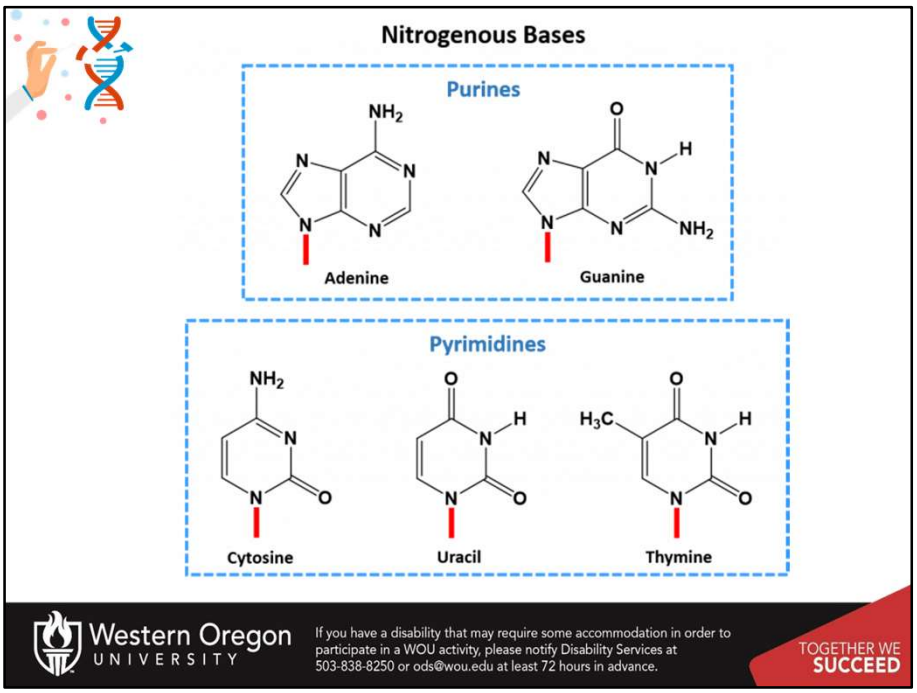
The sugar used in creating RNA molecules is ribose. The numbering of the carbon atoms in sugar molecules starts on the anomeric side of the molecule (or the carbon position where the carbon is bonded to two oxygen molecules. In the case of ribose, this is carbon-1'. This is followed by the 2', 3', 4' and 5' positions. Deoxyribose is used as the core for DNA monomers. The deoxy- position is located at the 2'-carbon. This small structural difference creates many of the physical differences between the RNA and DNA molecules, including things like degradation susceptibility.



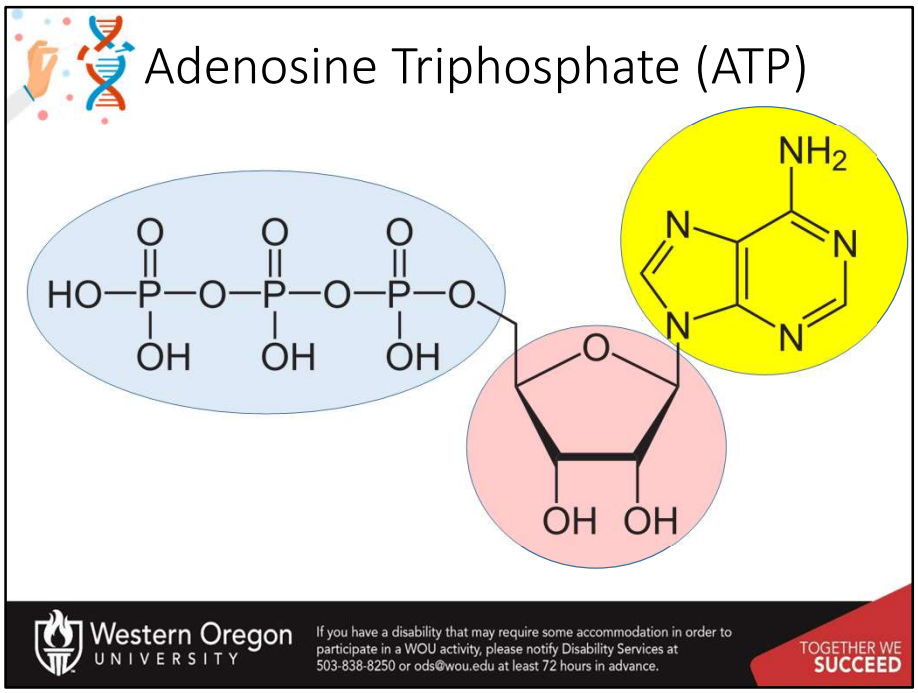
The core structure of a nucleic acid monomer is the ***nucleoside***, which consists of a sugar residue + a nitrogenous base that is attached to the sugar residue at the 1' position. When phosphates are attached to the unit, the building block is called a ***nucleotide***, with a T. The mnemonic that I use to help me remember this, is that a nucleo***SIDE*** only has a ***SIDE*** of sugar, whereas the nucleo***T***ide is the ***T***otal monomer. Note that the phosphate groups are attached at the 5'-position (hence the term 5'-phosphate, that you might already be familiar with). The 3'-OH group is the position where two nucleotides will be linked together when building a molecule of DNA or RNA. Nucleotides can exist as monophosphate, diphosphates, and triphosphates. The triphosphate form is the building block that will be used for RNA and DNA synthesis. The trinucleotide here would be used for RNA synthesis, as the sugar residue is ribose, and contains the 2'-OH group.



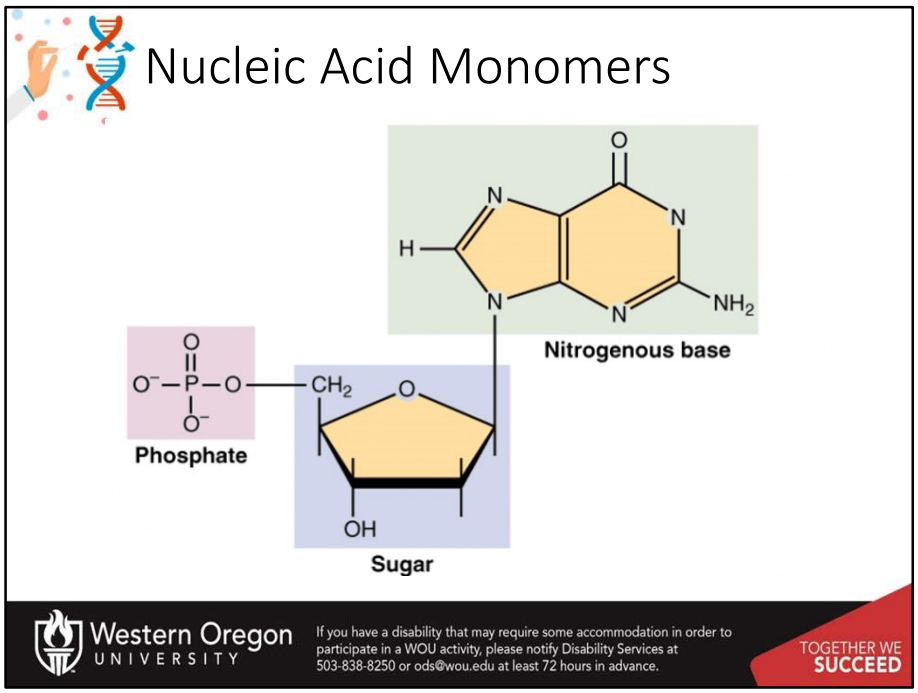
For the DNA molecule, there are four nitrogenous bases that are incorporated into the standard DNA structure. These include the ***Purines:*** Adenine (A) and Guanine (G), and the ***Pyrimidines:*** Cytosine (C) and Thymine (T). RNA uses the same nitrogenous bases as DNA, except for Thymine. Thymine is replaced with Uracil (U) in the RNA structure. The red lines indicate where each nucleotide is attached to the sugar molecule. Note that the purines (AdeNINE & GuaNINE) each have NINE in their name. I think this is a useful way to remember that they are the purines (or the bases that have two ring structures) Note that the two ring structures have NINE adjacent atoms. This is in comparison to the Pyrimidines, which contain single 6-membered ring structures. I like to link c***Y***tosine and th***Y***mine to the p***Y***rimidines by the ***'Y'*** in their names. Hopefully, this will help you identify them a little more easily! You will need to know how to recognize each of the bases by site.



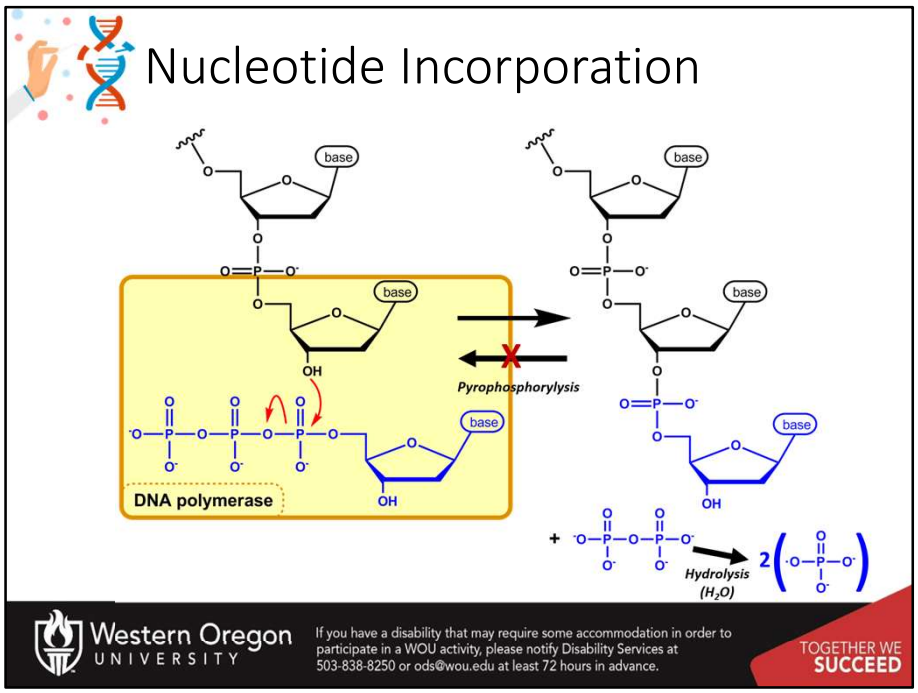
To further distinguish Adenine from Guanine, I use the alphabet. A comes ***before G***, and N comes ***before O***. Notice that ***Adenine contains only Nitrogens*** in the ring structure, whereas ***Guanine contains the Oxygen***. It also contains more Nitrogen, as well. But it is the presence of Oxygen that I look for, to identify Guanine. For Cytosine and Thymine, I use a similar strategy. C comes before T in the alphabet, and notably, Cytosine contains a Nitrogen atom that extends off of the ring structure. Thymine only contains Oxygen. Note now that Thymine also contains a methyl group (CH3). This methyl group can help you distinguish between Thymine and Uracil. That is the only difference between them. You can distinguish Cytosine from Uracil, by the Nitrogen contained on the Cytosine base. Practice identifying these residues! You will need to know them for the exam.



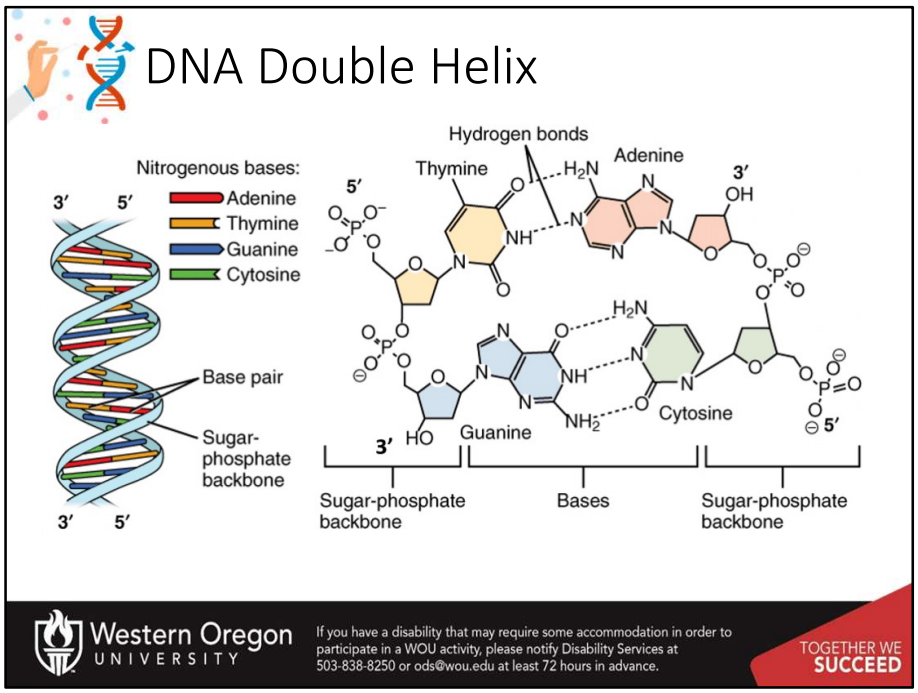
This is just a reminder that you have already seen ATP or Adenosine Triphosphate from chapter 1. Notice that it is the same building block used in RNA synthesis. Many of the nucleotides have dual functions within the cell. They are nucleic acid building blocks, but they also serve as cofactors in enzyme reactions, providing the needed energy to drive the reaction forward.



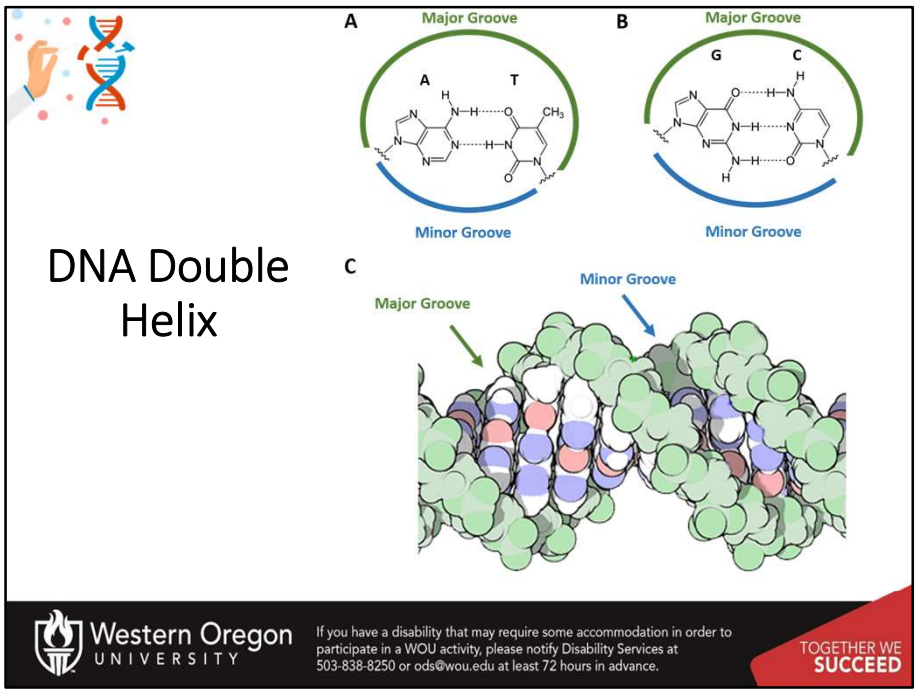
When the nucleotide triphosphates are incorporated into the nucleic acid structures, two of the phosphates are cleaved off, leaving the nucleotide monophosphate as the nucleic acid building block.



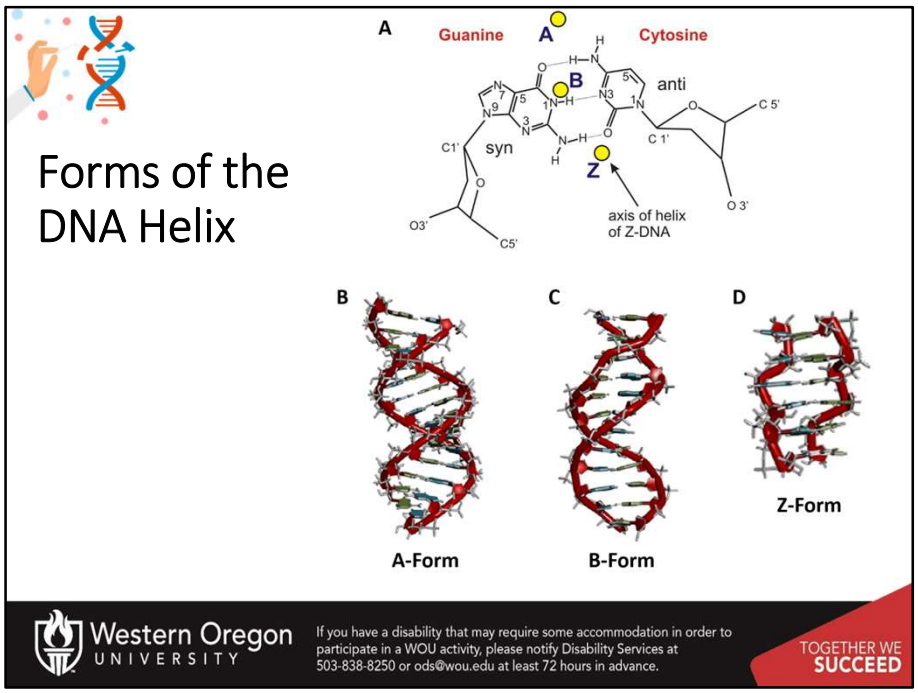
During nucleic acid biosynthesis, nucleotides are incorporated in a directional manner. They can only be added to the 3'-OH side of the nucleic acid strand. This is true for both DNA and RNA. The 3'-OH of the nascent strand, mediates nucleophilic attack on the phosphorous of the first phosphate position. This causes the other two phosphates to serve as a leaving group. The further hydrolysis of the diphosphate to the monophosphate, releases energy that drives the synthesis reaction forward. This energy release, provides a barrier that keeps the reaction from moving in the reverse direction. Thus, the growing nucleic acid is always built in the 5' to 3' direction. Remember 5'-phosphate and 3'-hydroxyl.



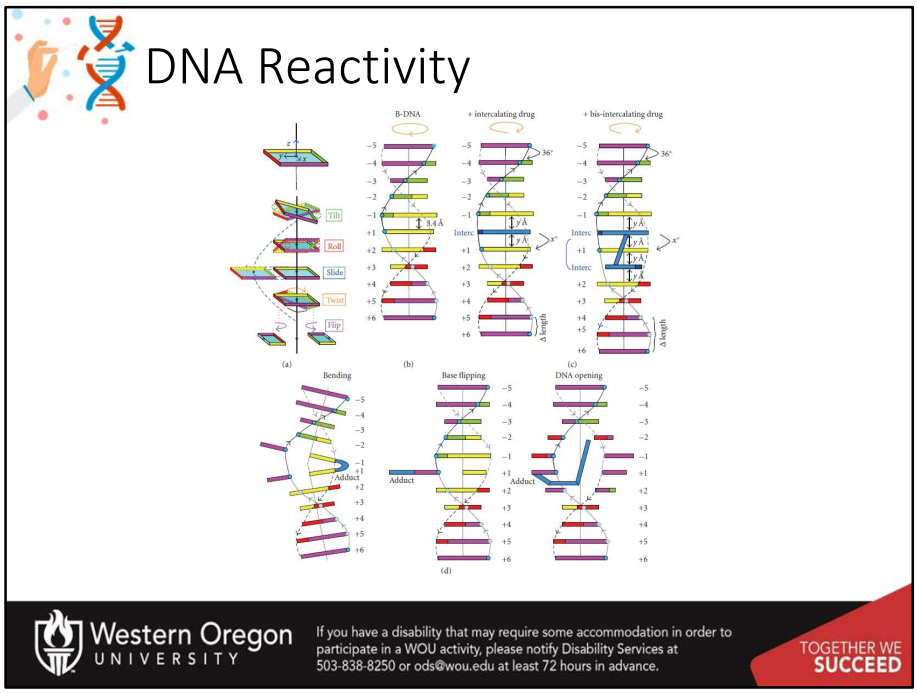
The overall DNA structure forms a double helix, where two separate strands of DNA align with one another in a regular twisting helical pattern. Notable features about the double helix, is that the two strands lie in antiparallel fashion or head to tail, with the 5'-side of one strand adjacent to the 3'-side of the other strand. The DNA bases also face inward towards the center of the helix, where they form hydrogen bonds that hold the helix together. In DNA, the adenine base always pairs with the thymine base forming 2 hydrogen bonds, whereas guanine always pairs with cytosine forming 3 hydrogen bonds.



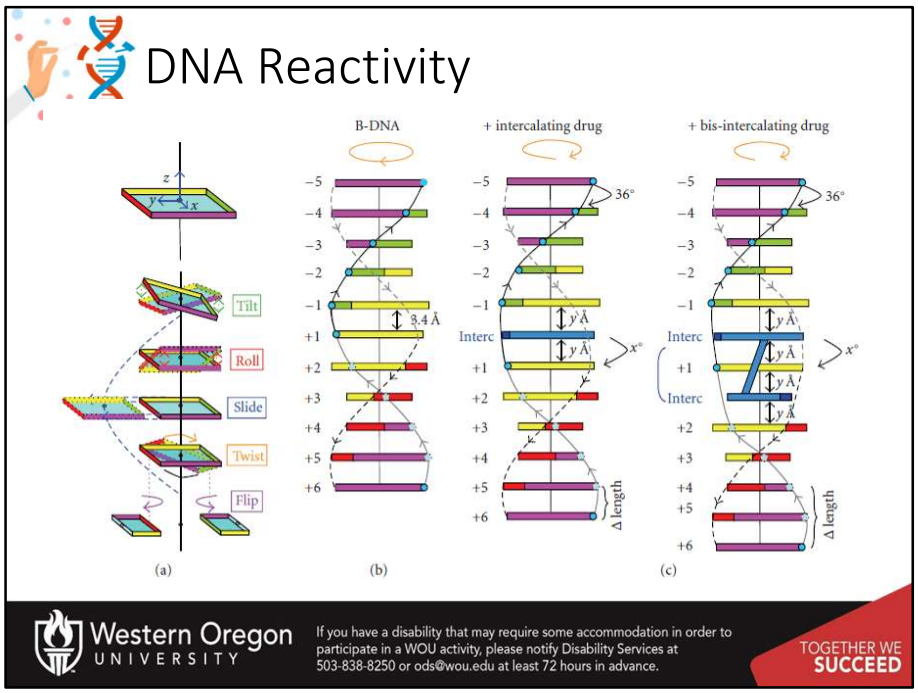
The structural elements of the DNA double helix form repeating three dimensional features in the helix. Most notably the formation of major and minor grooves inherent to the nature of the A-T and G-C base pairing. Here you can see the major groove is much larger than the minor groove, and that a major groove is always flanked by a minor groove on the other side of the helix, such that if we could see the backside of this molecule, there would be a major groove here, and a minor groove behind here. The major and minor grooves and minor variations found within them due to specific nucleotide sequences provide unique protein binding sites for transcription factors to recognize and bind in a very sequence-specific manner.



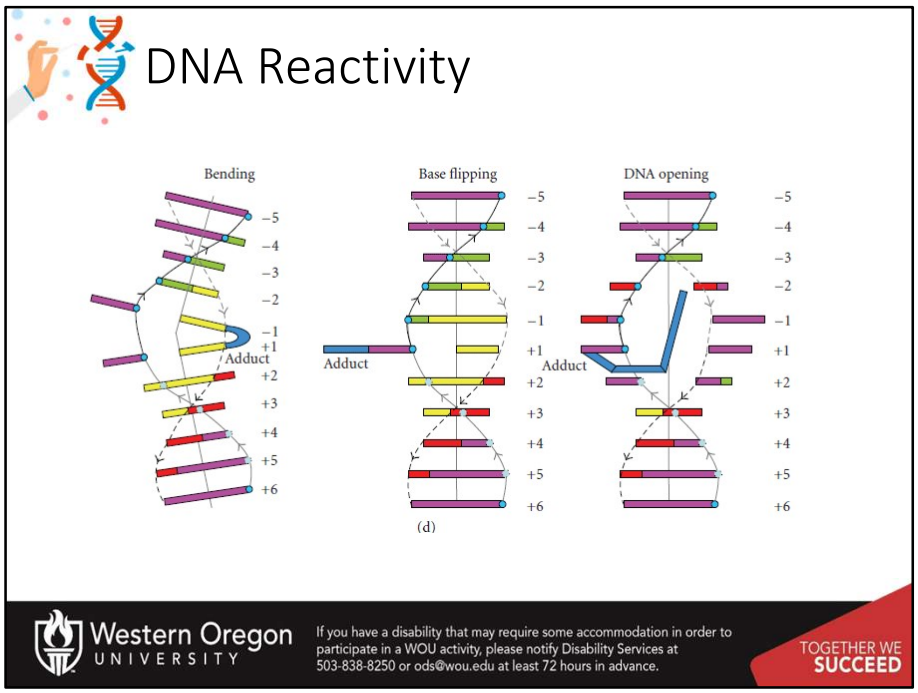
Depending on the hydration level of the DNA, and on regional sequence patterns, different forms of the double helix can arise. These include the ***A-form (A-DNA)***, the ***B-form (B-DNA),*** and the ***Z-form (Z-DNA).*** Both the A- and B-forms of the double helix are right-handed spirals, with the B-form being the predominant form found *in vivo*. The A-form helix arises when conditions of dehydration below 75% of normal occur and have mainly been observed *in vitro* during X-ray crystallography experiments when the DNA helix has become dessicated. the A-form of the double helix can occur *in vivo* when RNA adopts a double stranded conformation, or when RNA-DNA complexes form. The Z-form of DNA is a left-handed helix that is found rarely within DNA sequences, but highlights the flexibility of the DNA structure. Within the B-DNA structure, about 10 base pairs are required for 1 turn of the helix.



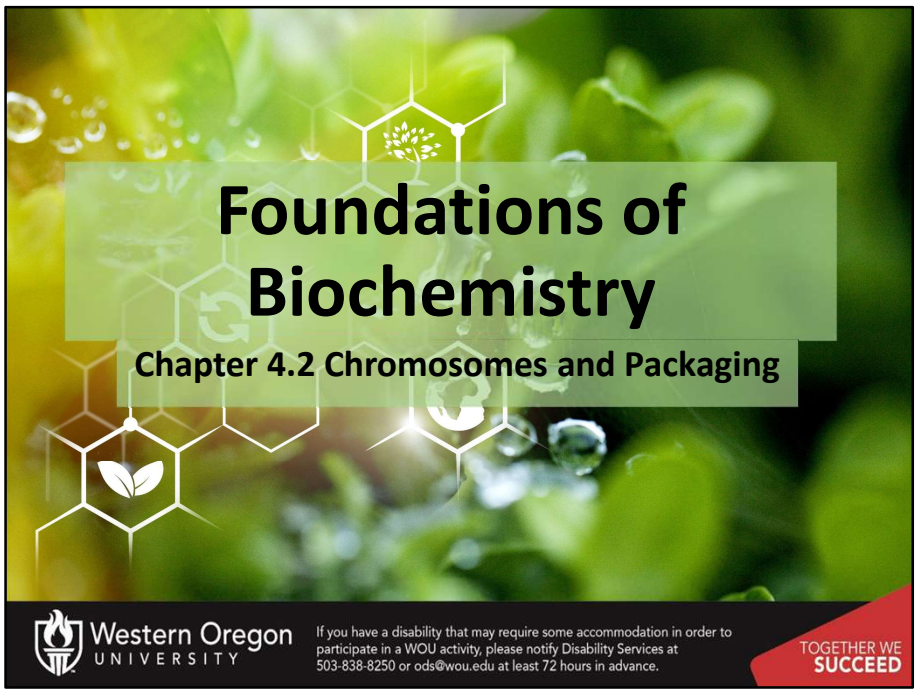
Because the DNA bases contain a lot of nitrogen and oxygen, they are fairly reactive and can be chemically and physically modified by interaction with external chemicals and proteins. This can lead to the regulation of DNA processes such as transcription and translation, but it can also lead to unwanted DNA modifications through base modification. These types of modifications can alter DNA structure.



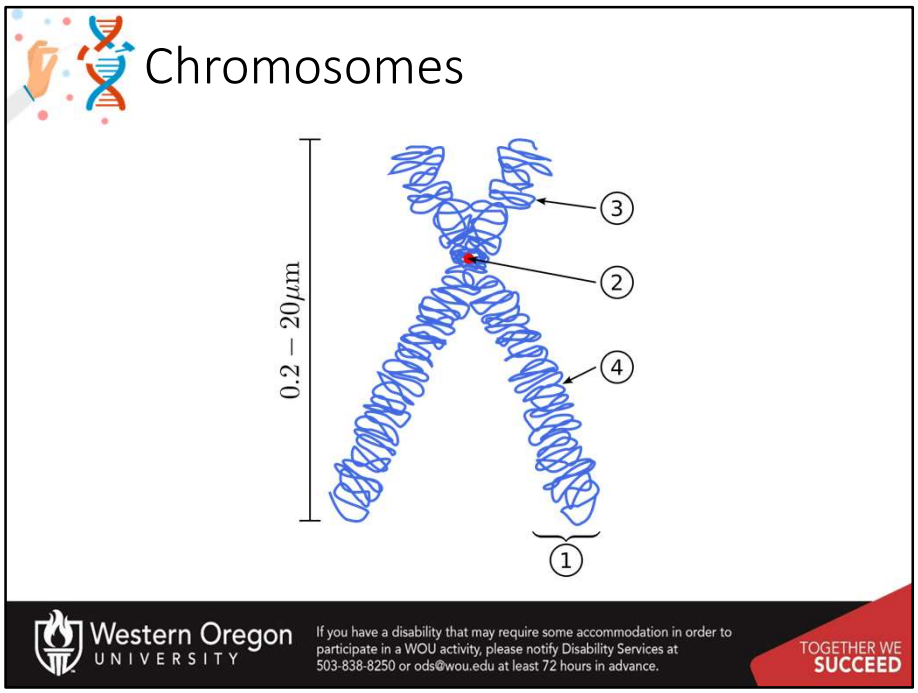
The diagram on the left, shows some of the structural modifications that can occur when DNA interacts with foreign, chemically reactive molecules. These can include shifts such as base tilting, rolling, sliding, twisting and even flipping. For example, intercalating agents are molecules that can slide inbetween DNA bases and are shown here in blue. When compared with the normal helical structure, in (b), intercalating agents can alter the length of separation between the base pairs and distort or elongate the major and minor groove structures.



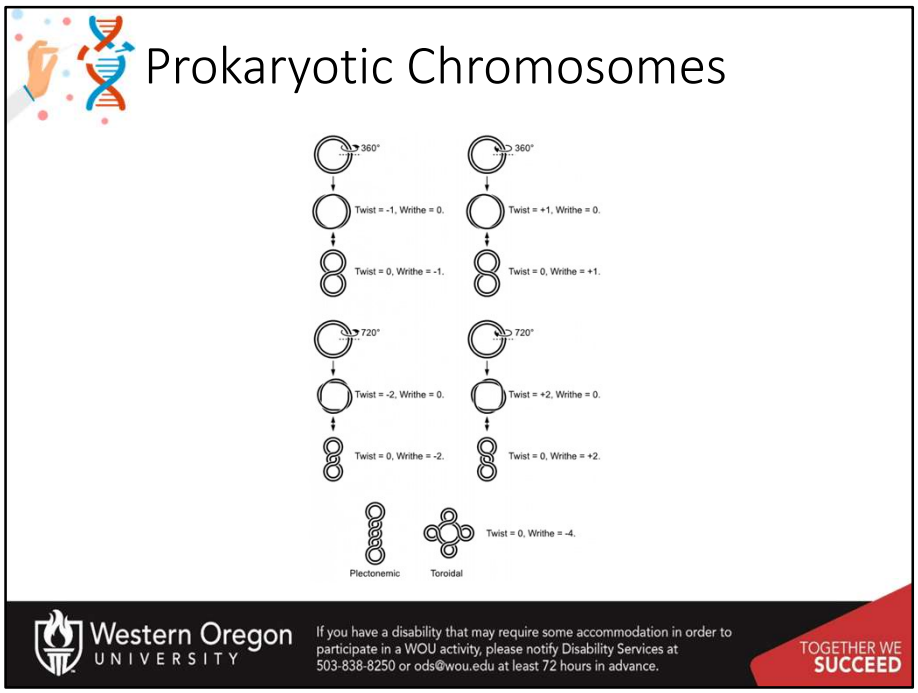
Other agents may react with and covalently bind with the DNA bases. When this happens the resulting lesion is called a DNA Adduct (shown in blue here). DNA adducts can cause the helix to bend, or it can cause base flipping, or even regional opening of the double helix. Lesions within the DNA can alter the ability of transcription to occur and potentially lead to the formation of sequence mutations if the DNA is replicated while lesions are present. In the next lecture, we will talk about how the double helix is packaged and protected within the nucleus of the cell to minimize the formation of these types of damaging lesions.



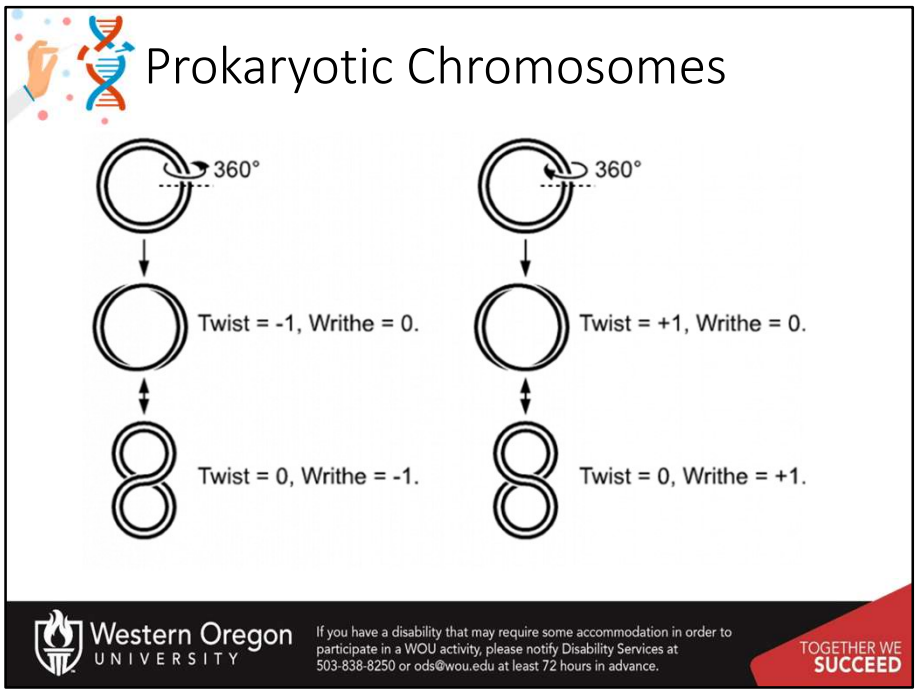
In this next section, we will cover chromosome structure and packaging.



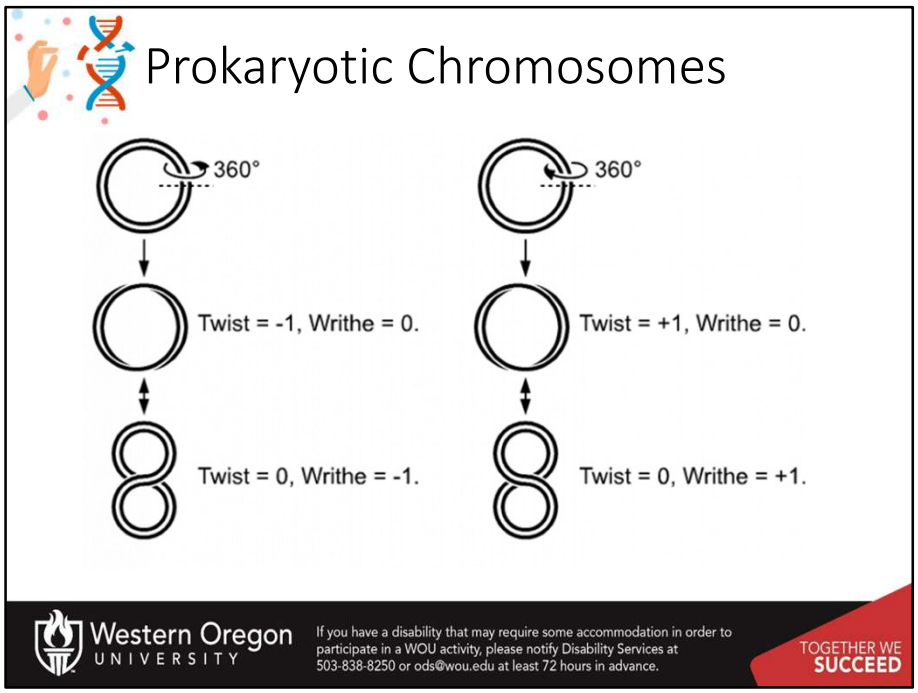
DNA is organized into long linear structures called ***chromosomes.*** ***Chromosomes*** consist of both DNA and the proteins used to organize and fold it. The diagram above shows a replicated and condensed chromosome that is getting ready to divide and be separated into to daughter cells. Each one of the two identical replicated chromosomes is called a chromatid. Chromatids are held together at the centromere of the chromosomes. From the centromere, the long arm of the chromosome is called the q-arm and the short arm is called the p-arm.



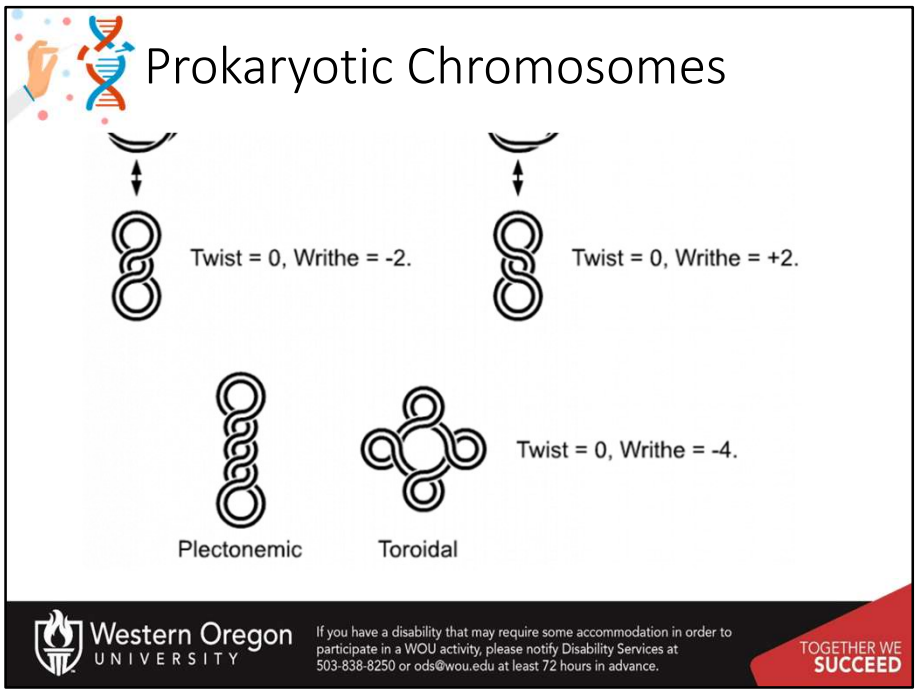
Prokaryotic cells do not contain any organelles, thus the chromosomal DNA exists as a circular structure in a region of the cell known as the nucleoid region.



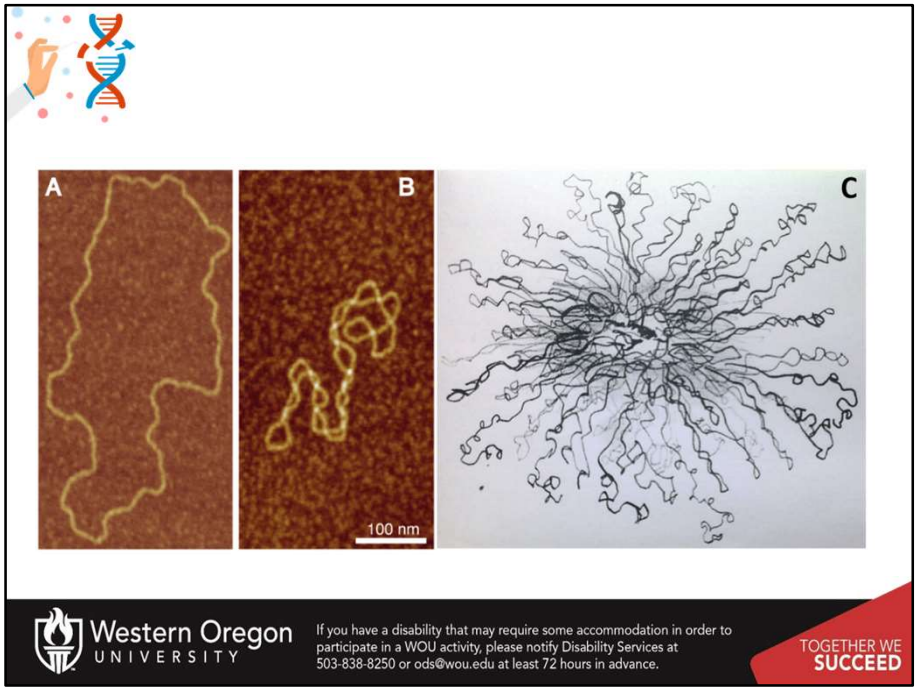
Prokaryotic DNA often exists in a supercoiled state, rather than as a relaxed circular DNA. ***DNA supercoiling*** refers to the over- or under-winding of a DNA strand, and is an expression of the strain on that strand (Figure 4.9). Supercoiling is important in a number of biological processes, such as compacting DNA, and by regulating access to the genetic code. ***DNA supercoiling*** refers to the over- or under-winding of a DNA strand, and is an expression of the strain on that strand (Figure 4.9). Supercoiling is important in a number of biological processes, such as compacting DNA, and by regulating access to the genetic code.



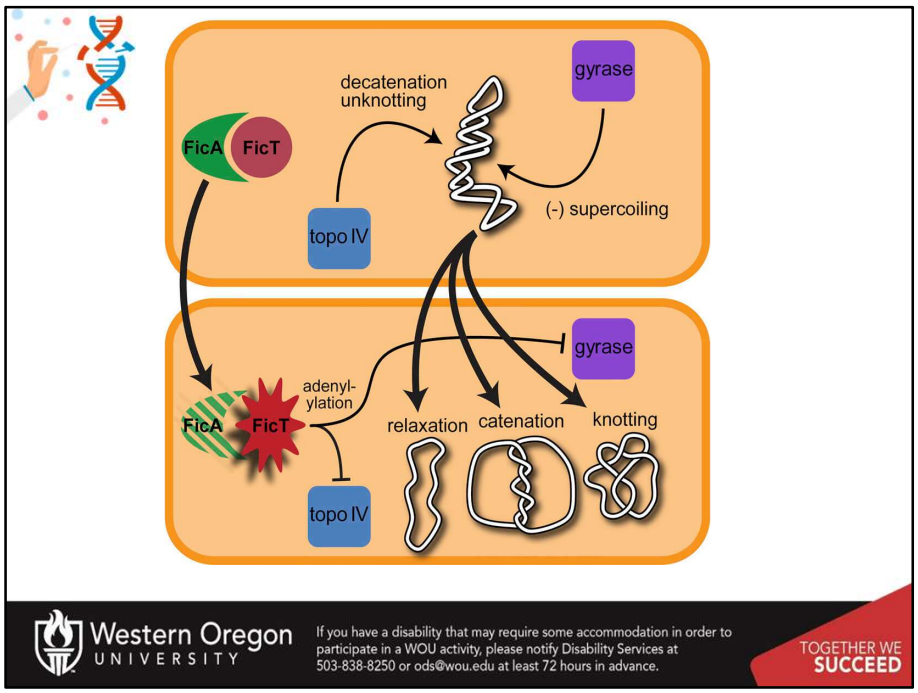
Supercoils can exists as twists with different amounts of writhing that occur in the righthanded or lefthanded orientation. Writhing is said to occur when distinct lobe structures form in the DNA



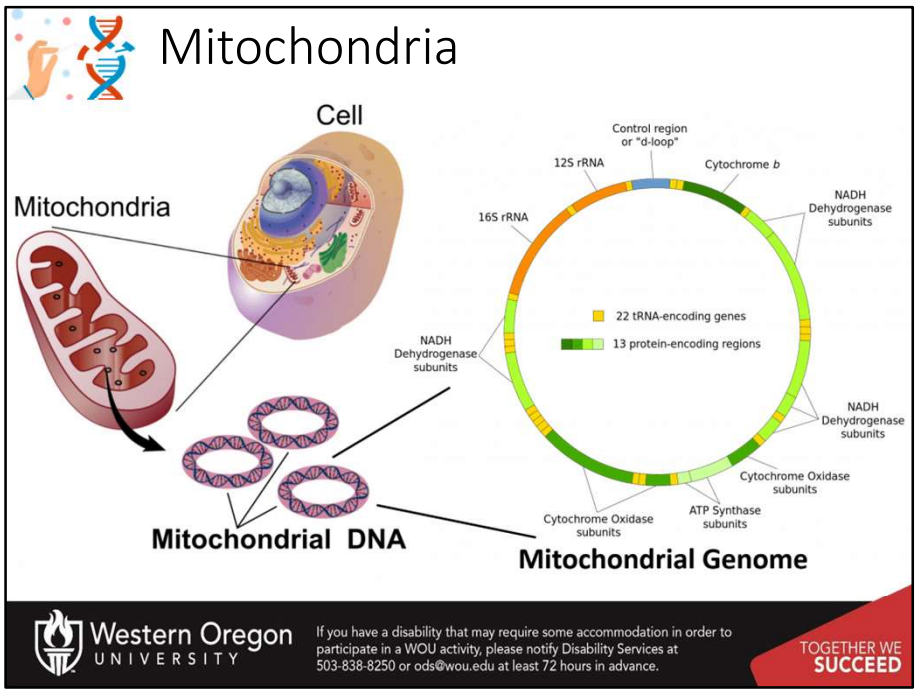
Supercoils can also be tightly would into plectonemic and toroidal conformations. Supercoiling is mediated by enzymes known as topoisomerases. Later this term, when we are focused on the processes of DNA replication and transcription, we will discuss topoisomerase enzymes in more detail.



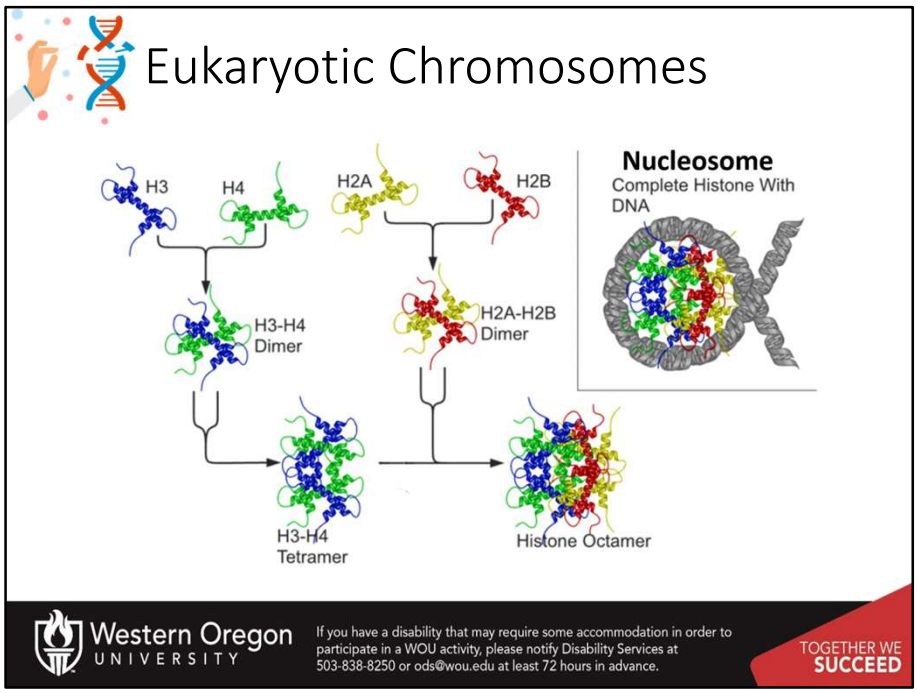
Atomic force microscopy can be used to visualize some of these different states, including the relaxed, plectonemic and toroidal states.



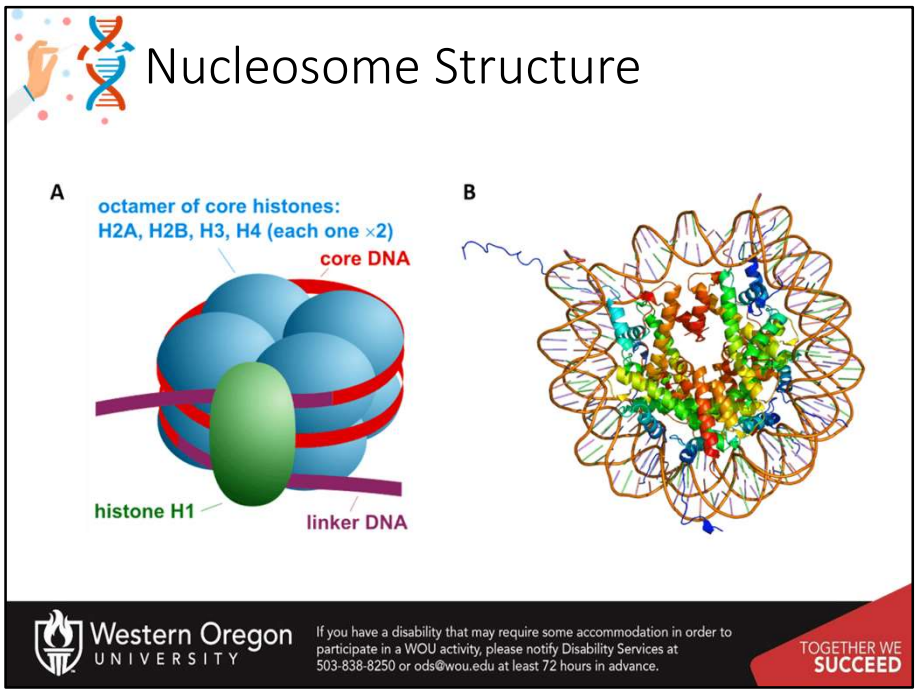
Sometimes during the process of replication, the two sister chromatids can get entangled with one another forming knots or becoming catenated. The topoisomerase enzymes (also called gyrases) are critical for unknotting and decatenating the DNA.



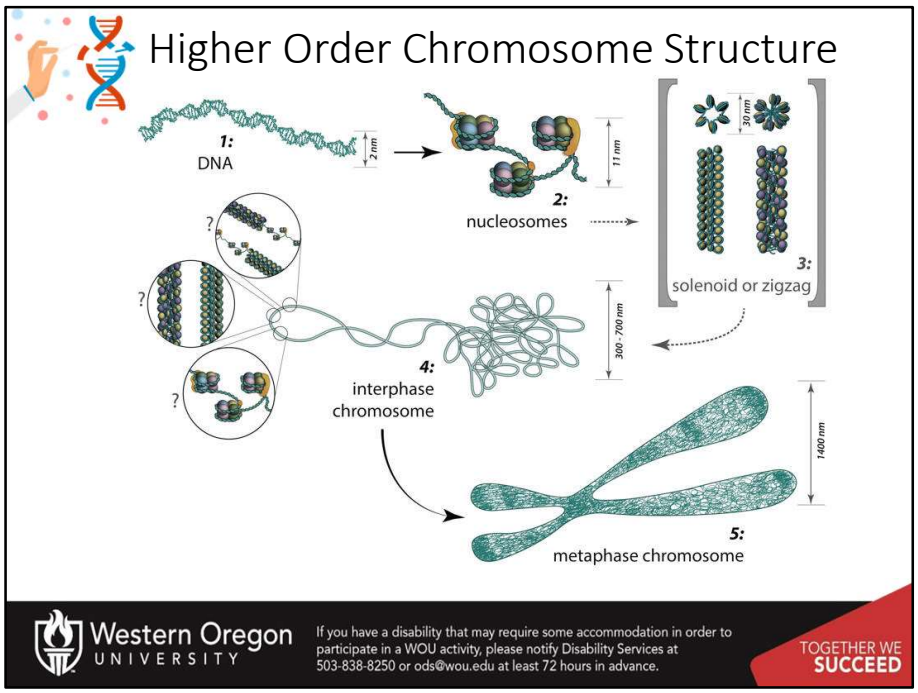
Circular chromosomal DNA is also found within the mitochondrial organelles of eukaryotic cells, supporting the hypothesis that they originated as prokaryotic organisms that became symbionts with an early eurkaryotic progenitor cell. The mitochondrial genome contains 13 protein encoding regions and 22 tRNA encoding regions and is involved in the biosynthesis of critical metabolic enzymes housed within the mitochondria.



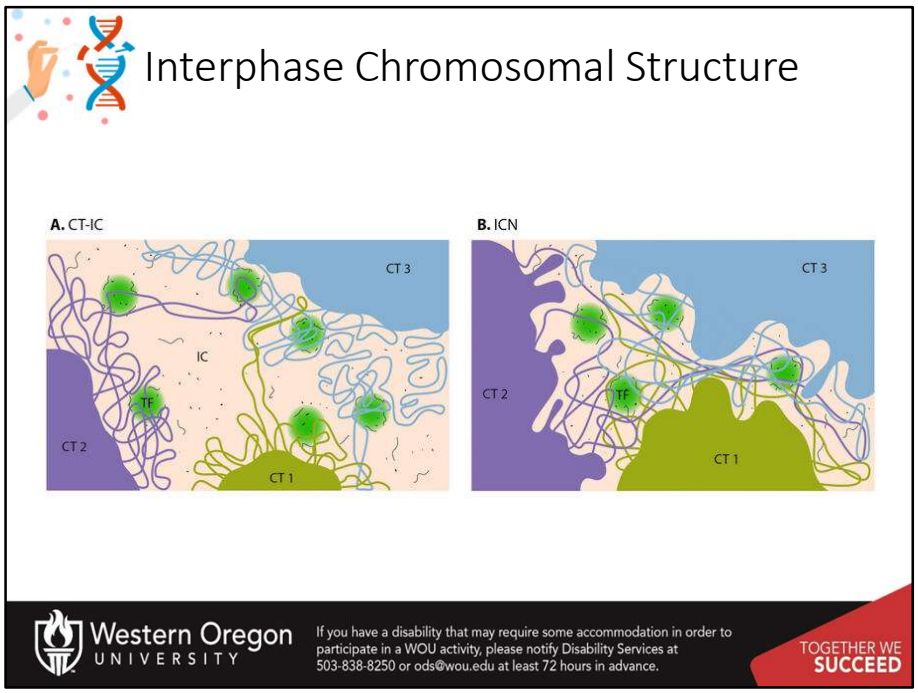
Eukaryotic chromosomes are housed within the nucleus of the cell and differ from prokaryotic DNA, in that the chromosomes exist as long, linear structures rather than being circular in nature. The major proteins that are associated with the DNA in the chromosome structure are known as histones. Histones are highly alkaline proteins (note that DNA has a lot of negative charge load due to the phosphate groups, thus, the alkaline nature or positive nature of the histones, keeps them attracted to the DNA.). Histones 3 &4 form a dimer, and histones 2A and 2B form a dimer, these counterparts come together to form tetramers and ultimately an octamer structure that provides a protein core that the DNA can wrap around. This DNA/histone core is called a ***nucleosome***



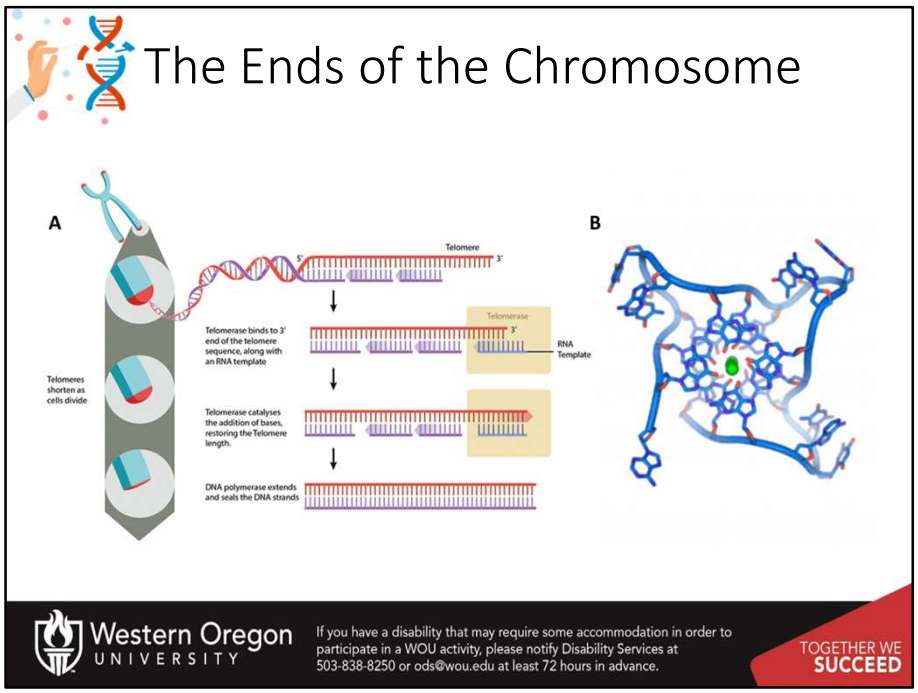
Here is a side view diagram and a top down diagram of the nucleosome core. Note that histones 2A, 2B, 3 and 4 form the octamer core, while Histone H1 is a linker that serves to seal the nucleosome structure, and also is a coordination point for higher order structuring of the DNA. The DNA is wrapped twice around the histone core and is shown in red on the right diagram



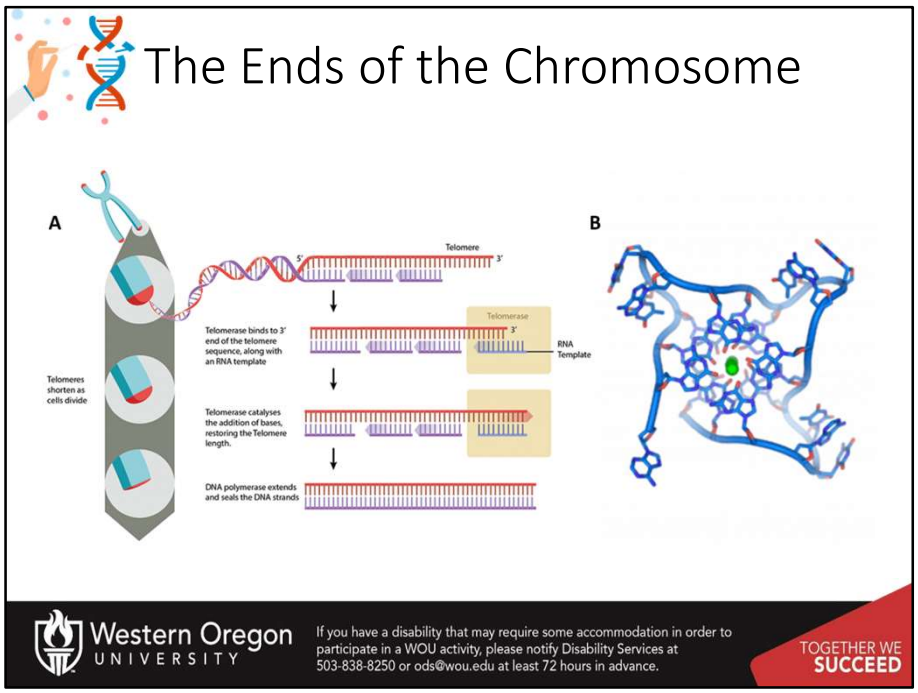
Once the DNA is wrapped onto the histone core to form nucleosomes, the nucleosomes can assemble into higher order structures such as solenoids or a zigzag-type structure. The solenoids and/or zigzags can be stacked into an organized structure. When the cell is not replicating the DNA or getting ready to divide, it is in a phase called interphase, and the DNA is relaxed and diffuse. In this state, regions of the DNA can be unwound from the histones to allow transcription, and hence gene expression to take place. During DNA replication and cell division, the DNA condenses into a very tightly wound structure forming the characteristic sister chromatid X structure. The DNA is typically NOT being transcribed in this state.



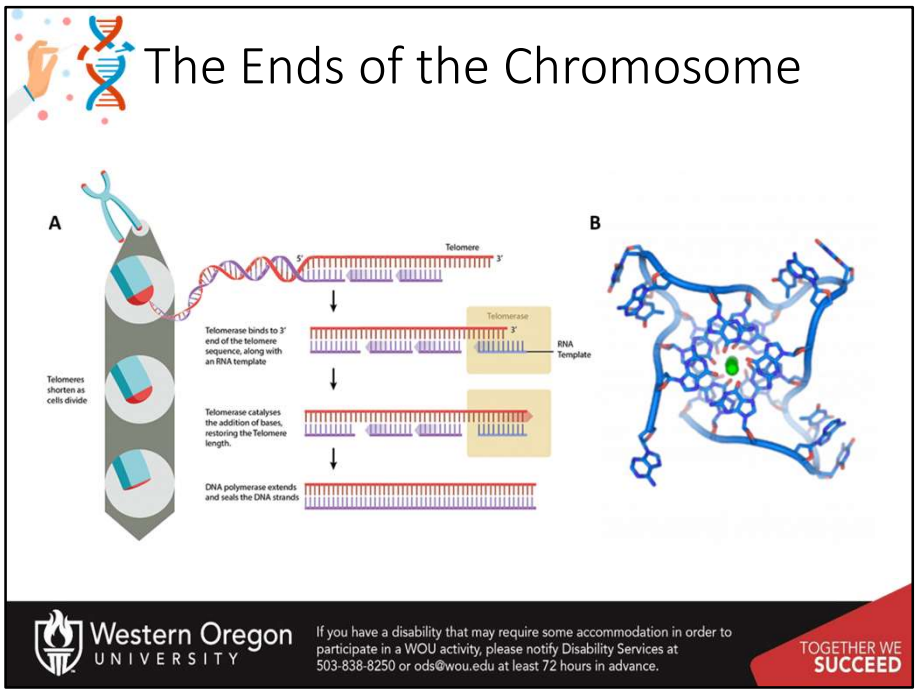
Chromosomes in interphase are much more dynamic than when they are condensed and highly structured for cell division. A number of different models have been proposed to describe structure of interphase DNA, especially differences in chromosomal DNA that is being actively transcribed vs. inactive DNA. You will not be required to know these different models



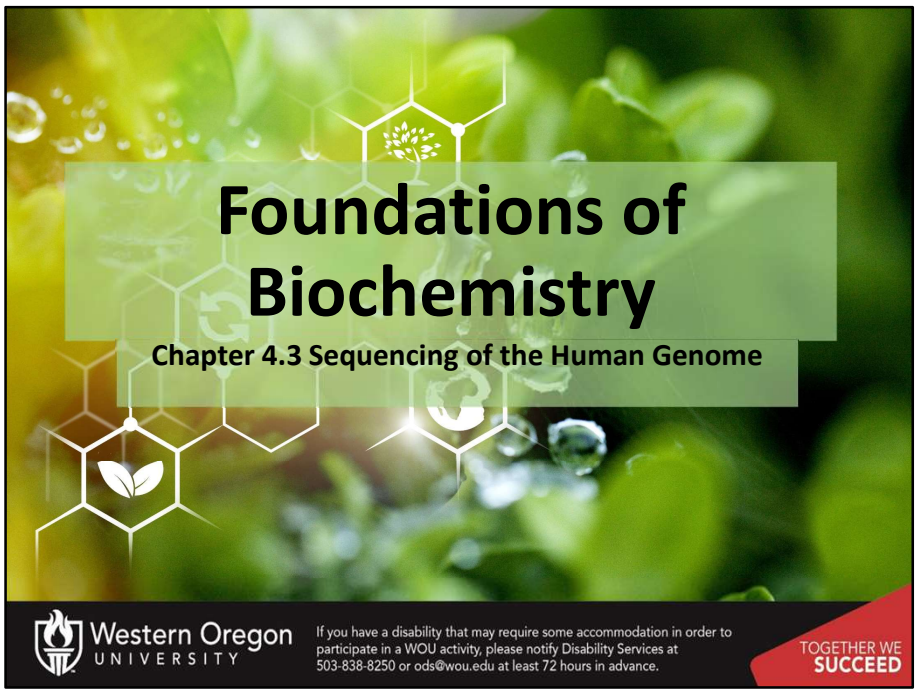
The ends of the chromosome pose a special problem for linear chromosomal DNA. They provide a region that is more susceptible to degradation, and cannot be fully replicated during DNA replication. To aid in the maintaining the stability of the chromosome and creating a mechanism for replicating the ends of the chromosome, special structures, called ***Telomeres*** exist at the ends of each linear chromosome. Telomeres are long regions of DNA (thousands of bases) that repeat the sequence TTAGGG. The telomerase enzyme, shown above in yellow, binds with a small RNA template molecule that is complementary with the TTAGGG sequence. Thus, it can base pair with this region and serve as a template, enabling the elongation of the chromosome ends.



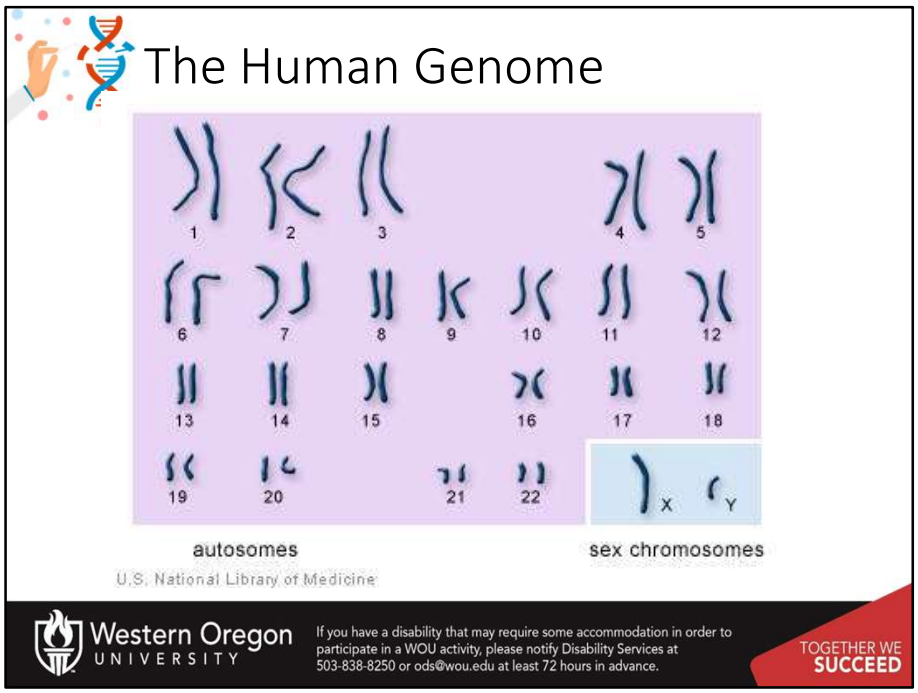
However, the telomerase enzyme required for this elongation process is only present within embryos and is turned off shortly after birth. Thus, over a lifetime, each round of DNA replication will result in a shortening of the chromosome. At some point, the telomeres become critically short. This attrition leads to cell senescence, where the cell is unable to divide, or apoptotic cell death. Telomeres are the basis for the ***Hayflick limit,*** the number of times a cell is able to divide before reaching senescence. While telomerase reactivation may seem to be the solution to longevity, it is also problematic, as overexpression of telomerase occurs often in tumor cells that have become cancerous. It is not clear if there can be safe expression levels of telomerase to maintain chromosomal length without causing harmful diseases.



Telomeres also stabilize the DNA by forming highly complex 3-D shapes at the ends of the chromosomes. Within this structure, four guanine bases form a flat plate, which then stack on top of one another forming a stable G-quadraplex structure. So this figure here is like we are looking down the center of the double helix showing the quadraplex structure here. Other structures such as T-loops, and D-loops can also sometimes form.



In this last section, we will cover some of the major findings revealed during the sequencing of the human genome, and preview the processes of transcription and translation that will be covered in detail later.



In humans, each cell normally contains 23 pairs of chromosomes, for a total of 46. Twenty-two of these pairs, called ***autosomes***, and look the same in both males and females. The 23rd pair, the sex chromosomes, differ between males and females. Females have two copies of the X chromosome, while males have one X and one Y chromosome. Each species has a unique chromosomal complement. For example, chickens have 39 pairs of chromosomes (78 total) with 38 autosomal pairs and one pair of sex chromosomes (Z and W). In this species, ZW chickens are female and ZZ chickens are male. The total length of the human genome is over 3 billion base pairs. The total length of the human genome is over 3 billion base pairs. The genome also includes the mitochondrial DNA

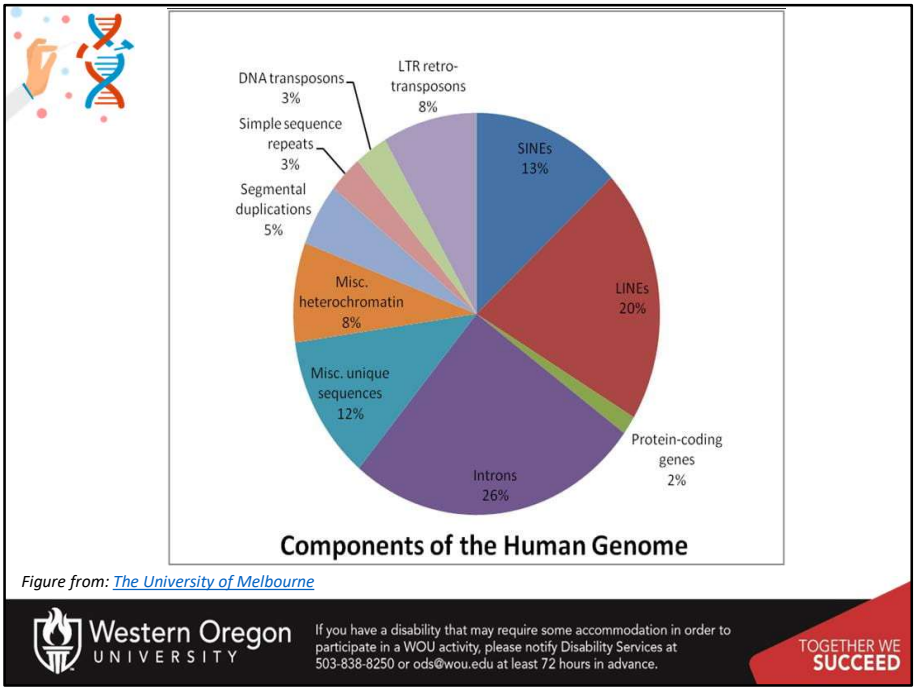


Figure from: The University of Melbourne

Amazingly, of the 3 billion bases of code, only about 2% of the genome encodes for proteins! So what is all of the 'extra' genome doing? About 26% of the sequence contain intron sequences between the gene coding exons. As we will see, the intron sequences are transcribed with the mRNA during protein expression, however, they are spliced out of the exon reading frames prior to protein translation. Somewhere between 8 – 20% of the genome is used for regulation and fine-tuning protein expression. A small part of the genome, such as the telomeres, play a structural role in the molecule and other types of non-coding RNA are also transcribed. About 50% of the genome contains repetitive sequences. These include Short and Long Interspered Nuclear Elements (SINEs & LINEs, respectively), which are known as transposable elements.

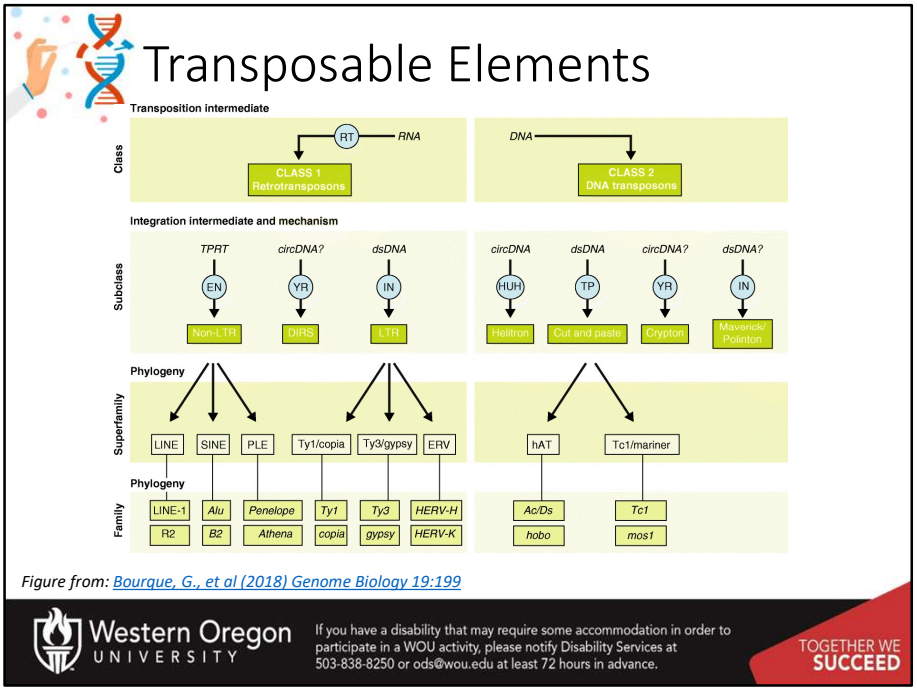


Figure from: [Bourque, G., et al (2018) Genome Biology 19:199]

Transposable elements (TEs) are DNA sequences that have the ability to change their position within a genome. As a result of their deep evolutionary origins and continuous diversification, TEs come in a bewildering variety of forms and shapes. TEs can be divided into two major classes based on their mechanism of transposition, and each class can be subdivided into subclasses based on the mechanism of chromosomal integration. Class 1 elements, also known as retrotransposons, mobilize through a 'copy-and-paste' mechanism whereby a RNA intermediate is reverse-transcribed into a cDNA copy that is integrated elsewhere in the genome. Class 2 elements, also known as DNA transposons, are mobilized via a DNA intermediate, either directly through a 'cut-and-paste' mechanism or, in the case of *Helitrons*, a 'peel-and-paste' replicative mechanism involving a circular DNA intermediate.

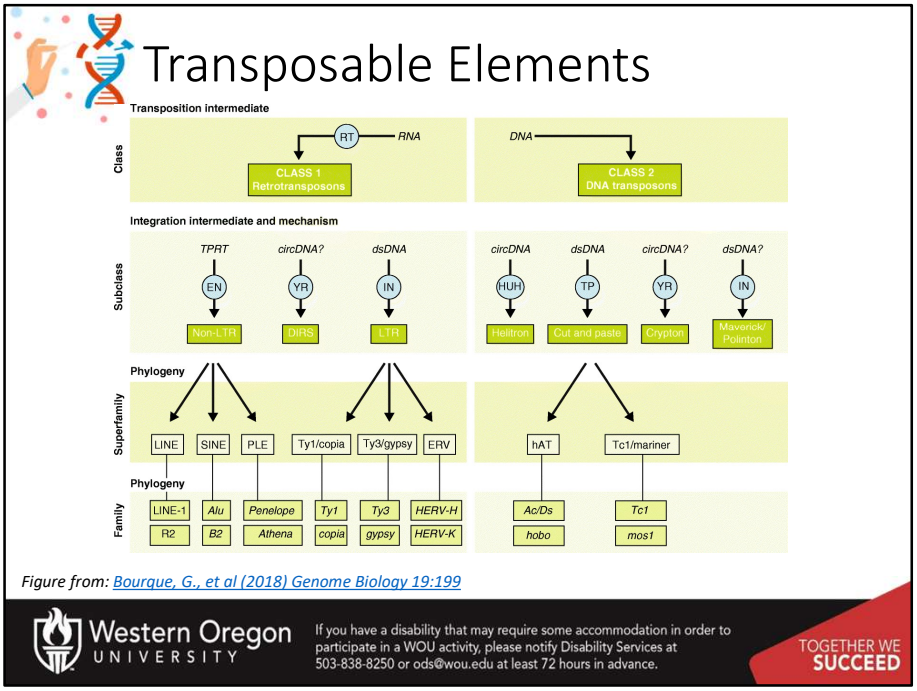
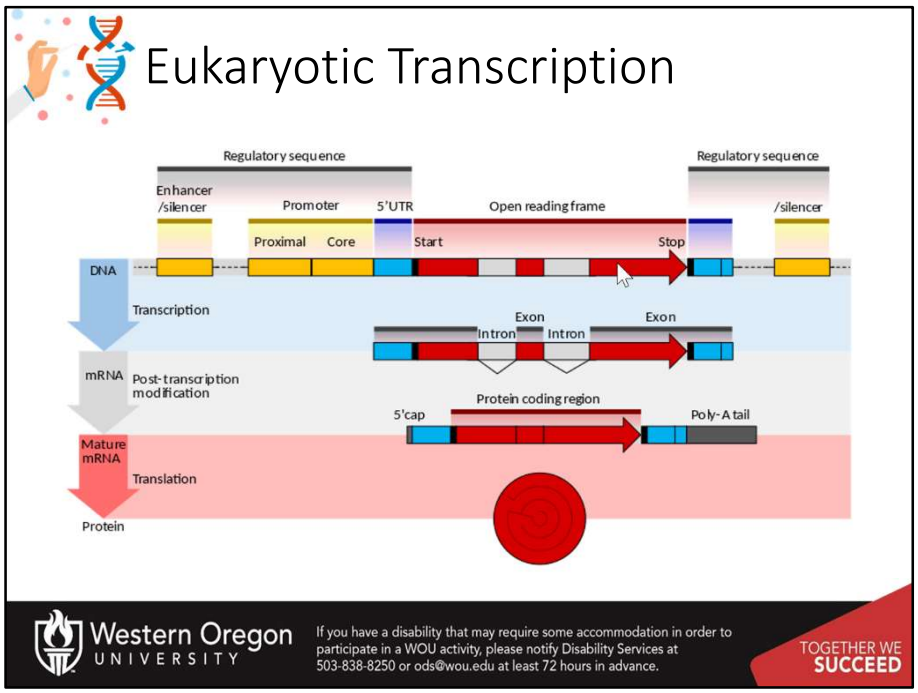


Figure from: [Bourque, G., et al (2018) Genome Biology 19:199]

Historically little attention has been given to transposition in somatic cells and its consequences, because somatic transposition may be viewed as an evolutionary dead end

for the TE, with no long-term consequences for the host species. Yet there is abundant evidence that TEs are active in somatic cells in many organisms. There are thoughts that they may play a role in processes like the diversification of neuronal cells or in early embryonic development. There is also evidence the transposon activity can be quite high in disease states such as cancer. Thus, more needs to be learned about these processes.

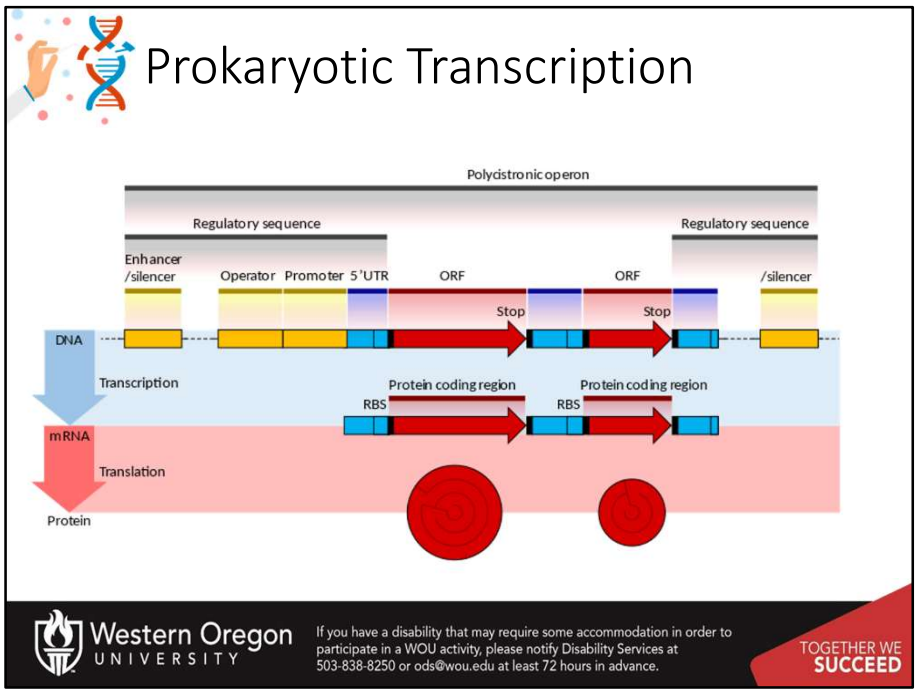


Sequencing of the human genome has enabled us to learn more about the structure of genes, or sequences of DNA or RNA that code for a molecule that has a function.

The process of transcription leads to the production of an RNA molecule from a DNA template. The open reading frame of the gene is usually represented as an arrow, indicating that direction in which the sense strand is read. An RNA Polymerase enzyme will actually bind to the template strand of the DNA rather than to the sense strand, and it will produce an RNA molecule that looks like a copy of the sense strand and is oriented in the same 5' to 3' prime direction as the open reading frame. Regulatory sequences are located at the extremities of the gene. These sequence regions can either be next to the transcribed region, where they're called the promoter, or separated by many kilobases where there called enhancers or silencers. The promoter is located in the 5' end of the open reading frame and is composed of a core promoter sequence and the proximal promoter sequence. The core marks the start for transcription by binding RNA polymerase and other proteins necessary for copying DNA to RNA. The proximal promoter region binds transcription factors that modify the affinity of the core promoter for the RNA Polymerase. Genes may be regulated by multiple enhancer or silencer regions that might be located many thousands of base pairs away. The binding of different transcription factors, therefore, regulates the rates of transcription initiation, at different times and in different cells.

A key feature of eukaryotic genes is that their transcripts are typically subdivided into exons and introns. Exon regions are the coding regions of the mRNA and are retained in the final and mature

mRNA, while the introns are non-coding and are spliced out or excised during post transcriptional processing. Indeed, the intron regions of the gene can be considerably longer than the exon regions. Notably making up 26% of the overall genome. Once they are spliced together, exons form a single continuous protein coding region and the splice boundaries are no longer detectable.
Eukaryotic post transcriptional processing also adds the 5' cap and the poly-A tail to the end of the mRNA. These additions stabilize the mRNA and direct its transport from the nucleus to the cytoplasm where it can undergo the translation process and the production of the protein.



The overall organization of prokaryotic genes is markedly different from that of the eukaryotic cells. The most obvious difference is that prokaryotic open reading frames are often grouped together into a structure that is called a polycistronic operon, which is under the control of a set of shared regulatory sequences. These ORFs are all transcribed as a single mRNA molecule. Some operons also display translational coupling and will be translated together. The operator switch, next to the promoter, is the main regulatory element in prokaryotic cells. Repressor proteins bound to the operator may physically block the promoter in the association of the RNA polymerase enzyme preventing transcription. Riboswitches are another important regulatory sequence commonly present in prokaryotic untranslated regions. These sequences switch between alternative secondary structures in the RNA, depending on the concentration of key metabolites. The secondary structure can either block or reveal important sequence regions, such as ribosomal binding sites, RBSs. Introns are extremely rare in prokaryotes and, therefore, do not play a significant role

in prokaryotic gene regulation.

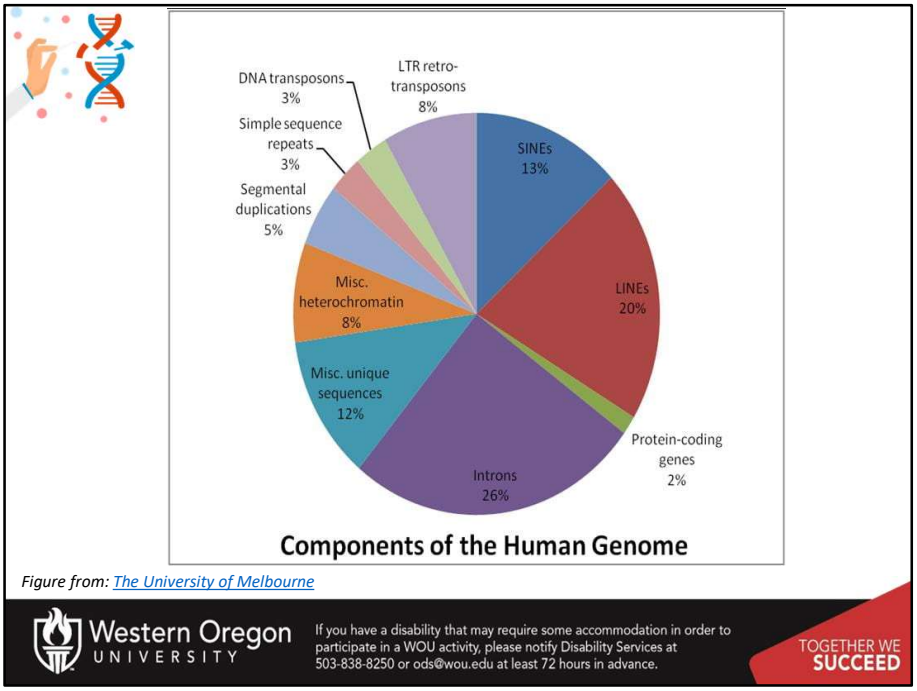


Figure from: The University of Melbourne

So eventhough coding sequences only make up 2% of the human genome, this leads to the production of over 46,000 proteins and another 2300 microRNA molecules.

Thus, in biochemistry, we have a lot to learn! This class really only brushes the surface. We will spend the rest of the term learning about the function of proteins and the processes

regulating gene expression.