Overview of Genetic Variation at Molecular Levels

Introduction

Genetics is the study of genes and genetic material in organisms. Genetic material is broad and expansive in humans, causing everyone to look so different from each other. There are various ways that genetic material can vary and many different mechanisms that are responsible for the differences. The list of these mechanisms is exhaustive, but here only the most important will be covered. To understand variance in human phenotypes, you must understand the structure of DNA, DNA folding, the Central Dogma, splicing, and mutations.

Structure of DNA

All of your genetic material is stored in DNA, so DNA must be structured in the most preservative way possible. This best structure is commonly known as the double helix and is how DNA is commonly represented. The double helix structure can be attributed to three components, the phosphate group/phosphodiester bond, the deoxyribose sugar, and the nitrogenous base. The backbone of DNA is made by the phosphate group, phosphodiester bond, and deoxyribose sugar. The structures align themselves into the backbone as a phosphate group attached to a phosphodiester bond, attached to the five prime positions of the deoxyribose sugar, with another phosphodiester bond attached at the three prime ends, attached to another phosphate group. This structure repeats fully along the DNA strand. Your genetic material is actually in the form of nitrogenous bases. The nitrogenous bases in DNA are known as Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). At this point, it is important to understand that DNA is double-stranded, with both backbones being composed of the same elements. Since the DNA is double-stranded, the nitrogenous bases pair with each other across strands, usually in the pattern of A to T and G to C. The nitrogenous bases can be present on any strand, but they must pair with their respective base. The bases pair with each other through hydrogen bonds, which are the weakest bonds we see in the DNA structure. This is because during replication, the strands must be unzipped, and hydrogen bonds allow this to happen easily. The number of bonds varies for the bases, as A and T pair with two hydrogen bonds, and G and C pair with three. This difference is because the structure of the bases themselves differs, as A and T are known as purines, and G and C are pyrimidines. For now, this distinction is only important in knowing how many hydrogen bonds are between two bases. DNA is also antiparallel. Thinking back to the earlier description of the structure of the backbone, notice how the positions of the deoxyribose sugar are mentioned. This is because one strand will run five prime to three prime, with the five prime of the deoxyribose sugar open on one end of the strand, and the three-prime end of the sugar open on the other end of the strand. The other will be upside down in comparison. So, the second strand will have its three prime ends open at the same end that the first strand has its five prime end open. The same is true for the other end of the strand, one will have its five prime end open, and the other will have its three prime end open. Acknowledging this property moves us into the territory of deciphering which strand will be leading and which is lagging. However, next, we

will cover further genetic variables, genes. For further understanding, see Figure 1 as it acts as a visual aid for what a nucleotide looks like.

Genes and Gene Architecture

The nitrogenous bases that were just mentioned code for all of your genetic information in the form of genes. Genes are spread all over your genome, and humans have a lot of them. They can be on either strand of your DNA, but when you are looking at one specific gene, you will call the stand with the gene on it the antisense strand, and the gene without the gene on it is the sense strand. This does seem backward, but when you think about replication, the strands are separated, and the strand is copied with the exact complementary bases. If you were to code the strand with the gene on it, the replicated sequence would be the sequence complementary to the gene, not the gene. Likewise, the code complimentary for the sense strand will be the gene. These genes code for amino acid and protein sequences. These protein sequences affect the resulting protein, many of which affect your physical traits. The discussion of genes brings up the discussion of Central Dogma, which will come up later. But, to best understand how genes are expressed, it is important to understand the architecture of the gene and the area around it. Every gene has a promoter region upstream (toward the 3' end) that will call the transcription machinery to the gene when it is ready to be transcribed. See Figure 2 for a visual of the architecture of genes.

DNA Folding

Although this topic does not specifically contribute to the concepts we will cover later in this chapter, it is important because specific mechanisms of folding can prevent any of the following steps we will discuss from happening. DNA is incredibly long, and if it was not packed into more condensed forms it would not be able to fit into our cells. For this reason, DNA goes through several forms of folding. First up, DNA is wrapped around packing proteins known as histones, into a formation known as a nucleosome. When DNA needs to be accessed, it is unfolded from histones through a cellular mechanism on the histone tail. This process could be prevented if there were methyl groups on CG-rich regions inside the promoter region of a gene. If these spots on the DNA are methylated, there is a malfunction in the unfolding mechanism, so the DNA remains tightly wrapped around the histone and can not be replicated. A later form of DNA folding happens when several sections of DNA are wrapped around histones and then tightly wrapped into a formation known as heterochromatin. This mechanism is mediated by methylation or acetylation. Methyl groups added onto the nucleosomes will tighten the DNA into heterochromatin, and similar to the earlier example on a smaller scale if those methyl groups cannot be de-methylated, the DNA packed into that formation will not be able to be replicated and no genes in that area will be able to be expressed.

The Central Dogma

For genes to be expressed and for humans to grow. Our cells undergo a process known as the Central Dogma, which is three mechanisms known as replication, transcription, and translation. There are several opportunities for genetic variation to be introduced during these processes, but first, you must have a basic understanding of each of them.

Replication

Replication is the process of all the genetic material in your cells replicating itself during a growing period. Several mechanisms are required for replication to happen, but the specific pieces don't need to be discussed now. The best understanding of replication is to understand that the DNA double-helix is unzipped, and is then replicated 5' to 3'. Both of the strands are replicated, causing DNA to have a semiconservative property. This means that when DNA is replicated, one of the strands is old, and the other strand it is bonded with is new. Through this property, two strands are created.

Transcription

Transcription is initiated when transcription complexes bind to the promoter region on a gene. During transcription, a certain gene is chosen to be changed from DNA into messenger RNA, or mRNA. RNA differs from DNA in three ways, first, RNA is only single-stranded, second, RNA replaces Thymine with Uracil (U), and third, the base sugar is ribose instead of deoxyribose. mRNA plays a role between transcription and translation. It is important to note that transcription results in a pre-mRNA that must be spliced to mature. This mature mRNA then will be able to leave the nucleus to be translated.

Translation

The mRNA created from transcription is then exported outside of the nucleus. A ribosome will then half assemble onto the mRNA to initiate translation. This is where transfer RNA comes into the game. These tRNAs, when charged, hold an amino acid that is coded for by a specific codon on the mRNA. When a ribosome assembles, it has three active sites that are crucial to understanding how translation happens. The first is the A site where the tRNAs enter the ribosome to try to bind to the mRNA. The second is the P site, where the amino acid attached to the charged tRNA is added to the growing amino acid chain. Finally, the third site is the E site, where the uncharged tRNA, meaning without the amino acid attached, is released from the ribosome. After initiation, there are usually two tRNAs in the ribosome, one in the P site holding the growing amino acid chain, and one binding to the A site to be added to the chain. After translation is completed, the ribosome detaches, and the product is an amino acid chain that codes for a protein.

Post Transcriptional Modifications

Now that the mechanisms and end products of transcription are understood, we can delve deeper into what transcription allows for genetic variability. In transcription, splicing heavily impacts

how a gene looks before it is translated. To best understand splicing we have to cover the concept of exons and introns. When a gene is transcribed, it is first characterized as pre-mRNA, because it still has both its introns and exons. This is where splicing occurs, because usually the introns will be taken out of the gene to stay in the nucleus, and the exons will be spliced together to make a mature mRNA. Splicing a gene is cutting and pasting different parts of a gene out and then putting everything back together. Through splicing, one gene could have several different forms because different exons or introns are put together in certain ways. This is because there are different splice sights on genes that give many options for the splicing of the same gene. This means, certain coding regions may or may not be present when the mature mRNA exits the nucleus, so the resulting amino acid sequence could be different.

Mutations

Mutations introduce a large probability of genetic variance and are the main cause of genetic variance at the gene level. Mutations can happen at either the DNA level or the chromosomal level and can have many different effects. First, we will review the types of mutations, and then we will examine these mutations at both levels and their effects.

Types of Mutations

The known point mutations are deletions, substitutions, and insertions. These mutations are often intragenic, meaning they are occurring inside a gene. Other mutations, happening at the structural, or chromosomal level, are insertions, deletions, inversions, translocations, fission, fusion, whole genome duplication, and whole chromosome duplication. Most of what the mutations do is pretty easy to guess based on their names, but some can be confusing. It is important at this point to understand that many mutations are spontaneous and are not caused by anything specific, although some mutations can be caused by UV radiation, we won't cover that specifically.

Intragenic Mutations

Intragenic mutations occur within a gene, so they happen on a small scale. Usually, these mutations are known as point mutations, because they are more than likely only one or two nucleotide changes made. Intragenic mutations are deletions, substitutions, or insertions. Based on their names you are probably able to tell that these mutations are the deletion, insertion, or substitution of a nucleotide. While this may not sound like much, single changes can lead to several different causes. First, deletions. Deleting one nucleotide shifts the reading frame, or what order the codons are in while they are read. For example, let's take the sequence 5'AUG-GGC-AAU-UGC-CGA-UGA 3'. The normal amino acid sequence would be 5'Met-Gly-Asn-Cys-Arg-Stop 3'. If we deleted the first Adenine, the sequence would instead read 5' UGG-GCA-AUU-GCC-GAU 3', and the new amino acid sequence would be 5'Trp-Ala-Ile-Ala-Asp3'. This amino acid sequence is different and probably changes the protein the amino acid sequence becomes. The same example could be used for insertion, except you add a new nucleotide

instead of deleting it. Substitutions are categorized as synonymous or non-synonymous mutations, meaning they either change or do not change the resulting protein sequence. Let's take the same sequence from earlier, 5'AUG-GGC-AAU-UGC-CGA-UGA 3', which codes for the amino acid sequence 5'Met-Gly-Asn-Cys-Arg-Stop 3'. If we were to substitute the G in the second codon to an A, the sequence instead would be 5'AUG-AGC-AAU-UGC-CGA-UGA 3', and it would instead code for the amino acid sequence 5'Met-Ser-Asn-Cys-Arg-Stop 3'. This is a non-synonymous mutation because the protein sequence is changed. However, if instead of the G, we changed the C to an A, the resulting sequence would be 5'AUG-GGA-AAU-UGC-CGA-UGA 3', and the amino acid sequence would remain as it was original, making this substitution synonymous, as the protein sequence was not changed. Although these mutations may make changes to the protein sequence, it is important to acknowledge that not every mutation is harmful and causes issues for the organism. Many could even be beneficial to the organism and increase its evolutionary fitness. Alongside this, it is possible that mutations can also be neutral until there is a reason for them to be seen as beneficial or harmful, as in a period of selection.

Chromosomal Mutations

Chromosomal mutations cover much larger areas of genetic material considering there is more genetic material available to them. Some chromosomal mutations are more beneficial or less harmful than others, as certain mutations cannot be tolerated by specific organisms. First up, insertions and deletions. Chromosomes carry a whole bunch of genes on them, so usually, when we mention deletions and insertions of chromosomes, we mean deletions and insertions of genes. There won't be any sequence examples, but the mutation of a deletion would be the deletion of one or more genes off of a chromosome, so the whole gene would no longer go through the Central Dogma process, and would not be expressed. Insertion is the opposite, we insert a new gene into a chromosome which would then be expressed in tandem with other genes that it may not normally interact with. Inversions are also fairly similar, if on a chromosome the sequence of genes, from top to bottom, was Gene A, Gene B, Gene C, and Gene D, in an inversion mutation, the sequence instead from top to bottom would be Gene D, Gene C, Gene B, and Gene A. This kind of change would affect what order genes are transcribed, but there is a possibility of issues with promoters that could cause issues with the genes being expressed. Transposable elements can be a bit more tricky, but they are essential elements that can move from their place on one chromosome and jump to another. Depending on where they land they hold the possibility to disrupt other genes. Fission and fusion are opposites of each other. In fission, one chromosome breaks into two chromosomes separating genetic material. In fusion, two chromosomes bind to one another to create one chromosome. Finally, this brings us to whole genome duplication. Humans are diploid, meaning we have two copies of each chromosome. In whole genome duplication, an extra copy of each chromosome is added, which would make up a triploid, with three copies of each chromosome. Few vertebrates can tolerate this change very well, but this is a version of speciation in plants. Then whole chromosome duplication, where one chromosome gets duplicated, leaving the organism with three copies of one chromosome. This type of

mutation is most prevalent in humans with down syndrome or trisomy 21 syndrome. Chromosomal mutations seem a lot scarier than point mutations simply because they happen on a larger scale, but it is important to remember that, like intragenic, the mutations made at this level are not always harmful and can often be beneficial or neutral.

Conclusion

Although the topics covered here are certainly not the only causes of genetic variation, they are the causes of genetic variation that happen at a more genetic level instead of cellular or allelic. Genetic variation at the molecular level is incredibly prevalent and a large cause of physical differences amongst species. Here, we have gone over DNA structure, DNA folding, the Central Dogma, Post Transcriptional Modifications, and Mutations. Most of the genetic variance is introduced in DNA folding, post-transcriptional modifications, and mutations, but the other topics are necessary to successfully understand the mechanisms needed for those variances.

Figures

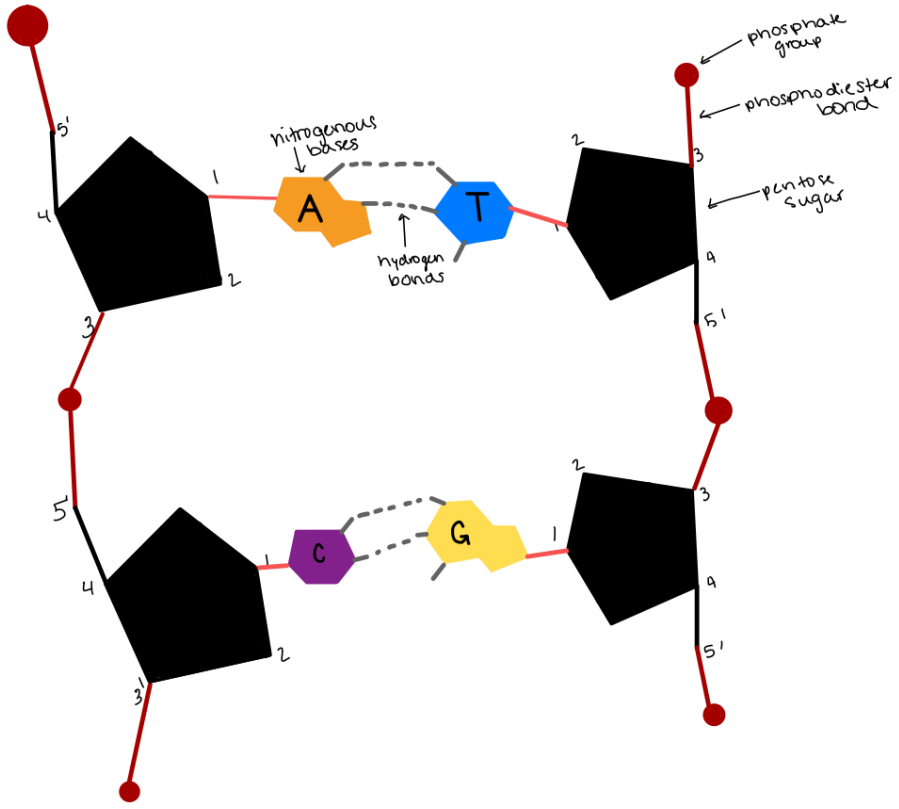

Figure 1: Depicts two nucleotides with the pentose sugar, phosphodiester bond, phosphate group, hydrogen bonds, and nitrogenous bases labeled. The nitrogenous bases are A (adenine), T (thymine), C (cytosine), and G (guanine)

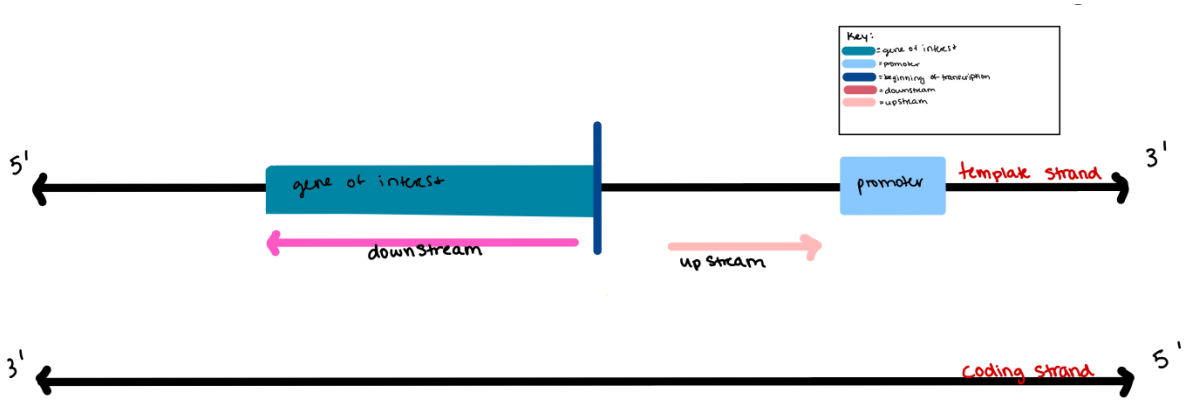

Figure 2: depicts the architecture of a gene and how it looks on a DNA strand. The template and coding strands are labeled, as well as the gene, the promoter, the start of transcription, and the terminology of location.

Resources

Deyholos, Michael, and Mike Harrington. *Open Genetics Lectures, Fall 2017*. Edited by John Locke and Mark Wolanski, 2017. https://canvas.okstate.edu/courses/151460/pages/textbook-pdfs-with-bookmarks.