**DNA, DNA Sequencing, and Genomes: A Look at What Makes Us, Us**

**Learning Objectives:**

* Understand what DNA is and the events that led to the discovery of its structure.
* Understand what DNA sequencing is and some of the different methods that are used.
* Understand what genomes are, the advancement of whole genome sequencing, and it’s application.

**DNA and its Discovery**

**Deoxyribonucleic acid**, or DNA for short, is a long molecule that contains the **genetic code** that makes you, you. It carries the instructions for growth, reproduction, and functioning for all life. DNA wasn’t discovered until the late 1860’s by a Swiss physician named **Friedrich Miescher**. He isolated many phosphate-rich chemicals from the nuclei of white blood cells, which he called “nuclein.” (Nuclein was later changed to “nucleic acid” and eventually deoxyribonucleic acid.) The importance of the discovery went unnoticed for years but it helped pave the way identifying DNA as the carrier of inheritance.

**DNA Structure**

The next big breakthrough for DNA was made in 1919 by Russian biochemist **Phoebus Levene**. He is credited for discovering the order of the three major parts of a single **nucleotide**. Levene suggested that nucleic acids, (DNA), is made up of a series of nucleotides and that each nucleotide is composed of one of four nitrogen bases, a sugar molecule, and a phosphate group. We now know that the sugar molecule found in the structure of DNA is **deoxyribose**. Also, there are two categories for the nitrogenous bases that make up DNA: the **purines**, which are **adenine** (A) and **guanine** (G), and the **pyrimidines**, which are **cytosine** (C) and **thymine** (T). Furthermore, after a scientist named **Erwin Chargaff** expanded on Levene’s work, we now know that A bases are always paired with T’s and G bases always pair with C’s. The order of these four nitrogenous bases make up the genetic code.

These discoveries about DNA helped contribute to **James Watson** and **Francis Crick’s** 3-D, double-helix model for the structure of DNA in 1953. Although scientists have made some small changes to the Watson and Crick model since its groundbreaking discovery, these four major features of the model remain the same:

* DNA is composed of two strands that appear twisted, referred to as a **double helix**. The two strands are connected by hydrogen bonds between the two pairing nitrogenous bases.
* Almost all DNA double helices are **right-handed**.
* DNA runs **anti-parallel**, which means the two strands run in opposite directions. The 5’ end of one strand is paired with the 3’ end of the other strand.
* The outer edges of the nitrogenous bases are available for hydrogen bonding also. This provides easy access to the DNA for other molecules that are involved in DNA processes.

**Sanger DNA Sequencing**

Soon after the structure of DNA was determined and the importance of DNA began to come to light, scientists started finding techniques to figure out the order of nucleotides in DNA strands. This is known as **DNA sequencing**. A British biochemist named **Frederick Sanger** is credited with most of the basic groundwork for DNA sequencing. The **Sanger method**, or dideoxy method, was developed in 1977 and was the most widely used DNA sequencing method for almost 40 years. This method relies on a special kind of nucleotide called dideoxyribonucleotide triphosphate (**ddNTP**.) ddNTP’s are kind of similar to normal DNA nucleotides, however, they are different enough to where they stop DNA replication. Once a ddNTP is added to a growing DNA molecule and replication is stopped, DNA fragments are made. After the fragments are made, they are heat denatured and separated by size using gel electrophoresis. As the fragments pass through the machine, a laser reads a fluorescent tag on the ddNTP’s, which reveals the order of nucleotides.

**Shotgun and Next-Generation Sequencing**

Although the Sanger method has stood the test of time, it can only be used to sequence pretty short strands of 100 to 1000 base pairs long. A different method that’s been used to sequence long DNA strands is called **shotgun sequencing**. Shotgun sequencing randomly breaks up DNA into numerous segments, and is then sequenced using the Sanger method to collect “reads.” After performing several series of fragmentation and sequencing, multiple overlapping reads are collected. These sequenced fragments are then but into computer programs that use the overlapping reads to put together a continuous sequence.

This classical shotgun sequencing was based on the Sanger method. The shotgun method is still used today, but it typically uses newer and more advanced sequencing technologies called high-throughput methods, or **next-generation sequencing**. These next-generation methods create shorter reads but can produce thousands or even millions of reads in a reasonably short amount of time. This results in a high amount of data, but assembling the full sequences is much more computationally rigorous. This makes the next-generation sequencing methods immensely superior to Sanger sequencing.

**Genomes**

A **genome** is an organism’s complete set of DNA. It includes the genes, the noncoding DNA, as well as the genetic material of the mitochondria. The genome includes all of the information needed to maintain and build that organism. The size of the genome depends on the complexity of the organism. Bacteriophages can have a genome that is only thousands of base pairs long while the human genome contains more than 3 billion base pairs. That vast amount of genetic information is contained in all cells that have a nucleus.

**Whole Genome Sequencing**

**Whole genome sequencing** is the process of determining the entire DNA sequence found in an organism’s genome. Bacteriophage ΦX174 was the first full DNA genome to be sequenced, and it contained 5,386 nucleotides. Frederick Sanger and his team completed this feat in 1977 by using a “Plus and Minus” technique, a method even more primitive and rigorous than the Sanger method (which was shortly invented after.) Perhaps the next big event in genome sequencing was in 1995 when an American biochemist named **Craig Venter** and his team successfully sequenced the entire genome of the bacteria *Haemophilus influenza*. This was the first free-living organism to have its entire genome sequenced, and its genome consists of 1,830,140 base pairs. This was important because it showed for the first time that random shotgun sequencing could be used on whole, large genomes with accuracy and speed. A couple months after the completion of the *H. influenza* project, the whole genome shotgun method was successfully used to sequence the genome of bacteria *Mycoplasma genitalium*. This method has been used to sequence the genomes of many organisms since then.

**Human Genome Project**

**The Human Genome Project** (HGP) was a 15-year-long international scientific research project with the goal of determining the DNA sequence of the entire euchromatic human genome. The HGP resides as the world’s largest collaborative biological project to date. The United States government picked up the idea in 1984 but wasn’t formally launched until 1990. This project included sequencing a couple of individuals and then assembling it together to get a complete sequence for each chromosome. Therefore, the finished human genome the project published in 2003 does not represent the genome of a single person. Also, the HGP did not technically sequence all of the DNA found in human cells, it only sequenced the euchromatic regions of the genome. The euchromatic regions of the genome make up 92% of the human genome. The other regions that were not sequenced are called heterochromatic, which are found in telomere and centromere regions of the chromosomes.

**Whole Genome Sequencing Human Application**

The cost of whole genome sequencing has drastically lowered in recent years, making it a more practical application in DNA analysis. It can be used to reveal large amounts of genetic information in a person such as genetic risk factors for onset diseases, carrier status for autosomal disorders, and other medical information that is not yet fully understood. It also creates huge potential for therapeutics and diagnostics. The end goal with applying whole genome sequencing on people is to create precise and **personalized medicine**. The more you can understand a person’s genes and the information on the genes, the better you can treat that person.

**References**

* Pray, L. (2008.) “Discovery of DNA structure and function: Watson and Crick.” *Nature*

*Education* 1(1):100

* Kratz, R. F. (2018.) “How Biologists Read a Gene with DNA Sequencing.” Retrieved

March 06, 2018, from http://www.dummies.com/education/science/biology/how

biologists-read-a-gene-with-dna-sequencing/

* Genome Unlocking Life's Code. (2017.) “Haemophilus influenzae becomes first

bacterium genome sequenced.” Retrieved March 06, 2018, from

https://unlockinglifescode.org/timeline/17

* National Human Genome Research Institute. (2015.) “All About The Human Genome

Project (HGP).” Retrieved March 06, 2018, from

https://www.genome.gov/10001772/all-about-the--human-genome-project-hgp/