**GENOMICS - GLOSSARY**

Genomics is an essential subfield of bioinformatics, and a major force in expanding human knowledge of genetic associations with disease and other traits. It is an interdisciplinary field, including the development of methods for DNA sequencing, as well as for big data analysis of genomic sequences. **Next-Generation Sequencing** (NGS) techniques allow for whole-genome sequencing, and analysis of epigenetic factors such as DNA-protein interactions and DNA methylation with unprecedented efficiency. Massive databases for biological sequence data such as**GenBank, EMBL,**and**SRA** allow for data-driven research and knowledge discovery. This course serves as an introduction to the bioinformatics sub-discipline of genomics. Students will be familiarized with the **biology of genetics**and **genetic variation, while considering practical applications.**

In this blog, we have listed all the terminologies and their meanings for your understanding, from the [**Genomics**](https://learn.omicslogic.com/courses/course/course-3-genomics) course. Clicking on the lesson heading will redirect you back to the lesson where these terms are used.

[**01 - DNA Structure & Variants**](https://learn.omicslogic.com/Learn/course-3-genomics/lesson/01-dna-structure-and-variants)

In this lesson, you will get an overview of one of important biomolecule, i.e. Deoxyribonucleic Acid (DNA). You will get familiarized with its structure, organizations and genomic variations.

* **Trait:**Trait is a specific characteristic of an individual. For example, their hair color or their blood type. Traits are determined by genes, and also they are determined by the interaction with the environment with genes. And remember that genes are the messages in our DNA that define individual characteristics.
* **Sequencing:**Sequencing technologies include a number of methods that are grouped broadly as template preparation, sequencing and imaging, and data analysis. The unique combination of specific protocols distinguishes one technology from another and determines the type of data produced from each platform.
* **Annotation:**An annotation is extra information associated with a particular point in a document or other piece of information. It can be a note that includes a comment or explanation. Annotations are sometimes presented in the margin of book pages.
* **Exons:**Exons are coding sections of an RNA transcript, or the DNA encoding it, that are translated into protein. Exons can be separated by intervening sections of DNA that do not code for proteins, known as introns. Splicing produces a mature messenger RNA molecule that is then translated into a protein.
* **Introns:** Introns are noncoding sections of an RNA transcript, or the DNA encoding it, that are spliced out before the RNA molecule is translated into a protein. The sections of DNA (or RNA) that code for proteins are called exons. ... Introns are also referred to as intervening sequences.
* **Mutations:** A mutation is a change in a DNA sequence. Mutations can result from DNA copying mistakes made during cell division, exposure to ionizing radiation, exposure to chemicals called mutagens, or infection by viruses. Germline mutations occur in the eggs and sperm and can be passed onto offspring, while somatic mutations occur in body cells and are not passed on.
* **Point Mutations:**Point mutations are a large category of mutations that describe a change in single nucleotide of DNA, such that that nucleotide is switched for another nucleotide, or that nucleotide is deleted, or a single nucleotide is inserted into the DNA that causes that DNA to be different from the normal or wild type gene
* **Codon:**A codon is a sequence of three DNA or RNA nucleotides that corresponds with a specific amino acid or stop signal during protein synthesis. ... Each codon corresponds to a single amino acid (or stop signal), and the full set of codons is called the genetic code.

[**02 - Working with DNA Sequences in R**](https://learn.omicslogic.com/R-Code/course-3-genomics/lesson/02-working-with-dna-sequences-in-r)

In this lesson, you will learn to load, view and manipulate sequences. We will start from learning about DNA and then turn to sequence alignment to find mutations that can help treat patients better.

**Sequence alignment**

In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Sequence alignments are also used for non-biological sequences, such as calculating the distance cost between strings in a natural language or in financial data.

**Pairwise Alignment**

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

**Multiple Alignment**

Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides. ... MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex.

[**03 - Introduction to Genomics**](https://learn.omicslogic.com/courses/course/course-3-genomics)

In this lesson, you will get an overview about the bioinformatics sub-discipline of genomics. Here, you will be familiarized with the biology of genetics and genetic variations.

**Introduction**

* **Genomics:**Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes. A genome is an organism's complete set of DNA, including all of its genes as well as its hierarchical, three-dimensional structural configuration
* **Bioinformatics:** Bioinformatics is defined as the application of tools of computation and analysis to the capture and interpretation of biological data. It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology
* **Genetics:** Genetics is the study of heredity in general and of genes in particular. Genetics forms one of the central pillars of biology and overlaps with many other areas, such as agriculture, medicine, and biotechnology.
* **Genome-wide association studies:**A genome-wide association study (GWAS) is an approach used in genetics research to associate specific genetic variations with particular diseases. The method involves scanning the genomes from many different people and looking for genetic markers that can be used to predict the presence of a disease.
* **Personalized medicine:** Personalized medicine is an emerging practice of medicine that uses an individual's genetic profile to guide decisions made in regard to the prevention, diagnosis, and treatment of disease.
* **Agrigenomics:** Agricultural genomics is a rich field that has been contributing to advances in crop development for decades. From sequencing reference genomes to genotyping for genome-wide association studies to genomic prediction, advances in technology and applications have led to breakthroughs in crop improvement. Genomics can reduce the trials and failures involved in scientific research to a certain extent, which could improve the quality and quantity of crop yields in agriculture. Linking traits to genes or gene signatures helps to improve crop breeding to generate hybrids with the most desirable qualities.
* **Metagenomics:** Metagenomics is the study of genetic material recovered directly from environmental samples. The broad field may also be referred to as environmental genomics, ecogenomics or community genomics.

**DNA molecules**

* **Reverse complement:**Reverse Complement. Reverse Complement converts a DNA sequence into its reverse, complement, or reverse-complement counterpart. You may want to work with the reverse-complement of a sequence if it contains an ORF on the reverse strand. Paste the raw or FASTA sequence into the text area below.
* **Transcriptase:** An enzyme which catalyses the formation of RNA from a DNA template during transcription, or ( *reverse transcriptase* ) the formation of DNA from an RNA template in reverse transcription.
* **Gene regulation:**Gene regulation is the process of turning genes on and off. During early development, cells begin to take on specific functions. Gene regulation ensures that the appropriate genes are expressed at the proper times. Gene regulation can also help an organism respond to its environment.
* **Transcription:**Transcription is the process by which the information in a strand of DNA is copied into a new molecule of messenger RNA (mRNA). ... The newly formed mRNA copies of the gene then serve as blueprints for protein synthesis during the process of translation.
* **RNA:** Ribonucleic acid (RNA) is a molecule similar to DNA. Unlike DNA, RNA is single-stranded. An RNA strand has a backbone made of alternating sugar (ribose) and phosphate groups. ... Different types of RNA exist in the cell: messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA).
* **Transcriptomics:**Transcriptomics allows identification of genes and pathways that respond to and counteract biotic and abiotic environmental stresses. The non-targeted nature of transcriptomics allows the identification of novel transcriptional networks in complex systems.
* **Epigenetics:**Epigenetics is the study of how your behaviors and environment can cause changes that affect the way your genes work. Unlike genetic changes, epigenetic changes are reversible and do not change your DNA sequence, but they can change how your body reads a DNA sequence.
* **Ploidy:** Ploidy refers to the number of sets of homologous chromosomes in the genome of a cell or an organism. ... The cell or the organism with two sets of homologous chromosomes, 2n, is described as diploid. Having multiple sets of paired chromosomes in a genome of an organism is described as polyploid.
* **Phenotypes:**The term "phenotype" refers to the observable physical properties of an organism; these include the organism's appearance, development, and behavior. An organism's phenotype is determined by its genotype, which is the set of genes the organism carries, as well as by environmental influences upon these genes.
* **Alleles:** An allele is a variant form of a gene. Some genes have a variety of different forms, which are located at the same position, or genetic locus, on a chromosome. ... Alleles contribute to the organism's phenotype, which is the outward appearance of the organism. Some alleles are dominant or recessive.
* **Epistasis:** The interaction of genes that are not alleles, in particular the suppression of the effect of one such gene by another.
* **Somatic cells:** A somatic cell is any cell of the body except sperm and egg cells. Somatic cells are diploid, meaning that they contain two sets of chromosomes, one inherited from each parent. Mutations in somatic cells can affect the individual, but they are not passed onto offspring.
* **Mitosis:** Mitosis is a process where a single cell divides into two identical daughter cells (cell division). During mitosis one cell? divides once to form two identical cells. The major purpose of mitosis is for growth and to replace worn out cells.
* **Meiosis:** Meiosis is a process where a single cell divides twice to produce four cells containing half the original amount of genetic information. These cells are our sex cells – sperm in males, eggs in females. During meiosis one cell? divides twice to form four daughter cells.

**Types of Genome Sequencing**

**Sequencing**

* **Genomic DNA:**Genomic deoxyribonucleic acid is chromosomal DNA, in contrast to extrachromosomal DNAs like plasmids. It is also then abbreviated as gDNA. Most organisms have the same genomic DNA in every cell; however, only certain genes are active in each cell to allow for cell function and differentiation within the body.
* **E**[**xome**](https://edu.t-bio.info/glossary/exome/)**sequencing:** Exome sequencing, also known as whole exome sequencing (WES), is a genomic technique for sequencing all of the protein-coding regions of genes in a genome (known as the exome). ... The second step is to sequence the exonic DNA using any high-throughput DNA sequencing technology.
* **Regulation:**Genome regulation encompasses all facets of gene expression, from the biochemical modifications of DNA, to the physical arrangement of chromosomes and the activity of the transcription machinery. The genome regulation programs that cells engage control which proteins are produced, and to what level.
* **Phenotype:**A phenotype is an individual's observable traits, such as height, eye color, and blood type. The genetic contribution to the phenotype is called the genotype. Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors.
* **Exome:** The exome is the part of the genome composed of exons, the sequences which, when transcribed, remain within the mature RNA after introns are removed by RNA splicing and contribute to the final protein product encoded by that gene.

**Genomics Variations**

* **Mutations:** A mutation is a change in a DNA sequence. Mutations can result from DNA copying mistakes made during cell division, exposure to ionizing radiation, exposure to chemicals called mutagens, or infection by viruses. Germline mutations occur in the eggs and sperm and can be passed onto offspring, while somatic mutations occur in body cells and are not passed on.
* **Point Mutations:**Point mutations are a large category of mutations that describe a change in single nucleotide of DNA, such that that nucleotide is switched for another nucleotide, or that nucleotide is deleted, or a single nucleotide is inserted into the DNA that causes that DNA to be different from the normal or wild type gene
* **SNP:** Single nucleotide polymorphisms, frequently called SNPs (pronounced “snips”), are the most common type of genetic variation among people. ... SNPs occur normally throughout a person's DNA. They occur almost once in every 1,000 nucleotides on average, which means there are roughly 4 to 5 million SNPs in a person's genome.
* **Biomarkers:**A biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. A biomarker may be used to see how well the body responds to a treatment for a disease or condition.
* **Gene regulation:** Gene regulation is the process of turning genes on and off. During early development, cells begin to take on specific functions. Gene regulation ensures that the appropriate genes are expressed at the proper times. Gene regulation can also help an organism respond to its environment.
* **Codons:** In DNA or RNA, a sequence of 3 consecutive nucleotides that codes for a specific amino acid or signals the termination of gene translation (stop or termination codon).
* **Phenotypes:**The term "phenotype" refers to the observable physical properties of an organism; these include the organism's appearance, development, and behavior. An organism's phenotype is determined by its genotype, which is the set of genes the organism carries, as well as by environmental influences upon these genes.
* [**Germli**](https://edu.t-bio.info/glossary/germline/)**n**[**e**](https://edu.t-bio.info/glossary/germline/)**:** In biology and genetics, the germline is the population of a multicellular organism's cells that pass on their genetic material to the progeny (offspring). In other words, they are the cells that form the egg, sperm and the fertilised egg.
* **Somatic cells:**A somatic cell is any cell of the body except sperm and egg cells. Somatic cells are diploid, meaning that they contain two sets of chromosomes, one inherited from each parent. Mutations in somatic cells can affect the individual, but they are not passed onto offspring.
* **Genomics:**Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes. A genome is an organism's complete set of DNA, including all of its genes as well as its hierarchical, three-dimensional structural configuration.
* **Genotype:**A genotype is an individual's collection of genes. ... The genotype is expressed when the information encoded in the genes' DNA is used to make protein and RNA molecules. The expression of the genotype contributes to the individual's observable traits, called the Phenotype

[**04 - DNA Replication and Reverse Complements in R**](https://learn.omicslogic.com/R-Code/course-3-genomics/lesson/04-dna-replication-and-reverse-complements-in-r)

In this lesson, we will learn about DNA and the DNA replication process. We will also learn about how we can compute nucleotide frequencies and create our own complementary strands of DNA on the computer using R.

**X-ray diffraction**

X-ray diffraction, a phenomenon in which the atoms of a crystal, by virtue of their uniform spacing, cause an interference pattern of the waves present in an incident beam of X rays. The atomic planes of the crystal act on the X rays in exactly the same manner as does a uniformly ruled grating on a beam of light.

**Semi-conservative duplication**

Semiconservative replication describes the mechanism of DNA replication in all known cells. ... This process is known as semi-conservative replication because two copies of the original DNA molecule are produced, each copy conserving (replicating) the information from one half of the original DNA molecule.

**Cyclin-dependent kinases**

Cyclin-dependent kinases (CDKs) are protein kinases characterized by needing a separate subunit - a cyclin - that provides domains essential for enzymatic activity. CDKs play important roles in the control of cell division and modulate transcription in response to several extra- and intracellular cues.

[**05 - Genomic Variation in NGS Data: Practical**](https://learn.omicslogic.com/Learn/course-3-genomics/lesson/05-genomic-variation-in-ngs-data-practical)

In this lesson, we will create a pipeline for variant calling using the T-BioInfo platform to process paired-end read data from each pair of matched tumor-normal cell lines, and identify variants between cancerous and noncancerous cell genomes.

* **Tumor suppressor genes:** Tumor suppressor genes are normal genes that slow down cell division, repair DNA mistakes, or tell cells when to die (a process known as apoptosis or programmed cell death). When tumor suppressor genes don't work properly, cells can grow out of control, which can lead to cancer.
* **Sequence Read Archive:**The Sequence Read Archive is a bioinformatics database that provides a public repository for DNA sequencing data, especially the "short reads" generated by high-throughput sequencing, which are typically less than 1,000 base pairs in length.
* **Array Express:** [ArrayExpress](https://www.ebi.ac.uk/arrayexpress/) is one of the major public repositories for functional genomics datasets. Most of the data is genome-wide gene expression data, measured on microarray or next-generation sequencing (NGS) platforms. A range of DNA assays are also hosted by ArrayExpress, such as ChIP-seq or genotyping.

The main object in ArrayExpress is the experiment. An experiment usually groups several assays belonging to one study or publication. Each experiment contains metadata describing the biological specimen and experimental procedures, as well as resulting data files. The definition of an assay depends on the experiment type. For microarray experiments an assay represents one hybridisation (of biological sample material to an array). For NGS experiments an assay is the read-out (sequencing) of one library.

* [**Paired-end**](https://edu.t-bio.info/glossary/paired-end/)**sequencing:** Paired-end tags are the short sequences at the 5’ and 3' ends of a DNA fragment which are unique enough that they exist together only once in a genome, therefore making the sequence of the DNA in between them available upon search or upon further sequencing.
* **S**[**ingle-end**](https://edu.t-bio.info/glossary/single-end/)**sequencing:**Single-read sequencing involves sequencing DNA from only one end, and is the simplest way to utilize Illumina sequencing. ... Single-read sequencing can be a good choice for certain methods such as small RNA-Seq or chromatin immunoprecipitation sequencing (ChIP-Seq).
* **FASTQ:**FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity. <https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/>
* **Confounding:** In statistics, a confounder is a variable that influences both the dependent variable and independent variable, causing a spurious association. Confounding is a causal concept, and as such, cannot be described in terms of correlations or associations.
* **GTF:**The Gene transfer format (GTF) is a [file format](https://en.wikipedia.org/wiki/File_format) used to hold information about [gene](https://en.wikipedia.org/wiki/Gene) structure. It is a [tab-delimited](https://en.wikipedia.org/wiki/Tab_key) text format based on the [general feature format](https://en.wikipedia.org/wiki/General_feature_format) (GFF), but contains some additional conventions specific to gene information. A significant feature of the GTF that can be validated: given a sequence and a GTF file, one can check that the format is correct. This significantly reduces problems with the interchange of data between groups.
* **Bowtie2:** Bowtie2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. ... Bowtie supports reads longer than 50bp and is generally faster, more sensitive, and uses less memory than Bowtie.
* **Strelka:**Strelka is an analysis package designed to detect somatic SNVs and small indels from the aligned sequencing reads of matched tumor-normal samples.
* **Single nucleotide variants:** A DNA sequence variation that occurs when a single nucleotide (adenine, thymine, cytosine, or guanine) in the genome sequence is altered. Single nucleotide variants may be rare or common in a population.
* [**Phred**](http://www.phrap.com/phred/)**quality score:**Base calling accuracy, measured by the Phred quality score (Q score), is the most common metric used to assess the accuracy of a sequencing platform. It indicates the probability that a given base is called incorrectly by the sequencer.
* **VCF:**The **Variant Call Format (VCF)** specifies the format of a text file used in bioinformatics for storing gene sequence variations. The format has been developed with the advent of large-scale genotyping and DNA sequencing projects, such as the 1000 Genomes Project. Existing formats for genetic data such as General feature format (GFF) stored all of the genetic data, much of which is redundant because it will be shared across the genomes. By using the variant call format only the variations need to be stored along with a reference genome.
* **SNV:** **Single nucleotide variants (SNVs)** occur when a single nucleotide (e.g., A, T, C, or G) is altered in the DNA sequence. SNVs are by far the most common type of sequence change, and there are a number of endogenous and exogenous sources of damage that lead to the single base pair substitution mutations that create SNVs.
* **Integrative Genomics Viewer:**Integrative Genomics Viewer (IGV) is a desktop application for the visualization and interactive exploration of genomic data in the context of a reference genome. A key characteristic of IGV is its focus on the integrative nature of genomic studies. Link to the software: <https://software.broadinstitute.org/software/igv/>
* **Tumor suppressor gene:** Tumor suppressor genes are normal genes that slow down cell division, repair DNA mistakes, or tell cells when to die (a process known as apoptosis or programmed cell death). When tumor suppressor genes don't work properly, cells can grow out of control, which can lead to cancer.
* **Allele:**An allele is a variant form of a gene. Some genes have a variety of different forms, which are located at the same position, or genetic locus, on a chromosome. ... Alleles contribute to the organism's phenotype, which is the outward appearance of the organism. Some alleles are dominant or recessive.

[**06 - Phylogenetic Analysis**](https://learn.omicslogic.com/Learn/course-3-genomics/lesson/06-phylogenetic-analysis)

In this lesson, we will get an overview about the phylogenetic analysis. Here, we will also learn how we can perform phylogenetic analysis based on NGS data using T-bioinfo server and derive biological significance.

**Taxonomy:** A taxonomy (or taxonomic classification) is a scheme of classification, especially a hierarchical classification, in which things are organized into groups or types. Among other things, a taxonomy can be used to organize and index knowledge (stored as documents, articles, videos, etc.), such as in the form of a library classification system, or a search engine taxonomy, so that users can more easily find the information they are searching for. Many taxonomies are hierarchies (and thus, have an intrinsic tree structure), but not all are.

Originally, taxonomy referred only to the categorisation of organisms or a particular categorisation of organisms. In a wider, more general sense, it may refer to a categorisation of things or concepts, as well as to the principles underlying such a categorisation. Taxonomy organizes taxonomic units known as "taxa" (singular "taxon")."

**OTHER IMPORTANT TERMS**

* **Cell division** :- Cell division is the process by which a parent cell divides into two or more daughter cells.
* **Genotype** :- A genotype is an organism's complete set of genetic material. Often though, genotype is used to refer to a single gene or set of genes, such as the genotype for eye color.
* **Phenotype** :- In genetics, the phenotype is the set of observable characteristics or traits of an organism.
* **Protein coding** :- Protein coding sequences are DNA sequences that are transcribed into mRNA and in which the corresponding mRNA molecules are translated into a polypeptide chain. Every three nucleotides, termed a codon, in a protein coding sequence encodes 1 amino acid in the polypeptide chain.
* **KRTAP12-4** :- KRTAP12-4 (Keratin Associated Protein 12-4) is a Protein Coding gene. Diseases associated with KRTAP12-4 include Deafness, Autosomal Recessive 98 and Autosomal Recessive Non-Syndromic Sensorineural Deafness Type Dfnb.
* **CNTNAP2 gene** :- This gene encodes a member of the neurexin family which functions in the vertebrate nervous system as cell adhesion molecules and receptors. This gene has been implicated in multiple neurodevelopmental disorders, including Gilles de la Tourette syndrome, schizophrenia, epilepsy, autism, ADHD and intellectual disability.
* **Regulatory regions** :- A regulatory sequence is a segment of a nucleic acid molecule which is capable of increasing or decreasing the expression of specific genes within an organism. Regulation of gene expression is an essential feature of all living organisms and viruses.
* **Multiple sequence alignment :-**Multiple sequence alignment (MSA) refers to the process or the result of sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor.
* **Variant calling :-**Variant calling is the process by which we identify variants (variations caused by mutations/changes) from high throughput sequence data.
* **Gene expression :-**Gene expression is the process by which the instructions in our DNA are converted into a functional product, such as a protein.
* **1,000 genome project :-**The 1000 Genomes Project (abbreviated as 1KGP), launched in January 2008, was an international research effort to establish by far the most detailed catalogue of human genetic variation. Scientists planned to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups within the following three years, using newly developed technologies which were faster and less expensive.
* **Omics technologies :-**'Omics' technologies are primarily aimed at the universal detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biological sample. Omics technologies have a broad range of applications.
* **Epigenomic regulation :-**Epigenomics focuses on the analysis of global epigenetic changes that provides important clues regarding mechanisms and function of gene regulation across many genes in a cell or organism.
* **Isoform expression :-**The idea that different binding partners confer different functional properties has been well studied in tissue-specific gene regulation. For example, the same transcription factor (TF) can direct gene expression in different tissues simply by binding with different TSSs in each tissue.
* **Non Coding :-** Some noncoding DNA regions, called introns, are located within protein-coding genes but are removed before a protein is made. Regulatory elements, such as enhancers, can be located in introns. Other noncoding regions are found between genes and are known as intergenic regions.
* **Non-Synonymous Mutation :-**Nonsynonymous mutations change the protein sequences and are frequently subjected to natural selection. The same goes for nonsense mutations that introduce premature stop codons into CDSs (coding sequences).
* **FGFS gene :-**Hitherto, the only known mutant gene leading to the long-hair phenotype in mammals is the fibroblast growth factor 5 (FGF5).
* **Genotype-Phenotype Associations :-**Genetic association can be between phenotypes, such as visible characteristics such as flower color or height, between a phenotype and a genetic polymorphism, such as a single nucleotide polymorphism (SNP), or between two genetic polymorphisms.
* **Oncogenes :-**When a proto-oncogene mutates (changes) or there are too many copies of it, it becomes a "bad" gene that can become permanently turned on or activated when it is not supposed to be. When this happens, the cell grows out of control, which can lead to cancer. This bad gene is called an oncogene.
* **Somatic :-**The majority of current somatic variant callers are designed to analyze matched tumor-normal samples from the same patient. The fundamental idea is to identify potential variants using the tumor and distinguish somatic variants from germline and loss of heterozygosity (LOH) variants using the matched normal sample.
* **Next Generation Sequencing :-**Next generation sequencing (NGS), massively parallel or deep sequencing are related terms that describe a DNA sequencing technology which has revolutionized genomic research. Using NGS an entire human genome can be sequenced within a single day.
* **Short reads :-**Short reads could be considered when you have sequences < 200-400 base pair.
* **Long reads :-**Long reads could be considered when you have sequences > 400 base pairs.
* **Pacific Biosciences :-** Pacific Biosciences of California, Inc. is an American biotechnology company founded in 2004 that develops and manufactures systems for gene sequencing and some novel real time biological observation.
* **Pre-processing** :- Raw sequence data is pre processed to produce analysis-ready files. This involves alignment to a reference genome as well as some data cleanup operations to correct for technical biases and make the data suitable for analysis.
* **Mapping :**- Mapping involves dividing the chromosomes into smaller fragments that can be propagated and characterized and ordering (mapping) them to correspond to their respective locations on the chromosomes. After mapping is completed, the next step is to determine the base sequence of each of the ordered DNA fragments.
* **Variant calling :-**Variant calling is the process by which we identify variants from sequence data.
* **Annotation :-**DNA annotation or genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining the function of those genes.
* **Genomic Visualization :-**A genomic visualization contains one or multiple coordinate systems applying a specific layout, partition, abstraction and arrangement (in case of multiple axes) of sequence coordinates. ... A visualization can consist of one or multiple views, each containing a set of aligned tracks.
* **Genome-wide association study :-**Genome-wide association study (GWAS) aims at identifying genetic variants (genotype) that are associated with specific traits (phenotype). GWA studies investigate genetic markers that cross the whole genome of a large number of individuals and predict genotype-phenotype associations by statistical analysis at population level.
* **Phylogenetic analysis :-** Phylogenetic analysis is important for gathering information on biological diversity, genetic classifications, as well as learning developmental events that occur during evolution. In phylogenetic analysis, branching diagrams are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor.
* **Apoptosis :-** Apoptosis is a form of programmed cell death that occurs in multicellular organisms. Biochemical events lead to characteristic cell changes (morphology) and death.