So Much "Junk"? The Complexities of the Human Genome

Luis Loor

Why do you look the way you do? What determines the color of your skin and eyes? What about your resemblance to your parents? In order to answer these questions, biologists have spent decades studying human genetics and the very intricate instructions that are located inside each of our cells. Safely stored inside the nucleus of cells lies the set of instructions that makes each and every one of us different, and also fundamentally the same: deoxyribonucleic acid, otherwise known as DNA. DNA are long, double-stranded molecules that look like twisted ladders. These molecular ladders are made from four types of nucleotide base units, called adenine (A), cytosine (C), guanine (G), and thymine (T). When these nucleotides are combined in sequence, they form our genome which contains instructions to direct all the processes in the cells and make it possible to build the components of our bodies and maintain them throughout life.

Naturally all of our cells maintain function and regenerate. In order to do this, all the cells follow the instructions encrypted in the DNA to undertake two very important processes, replication and transcription. During replication, DNA is copied in a very complex process. First enzymes called DNA polymerase attach to specific parts of the DNA. As the enzymes advance through the long DNA double helix, they split the DNA into two strands. The enzymes take both of those strands of DNA and use them as templates to create complementary strands, thereby replicating exactly the original DNA containing the same genetic information. The second process, transcription, describes how DNA is "read" to create proteins, which are the worker molecules that build cellular structures and maintain function. First the DNA is unwound, and then one of the strands passes genetic information to another molecule called ribonucleic acid (RNA). The RNA carries the information encoded in the desired gene, exits the cell nucleus, and, with the help of another enzyme system called ribosomes, processes the instructions on how to build complex protein.

Now you may ask, how is information organized in our DNA? To better visualize this we can say that each gene in our genome is a long series of codons, or a set of three-letter DNA words. Scientists have found that each gene contains on average 27,000 letters. If we compile a large set of long sequences of genes and the non-gene sequences between them, we get one chromosome. Humans

have a total of twenty-three pairs of chromosomes where we find the complete genetic data about a human, also called the human genome.

A genome can be compared to a "book" that holds the entire genetic information about an organism. This genetic information varies from one organism to the next. All told, the human genome is made up of about three billion nucleotide base pairs[1]. This sounds like a lot of genetic data and indeed it is! Imagine a large book like a dictionary with over a thousand pages. The human genome will fill 428 of those books! Now another question arises, one that scientists have asked themselves for a while: is all this information useful to our body?

The reason this question arises is because biologists are still trying to figure out where all the genes are and what they do. Science has made it evident that only about two percent of the human genome consists of DNA that codes for proteins[2], and as mentioned above, these proteins are essential to the creation of organs, tissues, and cells. Yes, protein-coding genes only make up a small part of the DNA. What about the other ninety-eight percent of genetic data that does not contain genes and does not code for protein? For some time scientists have called those parts "junk DNA," but recently they are changing their stance. They are finding that those parts considered useless actually control how and to what extent the genes are used.

How did the term "junk DNA" come about? In 1972, the geneticist Susumu Ohno came up with the term "junk DNA." In his paper "So Much 'Junk' DNA in our Genome," he used this term to describe "all noncoding DNA sections, most of which consist of repeated segments scattered randomly throughout the genome"**[2].** He was referring to the content of the human genome that does not contain functional genes. He compared the human genome to our planet: just as the earth contains many fossils of extinct species, our genome too is filled with the remains of extinct genes. His point of view greatly affected the study of genetics. Many biologists turned their backs on junk DNA and focused their research on the coding part of the human genome, which seemed more important. It wasn't until the 1990s that many scientists became interested in junk DNA and its repetitive elements. One can view these sections of DNA like advertisements in a magazine, which break up an interesting article. For instance, an article may start on page 5 and be followed by an ad on page 6, continue on pages 7 and 8 and then be followed by another ad on page 9. Even though the ads and the article are different types of information, both are important to the make-up of the entire magazine. In short, even though some genes don't transcribe for protein, it does not mean that they are rendered useless. This idea began to grow among scientists and led to many of them calling these sections of DNA "noncoding" instead of "junk."

As interest started to grow among biologists to try to understand the functions of the noncoding DNA, they noticed that these sequences "increase the ability of a species to evolve by serving as hot spots for genetic recombination and by providing important signals for regulating gene expression"[3]. Indeed these elements are components integral to our genome and can hardly be called junk anymore. Many scientists like John Rinn, professor at Beth Israel Medical Center,

have put a lot of time and effort into studying noncoding DNA. In an article in *Nature*, he explains that he "was not interested in looking at the map of known protein-coding genes on the chromosome, but rather everything else"[4]. He says, "We wanted to see if we could find biologically active molecules in the human genome that no one previously knew about"[5]. Because of his efforts and those of many other biologists, we are beginning to address the mystery of the function of noncoding genetic material. The way they see it is that, since noncoding sequences exist within the DNA, and since it makes up a large portion of the genome, it *has* to have a purpose. To illuminate this further, biologists have compared the human genome to a factory. Just as a factory has some devices that assemble parts and others that serve as controls and regulators, the genome has parts that encode for RNA and proteins, and other parts to control the process and serve structural roles. This is where noncoding DNA comes in: even though it does not directly code for protein products, it serves a purpose because a "substantial amount of noncoding DNA contributes to genome function"[6].

Scientists have now been able to classify certain noncoding sequences. First, we have the segments of DNA that exist within a gene itself because this is the area that has been most studied. In molecular biology the parts of the gene that code directly for protein are called exons, and the parts that don't are called introns. When the RNA is ready to be transcribed for protein, introns are removed and exons are linked up together in the process called splicing. This doesn't mean that introns are useless. The presence of introns in some genes allows for the gene expression enzymes to decide which introns to exclude from the final gene product. This process effectively increases variability of gene products without having to depend on mutations. Increasing variability means that the gene allows moderate changes done to itself by re-arranging the position of introns and therefore exons which results in the gene producing slight alternate versions of itself that can have positive benefits. Thanks to research like this it has become clear that the "changes in gene expression—primarily changes in noncoding DNA—have a tremendous impact on an organism"[7].

Another noncoding factor that is important in regulating gene expression is known as RNA interference. One known factor of RNA interference is the microRNA. This small RNA molecule made up of twenty-two nucleotides is known to be a post-transcription regulator. This means that its regulatory capabilities are used during translation of RNA into protein. MicroRNA originates from already transcribed RNA in the nucleus, and enters the cytoplasm along with the rest of the messenger RNA which will be translated. The microRNA's bind to a specific targeted part of the messenger RNA in order to regulate the amount of protein to be produced. This process is called alternative gene silencing because a specific part of the gene will be turned "off" or not used.

It has taken almost thirty-five years to clearly identify this type of RNA. Recent studies reveal that these microRNA's "suppress initiation of protein translation, promote messenger RNA degradation and turnover, and initiate transcriptional silencing"[8]. This is an important discovery because, as some

scientists imagine, microRNA's can act as tumor suppressors by regulating cellular replication of cancer cells.

Other types of regulatory sequences are located either close to a protein-coding gene or far away from it. These sequences that sit next to the gene are called *cis* regulatory sequences and they serve as markers by guiding the transcription enzymes by signaling where it can start the process of making RNA. The regulatory sequence itself does not get transcribed. The specific sections of DNA onto which the enzyme attaches are called transcription binding sites, and it is here that transcription of DNA begins.

As we can see, noncoding sequences do have important function, and they are required for many important cellular processes. Since some of these sequences regulate protein production, biologists use them to study variation in genetic sequences, like those that the introns are responsible for. In the past it was thought that "protein-coding variation was most likely the main source of disease susceptibility, mainly because most known genomics disorders are due to such mutations"[9], but now scientists are turning around and expanding their research to examine the vast amount of genetic data in noncoding sequences. Discoveries made in this area could help examine regulation of genes in a more precise and well-coordinated way. Heading in this direction could help uncover patterns of genetic sequences that may actually cause some genetic diseases. One way to illustrate the role of noncoding sequences is the following. Imagine holding in your hand a microchip that you want to insert in a computer. The chip itself can't do this task. It will require the use of one's arm muscles, bones and fingers to precisely place the chip in the right place. The same can be said about noncoding DNA sequences. Their mere size and complexity helps to support, manipulate and regulate the protein coding genes, which is represented by the microchip, in order for it to work properly. Damage to the supporting structure can therefore lead to malfunction, and in the case of the human genome, to disease.

Research on noncoding genes continues to this day, and it probably will for years to come. In the past it was challenging to analyze these long sequences due to the lack of computing power. With the advance in technology and in bioinformatics, it is now possible to study this part of the human genome in greater depth. It is becoming clearer that there exist complex interactions and functions between various elements of the human genome. Understanding the important connections between coding and noncoding sequences can help scientists discover new ways to prevent diseases, apply treatments, and perhaps even create cures. It is only when *all* genomic data is used that we can get a clearer picture of our complex genetic nature.

Notes

1. Bioinformatics Methods in Clinical Research, edited by Rune Matthiesen, 2010, p. 49.
2. A. Khajavinia, Ishafahan (May 2007). What is "junk" DNA, and what is it worth? *Scientific American*, p. 104.

3. Aria Pearson (July 11, 2007). Genomics: Junking the junk DNA. *New Scientist*. Retrieved from: http://www.science.org.au/nova/newscientist/078ns_005.htm
4. Rethinking Junk DNA (March 2009). *Nature Magazine* vol. 458 p. 24o.
5. Svetlana A Shabalina and Nikolay A Spiridonov (March 2004). The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology* vol. 5, issue 4, article 105, p. 105.2.
6. Emmanouil T. Dermitzakis (2008). Regulatory Variation and Evolution: Implications for Disease. *Advances in Genetics*, vol. 61, p. 195.
7. Cristian I. Castillo-Davis (October 2005) The evolution of non-coding DNA: how much junk, how much func? *TRENDS in Genetics* Vol.21 No.10.
8. Aaron T. Willingham and Thomas R. Gingeras. (June 2006). TUF Love for "Junk" DNA. *Cell 125,* p. 1215.
9. Emmanouil T. Dermitzakis (2008). Regulatory Variation and Evolution: Implications for Disease. *Advances in Genetics*, vol. 61, p. 303.

Nominating faculty: Professor Armando D. Solis, Biology 3350, Department of Biological Sciences, School of Arts & Sciences, New York City College of Technology, CUNY.

Cite as: Loor, L. (2013). So much "junk"? The complexities of the human genome. *City Tech Writer, 8,* 8-12. Online at https://openlab.citytech.cuny.edu/city-tech-writer-sampler/