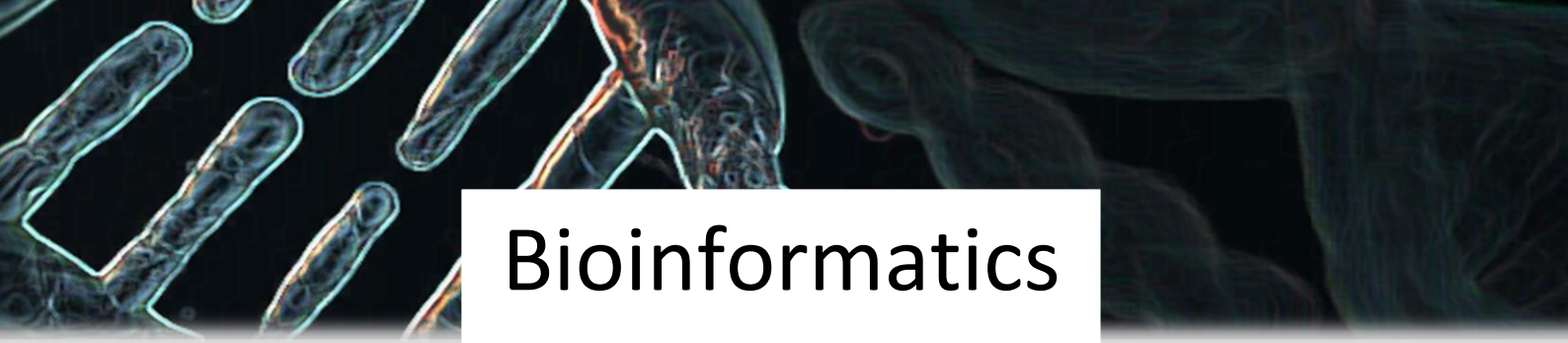


Bioinformatics

(Whole Genome & Exome Sequencing)

Outline

In every 1,000 births 10 individuals are affected by genetic disorder, and most of the affected individuals or carriers of genetic disorders have no family history of diseases and they do not aware of increasing risk of having affected newborns. Genetic disorder are caused by any changes in our genes. All living beings body are made up of millions and millions of cells and all the cells of the bodies contain genetic material called DNA which is packaged in structures called chromosomes. DNA is splited into thousands of genes and our genes are the sets of instructions that tell our bodies how to grow and function. Each and every gene is used to make a protein and proteins are what the cell use to perform their functions.

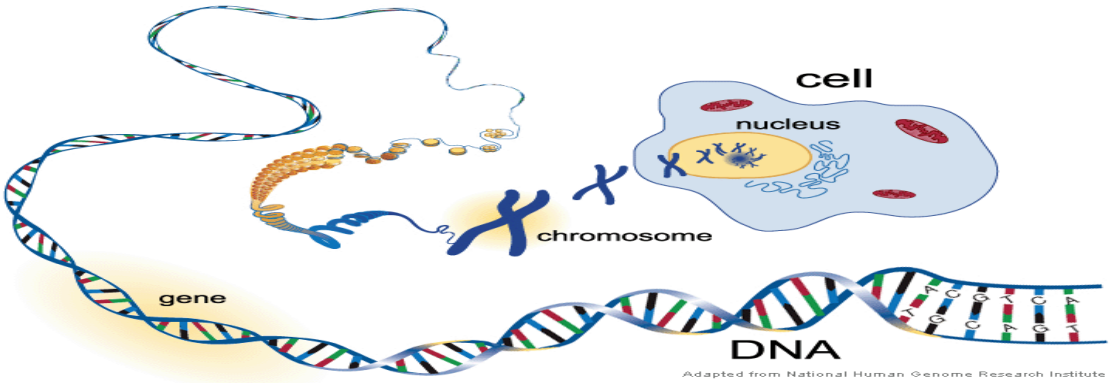
"Central dogma is the process in which the genetic information flows from DNA to RNA, to make a functional product protein."

A gene is made up of a sequence of letters in a particular order and the four letters used to make this sequence are a, t, g and c. Any change in the sequence of bases in DNA or RNA is called **mutation**. In order to understand the protein functions that are related to disease, it is important to detect the correlation between amino acid mutations and disease. Many mutation studies about disease-related proteins have been carried out through molecular biology techniques, such as vector design, protein engineering, and protein crystallization. However, experimental protein mutation studies are time-consuming, be it in vivo or in vitro.

1.1 Genomes and Exomes

As we all know our bodies are made up of cells**.** . The basic structural and functional unit of all forms of life is cell. Each and every cell contain cytoplasm which enclosed within a membrane, and contains macromolecule such as proteins,

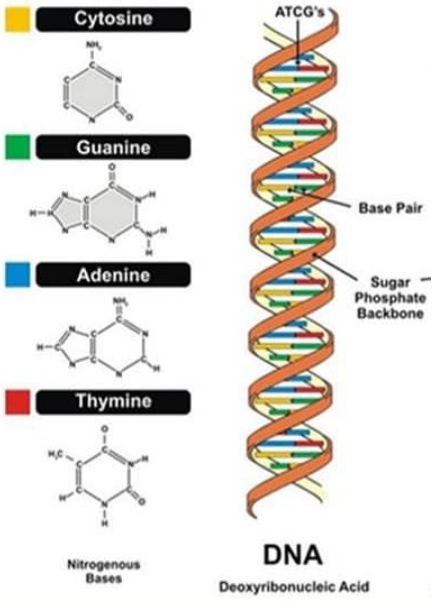
DNA and RNA, as well as many numerous molecules of nutrients and metabolites. Every single cell of the body has a copy number of DNA which carries the genetic information from one generation to next generation. To carry the genetic information from cell to cell there is a threadlike structures made of protein and a single molecule of DNA called **chromosomes.** Each chromosome contains hundreds of different genes on it and our genes what those are basically instruction manuals for the different things that our body needs to do.



https://useruploads.socratic.org/puDVmAVgSBy1dqrzo38g_cellsToDNA.gif

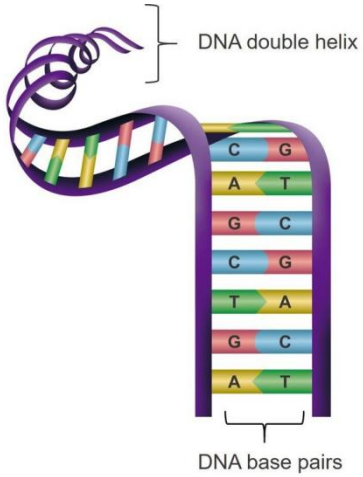
1.1.1 What is the Genome and Exome?

The genome is the 3 billion base pairs that make up our DNA.
As we all know a gene is just a small part of a genome. Its an organism entire genetic code stored in one long sequence of deoxyribonucleic acid- or DNA. The human genome is 3.2 billion letters long. Genome involves reading through the A's, T's, G's, and C's that makeup our DNA, and it has given a lot of information about what our genes are and how they are organized.



https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcS2Vuv689RTVoSDa13urqR6v5eKDFIWgDN0E3kN7Uff5GuJJiQMWliqBD8N_DDu0FheYIc&usqp=CAU

These four organic molecules called **Nucleotides,** chooses their pair and bind together as A-T, C-G. Adenine only binds with Thymine, and Cytosine only binds with Guanine.



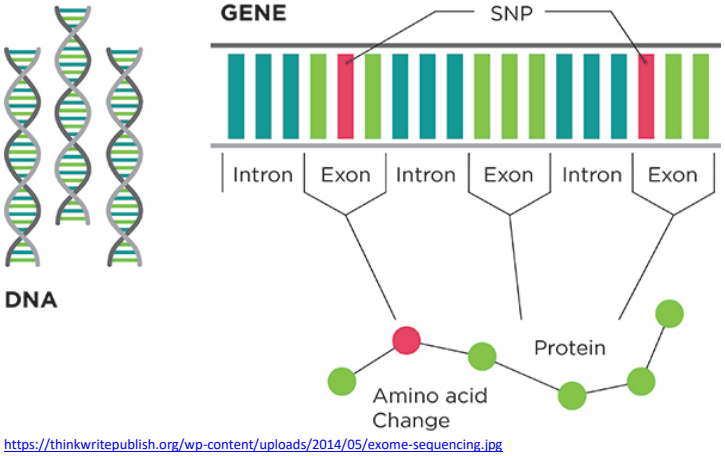
https://qph.cf2.quoracdn.net/main-qimg-0ff58dbeadeea2aa0d8279e861606bb9-lq

But not every single letter of DNA contains instructions for building life, only certain segments do- those are called Genes. The length of a gene can vary depending on the complexity of what it wants the body to build. They can be as short as few hundred letters long or as long as tens of thousands. Genes are passed from parents to children through something called a **chromosome.**

A chromosome is where the DNA is actually stored and humans have 46 of them, 23 from mother and 23 from father. Chromosomes are found in nucleus of every single one's cells.

The DNA needs a messenger called ribosome. To do this, DNA transcribes a 1/2 helix copy of itself called ribonucleic acid, or RNA which slips out of the cell's nucleus. Messenger RNA (mRNA) takes the gene construction to a ribosome, which processes the DNA by taking amino acids from the cytoplasm and turning them into proteins. Amino acids are sometimes referred to as the building blocks of life because the proteins they form go on to make cells, cells make tissues, tissues make organs and organs make living creatures or being.

The genome is the 3 billion base pairs or 3 billion letters that make up our DNA so its that entire store in DNA every single letter that makes it up that would be our **Genome** and the process for sequencing and analyzing a sample of DNA taken from beings blood is called genome sequencing. The Exome is just a portion of the genome, it's about one to three percent of the total genome and exome consists of exons that are the portions of the DNA that encode the proteins. **Exomes** are the protein coding region of the genome. Sequencing and analyzing of all the protein coding region of gene in a genome is called whole exome sequencing, exomes encode most known disease-related variants.



https://thinkwritepublish.org/wp-content/uploads/2014/05/exome-sequencing.jpg

Whole genome sequencing analyzes the entire genome including coding, non-coding and mitochondrial DNA. It helps to discover new genomics variants (structural, single nucleotide, insertion-deletion, copy number) and it helps to identify previously unknown variants for future targeted studies. Because the entire genome is being sequenced, changes in non-coding sections of DNA within genes, called introns, can also be determined. Under normal circumstances, introns are removed by RNA splicing during the post-transcriptional process, and changes in these regions can be consequential to whether DNA is transcribed into RNA or potentially into a truncated one, results in non-functional proteins.

Exomes make up only about 2% of the entire genome. Because the genome is very large, the exome can be sequenced at a very high depth (the number of times a given nucleotide is sequenced) at low cost. This greater depth provides greater confidence in the low frequency changes. Sequencing depth can be even greater for less cost by using targeted or "hot-spot" sequencing panels, which contain a select number of specific genes, or coding regions within genes that are known to harbor mutations. are known to contribute to the pathogenesis of disease, and may include clinically-functional genes of interest (eg, diagnostic, theranostic, etc.). These are often used in clinical care to provide greater confidence as well as keep costs down and a better chance for insurance reimbursement. However, whole-exome sequencing and targeted panels only see part of the story because they focus on fewer regions of the genome. Consequently, for some research projects or genetics testing, whole-genome sequencing may be beneficial.

1.2 What is Bioinformatics?

80 %, that's how similar our protein coding genes are to those of mice. Humans are larger, smarter and live longer than mice. The reason we are able to calculate our similarity to mice is because of massive effort to sequence the genomes of humans, an undertaking called the Human Genome Project. As well as from sequencing the genomes of the mice and many other organisms. Genome sequencing involves reading through the A's, T's, G's, and C's that makeup our DNA, and it has given a lot of information about what our genes are and how they are organized. And it has helped us improve how we diagnose and even treat human disease. But genome sequencing involves generating very large sets of data, so we need powerful tools to decipher all those ATGC's. This is where the rapidly growing field bioinformatics comes in.

Extremely powerful computers are being used to store and manipulate all of this data, and the people behind the computers are bioinformaticians, scientists who are often trained both in biology as well as maths or computer science. These multidisciplinary researchers develop methods and software tools to programme computers to dig through and make sense of all of this data.

1.2.1 How does Bioinformatian help us understand similarities between genes?

By looking for small sequences in one genome that match the other genome.

For example: Human: ACTGTAACGTT**ATTGCACGTCTA**CCTTCAA

Mice: ACTGTAACCTT**ATTGCACGTCTA**CCTAGGC

(Looking for small matching sequences '**ATTGCACGTCTA**')

Once matching area found, scientists design algorithms that can scan past the ends of both sequences to see just how far the matching regions extend.

So, in this case, the letters after the CTA do not match between the mouse and human sequences.

Human: ACTGTAACGTT**ATTGCACGTCTA**CCT**TCAA**

Mice: ACTGTAACCTT**ATTGCACGTCTA**CCT**AGGC**

Differences in sequences

As differences are present in sequences, scientists can analyze the following sequences and start to figure out which genes are involved in many of the traits that make us humans different from mice, like brain development and endurance.

Contrastingly, researchers can also find sequences of the genome that are highly similar across different species. Along with, sharing parts of our genome with mice, humans have genes in common with plants, flies and even microscopic bacteria. These regions are called **conserved genes**, and because their shared across many species, they likely code for proteins that are essential for life on earth.

Human: ACTGTAACGTT**ATTGCACGTCTA**CCT**TCAA**

Mice: ACTGTAACCTT**ATTGCACGTCTA**CCT**AGGC**

1.2.2 How Bioinformatian does helps understand disease?

The complete analysis of the human genome has revolutionized the way diseases are studied and treated. In addition to techniques such as next-generation sequencing, which allow scientists to study the genetic sequences included in the human genome, bioinformatics is also important in ensuring the reliability of these scientific results..

In addition to finding the similarities and difference between the genomes of different organisms, sequencing technologies has also allowed to pick up differences in DNA between different people. This has been particularly important because it helps us better understand human disease.

A specific DNA bases in a gene was different from person to person.

For example, sequence base between 50 healthy people and 50 unhealthy people. If 47 out of 50 unhealthy people have ‘A’, and only 5 out of 50 healthy people have ‘A’, that would be strong evidence of association between the ‘A’ variant and the disease. But importantly it does not mean the variant cause the disease.

Scientists or researchers have developed technologies to look across hundreds of thousands of sites across the genome for these kinds of single base differences and have looked for correlation between certain bases and disease. These experiments are known as genome-wide association studies.

Sequencing and bioinformatics analysis are also becoming increasingly important for the diagnosis and treatment of many dangerous diseases such as cancer. Cancer cells often have many mutations or changes in the nucleotide code compared to a patient’s normal cells.

By using algorithms to compare a patient’s tumor cells to the normal genome, as well as to the tumor of many other patients, doctors can quickly pinpoint the changes in the DNA that are causing the cancer cell to grow uncontrollably. This helps them to choose the best treatment for their patients.
As the ability to sequence genomes continues to increase, bioinformatics will need to continue to develop faster and more advanced algorithms to handle these massive data sets.
An important part of this field has been the development of large centralized databases of genome sequences that can be accessed by anyone. The challenge for the future will be to continue to grow these databases in a way that helps scientists make important new discoveries while preserving the privacy of patients.