Susanne B. Haga

The Book of Genes and Genomes

Springer

The Book of Genes and Genomes

Susanne B. Haga

The Book of Genes and Genomes

Springer

Susanne B. Haga
Duke University Center for Applied Genomics and Precision Medicine (CAGPM)
Durham, NC, USA

ISBN 978-0-387-70915-4 ISBN 978-0-387-70916-1 (eBook)
https://doi.org/10.1007/978-0-387-70916-1

This Springer imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

Swirling around in the core of our cells and those of every living creature is a twisted work of art. Physically twisted, that is. In all its simplicity, DNA is the common thread of all life-forms, big and small. If we look closely enough, we can find traces of our past and current life, as well as hints of our future. A personal barcode of sorts, each person's DNA is often regarded as the indisputable proof of identification. So, what does it all mean and why should I care? How does it affect me, my family, the environment, and society?

Skim the daily headlines on your phone or computer, and you are bound to see a story announcing some new discovery about a gene related to a disease, virus, behavior, political preference, or anything else you could imagine. Or maybe it is a story about the Neandertals and their relationship with modern humans, as deduced from genomic sequencing. Or maybe someone is proposing to bring back an extinct species, say the woolly mammoth (think Manny from Ice Age!). If you do not read the news, check out literature, television, and movies, and you will find that these fictional outlets have been keen to incorporate DNA and genetics into storylines: Jurassic Park's recreation of dinosaurs, CSI's and Law and Order's (and all of their spin-offs) reliance on DNA analysis to identify criminals, and genetically engineered (accidental or intentional) action movies (e.g., Minority Report, Wolverine, Jupiter Ascending, Spiderman). You can call it "science fiction," but it appears to have escaped that category, blending fiction and fact to create stories that aim to impress us with the precision and power of genetics and now genomics.

Therefore, it is almost impossible to avoid even if you have tried to—genetics and genomics are literally everywhere. It is actually possible that you have become desensitized to all of the news stories. Can everything possibly be linked to DNA you might wonder? I would argue probably yes, since DNA is present in almost every living organism on Earth (and we are now looking for it beyond Earth). But despite the ubiquity of DNA-related stories, scientists are still very much uncovering the secrets of DNA. It really was not until the

last 100 years that scientists got a handle on what exactly was passed on from generation to generation (DNA), what DNA looked like (a double helix), and finally the sequence of the unique code (the order of the four-letter alphabet). What most people do not realize is that scientists are really trying to read a book written in a language completely unlike any other. Every species has its own genetic "codebook," although there are shared instructions between all living organisms. To further complicate matters, every member of a given group will differ just slightly in their codebook. And, on top of that, the genetic codebook (or parts thereof) can have different meanings if "read" in different environments. Confused yet? Welcome to the world of genetics (and science for that matter).

Ironically, my interest in genetics was sparked by its apparent simplicity and precision (as it was presented to me in ninth grade). And as I continued with my education in human genetics during a time of rapid advances in scientific knowledge and technology, it became very clear that genetics is anything but simple and precise. I believe that this also inspired my passion in education to help others understand genetics, either very broadly as this book attempts to do or about one specific genetic test or application. DNA is a chemical, but not to worry—this is not a chemistry book. However, in order to describe some of the medical and nonmedical applications, I have ventured a little into the science of genetics that you may have avoided in school. By bringing together the wide range of applications based on DNA science and medicine in this book with a little scientific explanation, my ultimate hope is to leave you with some knowledge to make you a more critical reader about genetics and genomics and, if needed, to make informed decisions about your health or that of your family members.

Along with the scientific discoveries and the exciting applications developed thus far, we must also consider the ethical and legal concerns and the potential adverse consequences raised by this newfound knowledge and technology. Genetics has benefitted and suffered from the intense focus of recent decades and the negative history associated with eugenics. Thus, not surprisingly, it draws a range of public responses from fascination to fear and trepidation. With the rapid advances in genetics and this new field called genomics, both practical questions (who has access to new applications, who pays for it) and ethical questions (should we really be doing this) have followed the science, sometimes with unclear answers. We all must continue to ask these important questions. A better understanding of the science and related applications should promote more informed and greater public engagement.

Growing up in the pre-genomic era, before fancy sequencing machines could whip out DNA codes, I was in awe of the power of DNA. My

amazement continues to grow as I have witnessed how the knowledge of genetics and genomics has transformed medicine and society in such a relatively short period. It is my goal to share some of this excitement to the readers of this book and possibly create a lifelong follower of the field while imparting a bit of knowledge. Undoubtedly, something has been discovered or developed after I finished writing—that is the difficult part about writing a book on this field. There continues to be so much to learn, and I hope that this is just the start for you.

I am indebted to all of the wonderful teachers, mentors, and collaborators that I have learned from and been inspired by. May I continue to pay it forward.

Durham, NC, USA Susanne B. Haga

Contents

1

From Genes to Genomes in All Living Things

Although this book aims to present several applications of genetics and genomics, if I actually started off with that, I would have to keep stopping to explain the science a bit. Thus, this first chapter aims to provide readers with a short overview of the history of genetics and genomics, starting at a time when the words *genetics* and *genomics* did not yet exist to the present.

Most people are probably quite familiar with the words "gene" or "genetics." A word association game may yield words like family, health/disease, and identification (e.g., paternity, forensics). In contrast, the word "genome" (pronounced jee´nōme) is much newer to our lexicon and thus likely to be quite unfamiliar to many. The genome refers to the entire DNA content found in a given cell (as opposed to a gene, which is one very small part of the genome). Although the term genome was coined in 1920, from the words *gene* and chromo*some* (a condensed form of DNA to be explained more later), it really did not garner much attention by the scientific or medical community until fairly recently, beginning in the 1980s and 1990s, as will explain why later.

In the early part of the century, the knowledge and technology were not available to enable scientists to fully understand the human genome, let alone a single gene. The stepwise process of scientific research can seem painfully slow, but lots of information was being learned about basic cellular processes that are taken for granted today. Now we are able to analyze an unknown sample of DNA extremely rapidly and determine from which species the DNA was from, and potentially determine the exact individual human or animal it came from. But despite the huge advances afforded by new scientific technologies and the generation of a lot of data, scientists are still looking for

S. B. Haga, *The Book of Genes and Genomes*, https://doi.org/10.1007/978-0-387-70916-1_1

answers to questions regarding human health, environment, and other fields that may reside in genetic material (or DNA). Imagine trying to put a 3-billion-piece rocket ship together that you do not have the instruction manual for and then try to figure out how it is supposed to work. Now you will begin to get some inkling of the challenges faced by geneticists and genome scientists in trying to uncover the secrets stored and what each "part" does within a genome.

The 1800s

Flashback, if you will, to a time of intellectual curiosity, and of relative peace and prosperity, and the quiet sanctity of an Augustinian monastery. During the mid-1800s, the town of Brünn, Austria (now Brno, Czechoslovakia), was part of the Austrian Empire and later the Austro-Hungarian Empire. Brünn was a hub for textile industries and agriculture—particularly wool and fruit. In 1850, the town's population had reached about 47,000. Perhaps an unlikely place to begin a chapter on genetics and genomics, but this has been referred to as the origin of the theories of inheritance and genetics.

In 1822, Johann Mendel was one of five children born to a peasant family. His father was a farmer, managing his own crop of fruit trees as well as tending to the fields of the feudal lord whom he worked for 3 days a week. Mendel excelled in school but was shy by nature and often needed to return home due to illness caused by stress. During his pre-university studies, Mendel focused on physics and math. Due to a lack of funds though, he did not immediately continue his education at the university level. His physics teacher recommended Mendel to the Abbot at the Brno monastery for the novitiate (a sort of mentorship program for prospective candidates of a religious order who have not yet been admitted).

In 1843, he was admitted to the Augustinian monastery and was ordained in 1848. Although Mendel was not of deep religious faith, entry into the monastery was a way to continue his education and training in science. He took the name Gregor after entering the monastery. He first served in a role similar to a parish priest to a parish affiliated with the monastery, which included responsibilities like tending to the sick at a nearby hospital. However, the stress due to the constant suffering and pain he witnessed took an emotional tool and he became ill and depressed himself.

Recognizing his struggles in providing comfort to the ill, the Abbot reassigned him to teach math and science to seventh graders. In 1850, as required by law, he took the exam to be a teacher of natural history and physics, but

failed. In 1852, to address his apparent deficit of knowledge in the sciences, he attended the University of Vienna and learned from several well-known research scientists of the time. He recalled that the Abbot had mentioned that the mystery of heredity would only be solved through rigorous experimentation. Upon his return to the monastery in 1853, Mendel began studying pea weevil plants. He was familiar with the techniques of artificial fertilization learned during his childhood experiences with fruit trees.

In 1855, he began some experiments that would support his now famous work on the theories of inheritance, which he would publish 11 years later in the proceedings of the Brunn Society for Natural History. So what exactly did Mendel figure out with his simple pea plant experiments? He presented his findings in 1865, opening with the following introductory remarks:

> Experience of artificial fertilization, such as is effected with ornamental plants in order to obtain new variations in color, has led to the experiments which will here be discussed. The striking regularity with which the same hybrid forms always reappeared whenever fertilization took place between the same species induced further experiments to be undertaken, the object of which was to follow up the developments of the hybrids in their progeny.

To understand the significance of his work, it is helpful to backtrack a moment to provide some context of Mendel's work. For centuries before Mendel, it was well known, by farmers and ranchers in particular, that the qualities of the next generation's crop were dictated in part by the parents. Two parents with certain desirable traits would be mated to produce the next generation with the same desirable traits, giving rise to "pure-bred" strains. Experiments in the 1700s on plant hybrids were beginning to shed light on the inheritance of traits that were believed to have informed Mendel's thinking. Oftentimes (as was expected), the plant hybrids represented a mixture (blending) of the parental traits, but some would occasionally appear more like one parent than the other. The understanding of whether plants sexually reproduced still was not clear (as was readily obvious in animals).

More than 100 years prior to Mendel, the observation of predictable transmission patterns of human diseases, particularly those that affected only one sex, had been recorded by a number of scientists. In 1794, the English chemist John Dalton noted that he and several of his male relatives were affected with color blindness, a condition now understood to predominately affect males.

Mendel's success was due in part to his choice of organisms (the pea plant) and his selective study of traits (those with only two possible outcomes) rather

than more complex traits with multiple potential outcomes. Traits such as pea color (yellow or green) and texture (smooth or wrinkled) produced unambiguous results (Fig. 1.1). Based on repeated observations and tracking of multiple generations of peas for selected traits, Mendel deduced that each trait was due to the combination of two versions (later defined as an "allele") of a gene in each plant. He did not use the term "gene" since it had not been coined until the early 1900s. Instead, Mendel actually used the word "factor" to describe a unit of inheritance, and deduced that one factor was derived (or inherited) from each parent. During reproduction, these two copies would separate and each parent would donate one or the other copy to the offspring (Fig. 1.2).

The breeding of peas for certain traits for several generations revealed that some traits were dominant compared to others. In other words, some traits that appeared in the first generation of offspring were referred to as *dominant*, and those that appeared in the second generation as *recessive*. For example, if a round pea was bred (or crossed) with a wrinkled pea, all of the offspring in the next generation were observed to be round (and thus, the trait roundness was determined to be dominant). However, if two round peas from this first generation were crossbred, the offspring were a mix of round and wrinkled (defined to be recessive). Although Mendel was not the first person to describe

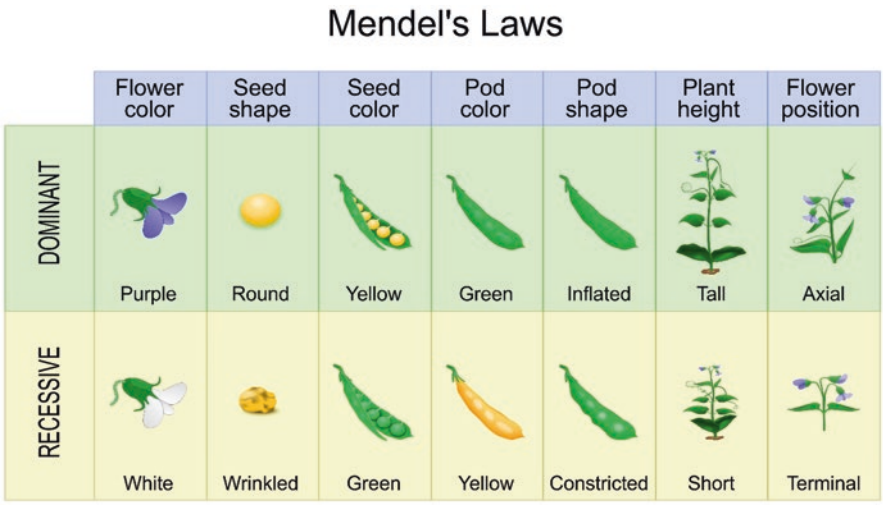

Fig. 1.1 Peas were an ideal organism to study the passage of traits from generation to generation given the wide array of distinct features and multiple combinations possible, e.g., seed form and flower color (source: Adobe Stock)

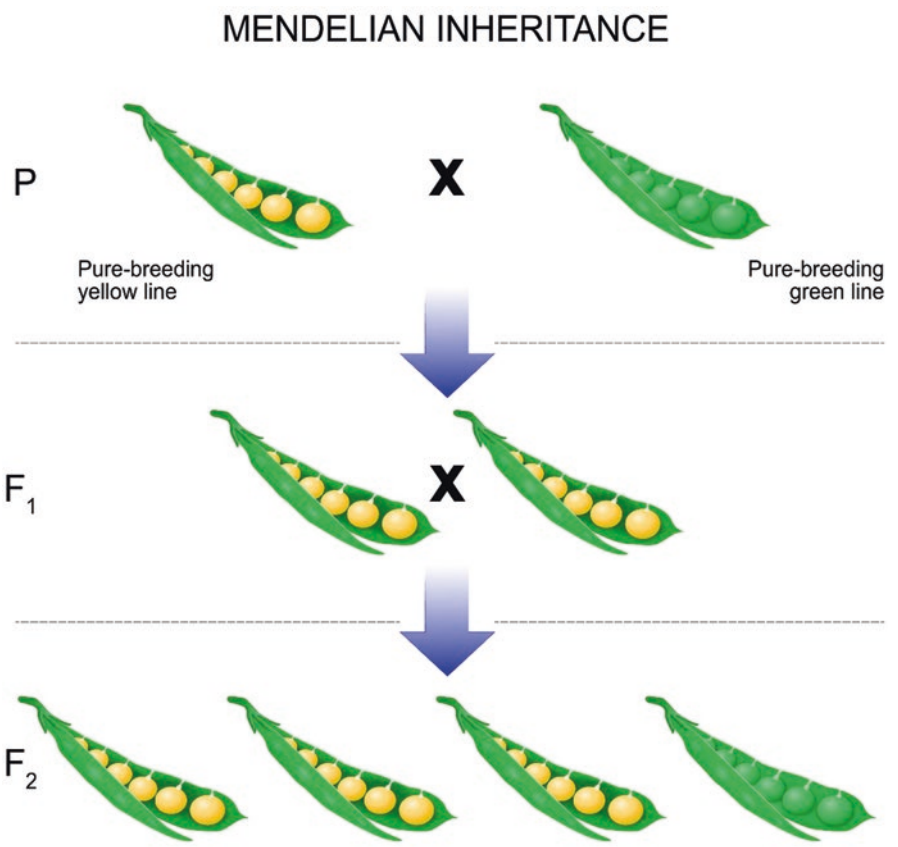

Fig. 1.2 Illustration showing the passage of traits from parents to offspring (P = parents; F1 = first generation; F2 = second generation) (source: Adobe Stock)

the idea of dominant and recessive traits, his experiments unequivocally proved the concepts.

We now know that genes were these "factors" that Mendel described and that they contain the instructions to make proteins, the molecules that actually do the work required of our cells and bodies that would give rise to certain traits or characteristics. If, by chance, one of the genes is mutated or altered, the protein produced from that gene may not work properly, correctly, and/or as efficiently. If the bodily function in which that gene participates requires the normal dose from two copies of the gene to perform its function, the absence of one copy due to a mutation (alteration in the gene) may lead to disease. In this situation, the disease is said to be dominant as only a single mutated copy of a gene will give rise to it. On the other hand, if both copies need to be mutated in order to give rise to a disease, it is referred to as

recessive. In this situation, both parents would likely be carriers of a mutated gene, but neither would be affected. Each parent would have a 50% chance of passing on the mutated gene to the offspring, for a 1 in 4 chance of having an offspring with the recessive disease (inheriting two copies of the mutated gene).

When Mendel studied two or more traits together (e.g., round and yellow peas vs. green and wrinkled peas), he also observed that each individual trait was passed on independently of the other trait and the outcome of one trait was not linked with the outcome of the other traits. All combinations were possible (round/yellow, round/green, wrinkled/yellow, wrinkled/green), although as with the single traits, each combination would appear with different frequencies in every generation. After collecting data on thousands of pea plants and using his assumptions about dominant and recessive alleles, genetic segregation, and independent assortment, Mendel eventually could calculate with precision what ratios could be expected in a given generation for a given trait.

Shortly after his publication, in 1868, Mendel was appointed as the sixth Abbot of the Brunn monastery. With much of his time now devoted primarily to administrative duties, he had far less time to spend on scientific experiments and observation, which appeared to have ceased in 1871. At the age of 61, Mendel died in 1884 without any recognition of the significance of his work by his peers.

More than three decades passed before the significance of Mendel's work was realized. Why did it take so long? One reason was that scientists were focused on other pressing issues of the time—namely that of Charles Darwin and his theories of evolution. It was a challenge, to say the least, to make sense of how Darwin's theories of natural variation and Mendel's theories of inheritance fit together, especially as each was still unclear on its own. More theories on heredity were developed between Mendel's publication and its discovery in the early 1900s, which greatly benefited from advancements at the time in laboratory experimentation, microscopy, and evidence that particular traits or diseases run in families; that is, multiple family members in multiple family generations were affected with the same condition. Eventually, Mendel's "factors" and theories were rediscovered, further advancing the young field of genetics and understanding of inheritance (Fig. 1.3).

The 1900s

As in any field of science, discovering the answer to one question only leads to more questions. Understanding the process of inheritance was only part of the

Fig. 1.3 Commemorative stamp in Czechoslovakia of Gregor Mendel (source: Adobe Stock)

puzzle to be solved. The 1900s were a time of rapid discovery (and rediscovery), when the puzzle pieces seemed to finally fall into place one by one.

The year 1900 marked the recognition of Mendel's work, when three scientists independently noted the significance of Mendel's work. During that year, the British zoologist William Bateson mentioned Mendel's work during a scientific presentation in London. In 1904, Bateson even visited Brno to learn more about Mendel, but no one was able to tell him much about the quiet man and his scientific experiments. In 1902, the British physician Archibald Garrod published his work on the biochemical disease known as alkaptonuria, and determined that this disease was inherited in a "Mendelian" recessive fashion. In 1906, Bateson coined the term "genetics" to describe the study of

heredity. In 1909, Wilhelm Johannsen introduced the word "gene" to denote the unit (the ambiguous term previously used by Mendel) that was passed from parent to offspring.

Confident in their understanding that something was being passed on from generation to generation in predictable patterns, the next big question facing scientists was to determine what exactly was being passed on from parent to offspring—i.e., what exactly was a gene made of? While the European and British scientists had made great contributions in establishing the foundation for the field of genetics, US scientists began making up for lost time with a series of experiments that would confirm that genes were made of DNA. In 1901–1902, separate experiments proved that half of the chromosomes were passed on from the mother and half from the father. Subsequent experiments conducted by US scientists in bacteria and viruses proved that the hereditary material was indeed DNA.

The next question was what exactly did DNA look like—or what was its structure? This was important to learn, as by understanding its structure, scientists might then be able to deduce how it actually functioned, replicated, and was passed on from parent to offspring. It was known that DNA was comprised of four chemical units, abbreviated A, T, C, and G. And it was also known that the number of As and Ts was equal to the number of Cs and Gs in a given sample of DNA. However, the AT/CG ratio differed between species. But how these chemical units were assembled was unknown.

In 1953, the stunningly simple structure of DNA was revealed by the American scientist James Watson and British scientist Francis Crick working together in Cambridge, England. Clues from a special type of X-ray photograph of DNA led them to hypothesize that the structure of DNA was some type of helix. After several attempts to arrange the As, Ts, Cs, and Gs through chemical models, they realized that DNA resembled a twisted ladder, whereby the chemical units were located on the "rungs." Based on the earlier observation about the proportion of As and Ts and Cs and Gs, Watson and Crick predicted that the A units paired only with the T units (A–T) and the Cs paired only with the Gs (C–T) (Fig. 1.4). When cells divide, the DNA has to make a faithful copy of itself to pass on to the next generation or "daughter" cells. Based on this double-helical structure, it was proposed that the DNA strands come apart like a zipper.

These "parent" single strands then serve as the template for two new strands. Thus, half of each new DNA molecule was composed of the original parent strand and a new strand. Because of the understanding that As only connect to Ts and Gs to Cs on the rungs of the DNA ladder, the newly formed DNA

Fig. 1.4 The pairing of DNA subunits (source: BioRender)

was the exact same code or sequence of chemical units present in the parent DNA.

In the years following the discovery of the structure of DNA, much research was done to understand how the DNA code was "read" by the cell's machinery, giving rise to the production of proteins, the molecules that are the "workhorses" of the cell. The chemical units of DNA by itself are not functional. It was eventually determined that the sequence or order of the DNA letters (A, T, C, and G) is read three letters at a time—each three letters encodes an amino acid, the building unit of proteins.

Can We Actually "See" DNA?

How do scientists actually study DNA—is it something that is visible? Yes and no—DNA is visible under some circumstances, but the actual order of the bases (As, Ts, Cs, and Gs) is not visible. When the concept for a microscope was described in the late 1500s and actually built in the mid-1660s by Antonie van Leeuwenhoek, this was a tool that enabled scientists to eventually see the structure of cells and the organelles (the "organs" of a cell) for the first time and to observe the changes that occurred as cells grew and divided. In 1882, scientists first visualized cellular structures called chromosomes (kromo-zomes), which is DNA condensed in coil-like structures (to be discussed in more detail in Chap. 2). Chromosomes resemble squiggly wormlike structures located in the center of cells. One could actually see that there are multiple chromosomes within cells and later determined that different organisms had different numbers of chromosomes. It was then determined that sperm and eggs had half the number of chromosomes observed in other types of cells (and thus, their union would reconstitute the full number of chromosomes). In 1889, it was proposed that the hereditary material was passed on in these chromosomes, but no experimental evidence had yet proven this theory. Thus, the strands of DNA were not visible, but DNA stored in the chromosomes

was helping scientists learn about DNA replication and movement each time the cell divides.

Furthermore, scientists did not need to "see" DNA to study it or to infer the consequences of genetic changes.

But how does one actually find a mutation? It turns out that the human genome is three billion units long, and contains approximately 20,000 genes. Therefore, searching for just one chemical unit that has been mutated out of 3 billion may take some time. The analogy "searching for a needle in a haystack" aptly describes the unbelievably difficult process of finding a gene linked to a particular disease until this century.

In the first half of the twentieth century, scientists were able to map (or determine the location thereof) a gene believed to be responsible for a specific trait (e.g., eye color) to a specific part of a chromosome. Each chromosome contains many genes, so narrowing the region in which the putative culprit gene was located was a very important first step. This type of mapping is somewhat analogous to determining that a house of interest is located in the city of Baltimore, and possibly narrowing down the part of the city to a certain part of town (e.g., North). But no detailed maps of each chromosome existed to further help navigate (a chromosome is linear), nor knowledge of what the region looked like or how far scientists would have to "walk" to find the culprit gene.

In the 1970s, a method to sequence DNA was worked out for the first time. Scientists could decode the A, T, C, and G's for a given piece of DNA (Fig. 1.5). Thus, once a culprit gene was narrowed down to a chromosome and to a particular "neighborhood" along the chromosome, scientists could "walk" up and down the chromosome by sequencing the regions of DNA to figure out what genes were located in a region and if a mutation resided in one. Several genes for Mendelian disorders were identified using this approach, though it could take years to discover since at that time, scientists were only able to "sequence" small segments of DNA.

In the 1980s, the idea to sequence or decode the entire human genome was introduced—the elucidation of the order of As, Ts, Cs, and Gs of the entire DNA content from a human cell. Considered to be the most ambitious scientific endeavor by many, the feasibility and utility of such an unprecedented effort were uncertain. Much of the genome was considered to be "junk" DNA or DNA of unknown significance. Genes actually accounted for a very small proportion of the genome, but they were the target for understanding the genetic causes of disease.

Fig. 1.5 Section of DNA sequence (source: Adobe Stock)

In the 1990s, technological advances enabled scientists to initiate a massive project to determine the full sequence of the human genome (all 3 billion letters). In particular, it was the development of automated sequencing machines and computational sciences required to analyze and store the DNA data that enabled the project to be completed. As a result, the sequence of the genome was completed, chromosome by chromosome, including the location of known genes, each with a specific coordinate (very similar to longitude and latitude that correspond to a unique location). The complete sequence of the human genome was finalized in 2003, 50 years after the structure of DNA was revealed.

What Have We Learned from Sequencing the Human Genome?

In the 1980's, an idea was being discussed amongst the scientific community about sequencing the human genome. After much debate and securing Congressional funding, the Human Genome Project was launched in 1990 with the singular but enormous goal of sequencing the human genome and creating a reference book of sorts for the scientific community. A draft of the