Biology & Immunology- main topics

September 18, 2019

1 Biology

- There are 20 types of **amino acids**. They are divided to five groups: uncharged polar, negative polar, positive polar, hydrophobics and special (Glycine which is incredibly fexible, Proline which is the opposite and Cysteine which has the potential to form disulfide bonds).
- **peptides** are short chains of amino acid monomers linked by peptide bonds (an amide type of covalent chemical bond linking two consecutive alpha-amino acids from C1 (carbon number one) of one alpha-amino "acid" and N2 (nitrogen number two) of another along a peptide or protein chain)
- proteins can be hundreds of amino acids long.
- **hydrophyllic**- water loving
- **hydrophobic**- water fearing
- **Adenosine triphosphate (ATP)** is a complex organic chemical that provides energy to drive many processes in living cells.
- **Glycolysis** is the metabolic pathway that converts glucose into pyruvate. The free energy released in this process is used to form the high-energy molecules ATP and NADH. Glycolysis is a sequence of ten enzyme-catalyzed. reactions.
- Pathways break down sugar to make energy stored in ATP and also use to do the biosynthesis of molecules such as amino acids which are then going to go on to make proteins.

- Three nucleotides encode an **amino acid**. Proteins are built from a basic set of 20 amino acids, but there are only four bases. Simple calculations show that a minimum of three bases is required to encode at least 20 amino acids. Genetic experiments showed that an amino acid is in fact encoded by a group of three bases, or **codon**.
- **gene**- discrete factors of inheritance.
- **alleles**- the alternative types of genes (A,a). Each of the organism has two alleles for each gene.
- **Genotype**- Two alleles carries by the organism at a particular gene. The sequence of your genes which determine how your cells function/body function and whether or not you have certain traits or diseases. the part of the genetic makeup of a cell, and therefore of any individual, which determines one of its characteristics (phenotype).
- **Homozygote**- Two of the same alleles.
- **Heterozygote**- Two different alleles.
- **Phenotype**- Appearance, Trait.
- Phenotype number 1 is **dominant** over phenotype 2 if the F1 hybrid between the two has phenotype number 1.
- **diploid** organisms have two copies of every chromosome, so they contain two alleles of every gene.
- **homolog**- the two copies of a chromosomes.
- genes are stretches of DNA, and an allele is a version of a gene. Each allele differs from the others by a small or large change in the DNA sequence.
- An **autosome** is a chromosome thats not a sex chromosome.
- human have one pair of sex chromosome and 22 pairs of autosomes.
- The **promoter** of the gene- where the transcription from DNA to RNA starts.
- **Transcriptional terminator**- where the transcription terminates.
- mRNA has thousands of bases.

- **introns**- the bits that are thrown out from the immature RNA to the mature RNA.
- **exons**- the bits that stay in.
- **gene expression**- the process by which information from a gene is used in the synthesis of proteins. Several steps in the gene expression process may be modulated, including the transcription, RNA splicing(During splicing, introns are removed and exons are joined together.), translation, and post-translational modification of a protein. Protein production starts at transcription (DNA to RNA) and continues with translation (RNA to protein). Thus, control of these processes plays a critical role in determining what proteins are present in a cell and in what amounts. The amounts and types of mRNA molecules in a cell reflect the function of that cell. In fact, thousands of transcripts are produced every second in every cell.

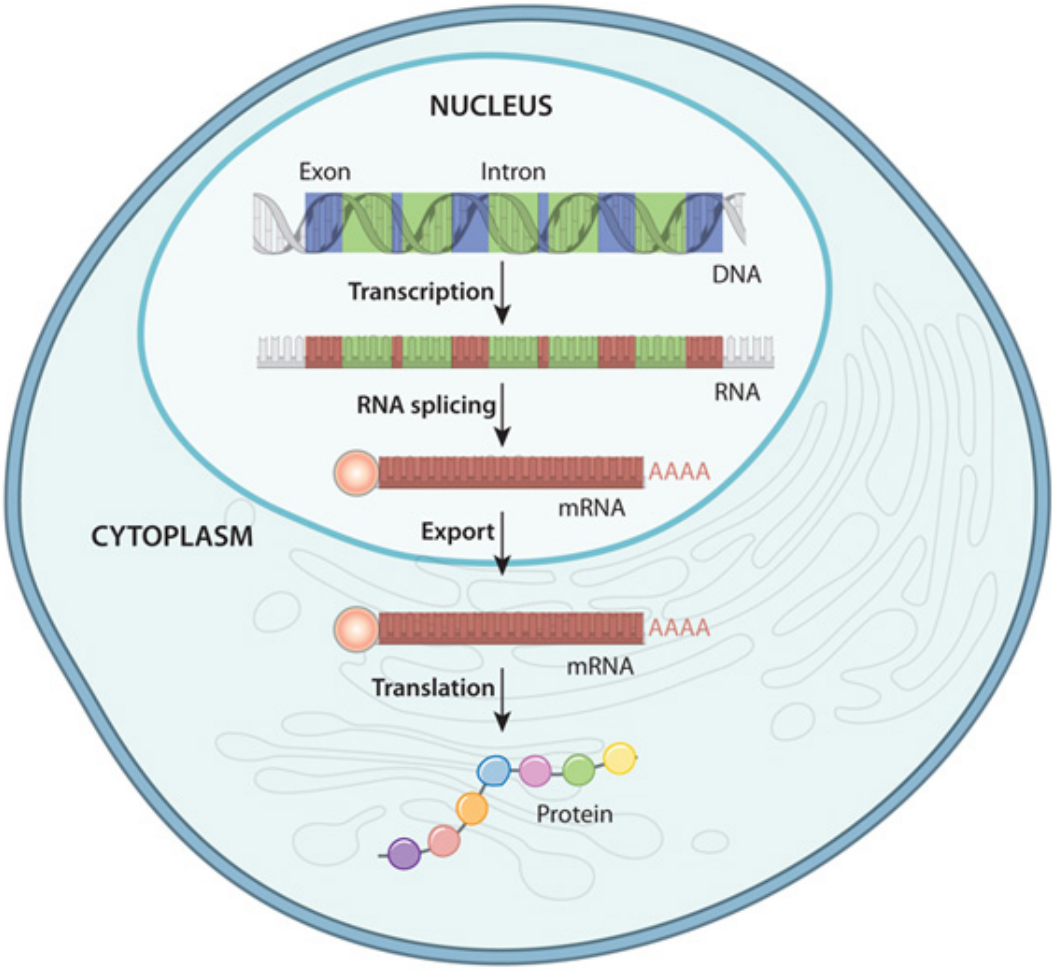

Figure 1: An overview of the flow of information from DNA to protein in a eukaryote

- **PCR (polymerase chain reaction)**- a method to make many copies of a specific DNA segment.
- **DNA polymerase**- an enzyme that synthesizes DNA molecules from deoxyribonucleotides, the building blocks of DNA. These enzymes are essential for DNA replication and usually work in pairs to create two identical DNA strands from a single original DNA molecule. During this process, DNA polymerase "reads" the existing DNA strands to create two new strands that match the existing ones. DNA is copied by a molecule called DNA Polymerase which takes free nucleotides(As, Cs, Gs and Ts) that are just floating around in your cell or you can

synthesize them so that they can be floating around in your test tube. And it uses them to copy the DNA that you're trying to sequence.

- **RNA polymerase**- an enzyme that synthesizes RNA from a DNA template.
- a very important feature of the RNA is that after transicription, the cell attaches a long string of A's to it.
- **RNA-seq**- this involves trying to capture all the genes that are being turned on in a cell, or in a collection of cells (capture mature RNA by poly(A) tail) and then doing a reverse transcriptase that gives us the DNA that matches the RNA that we have just cpatured.
- **ribosome**- The sequence of DNA, which encodes the sequence of the amino acids in a protein, is copied into a messenger RNA chain (by codons). It may be copied many times into RNA chains. Ribosomes can bind to a messenger RNA chain and use its sequence for determining the correct sequence of amino acids for generating a given protein. Amino acids are selected, collected, and carried to the ribosome by transfer RNA (tRNA) molecules, which enter one part of the ribosome and bind to the messenger RNA chain.
- A **locus** (plural loci) in genetics is a fixed position on a chromosome.
- **SNP-A single-nucleotide polymorphism**- a substitution of a single nucleotide that occurs at a specific position in the genome, where each variation is present to some appreciable degree within a population. For example, at a specific base position in the human genome, the C nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by an A. This means that there is a SNP at this specific position, and the two possible nucleotide variations C or A are said to be alleles for this position (it's the population variance).
- **Indel** is a molecular biology term for an insertion or deletion of bases in the genome of an organism.

2 Immunology

- **adaptive immune response**- a specific immune response, such as the production of antibodies against a particular pathogen or its products.

it is developed during the lifetime of an individual as an adaptation to infection with that pathogen. the adaptive immune system is capable of generating immunological memory, so that having been exposed once to an infectious agent, a person will make an immediate and stronger response against any subsequent exposure to it.

- **Antigens** - structures specifically bound by antibodies (BCR, TCR). They could stimulate antibody generation. An antigen is any molecule or part of a molecule that is specifically recognized by the highly specialized recognition proteins of lymphocytes

- **epitope**- the part of an antigen that is recognized by the immune system, specifically by antibodies, B cells, or T cells. For example, the epitope is the specific piece of the antigen to which an antibody binds. An individual antigen receptor or antibody recognizes a small part of the molecular structure of an antigenic molecule, which is known as an anti genic determinant or epitope.

- **antibody (Ab)**- Y-shaped molecules whose arms form two identical antigen-binding sites. These are highly variable from one molecule to another, providing the diversity required for specific antigen recognition. The stem of the Y is far less variable. There are only five major forms of this constant region of an antibody, and these are known as the antibody classes or isotypes. The antibody recognizes a unique molecule of the pathogen, called an antigen. Each antibody consists of two heavy chains and two light chains. The two heavy chains are linked to each other by disulfide bonds, and each heavy chain is linked to a light chain by a disulfide bond.

- **isotype**- The class, and thus the effector function, of an antibody is defined by the structure of its heavy chain. There are five main heavy-chain classes call isotypes.

- There are two major types of lymphocytes: B lymphocytes, which mature in the bone marrow and are the source of circulating antibodies, and T lymphocytes, which mature in the thymus and recognize peptides from pathogens presented by MHC molecules on infected cells or antigen-presenting cells. Each lymphocyte carries cell-surface receptors of a single antigen specificity. These receptors are generated by the random recombination of variable receptor gene segments and the pairing of distinct variable protein chains: heavy and light chains in immunoglobulins, or the two chains of T-cell receptors. The large

antigen receptor repertoire of lymphocytes can recognize virtually any antigen.

- **CD4, CD8**- lymphocytes are composed of two main classes, one of which carries the cell-surface protein called CD8 on its surface and the other bears a protein called CD4. These are not just random markers, but are important for a T cell's function, because they help to determine the interactions between the T cell and other cells. Cytotoxic T cells carry CD8, while the helper T cells involved in activating, rather than killing, the cells that they recognize carry CD4. T-cells fall into two major classes, which have different effector functions and are distinguished by the expression of the cell-surface proteins CD4 and CD8. CD8 is carried by cytotoxic T cells, while CD4 is carried by T cells whose function is to activate other cell.

  CD8 recognizes MHC class I molecules and CD4 recognizes MHC class II. During antigen recognition, CD4 or CD8 (depending on the type of T cell) associates on the T-cell surface with the T-cell receptor and binds to invariant sites on the MHC portion of the composite peptide:MHC ligand, away from the peptide-binding site. This binding is required for the T cell to make an effective response, and so CD4 and CD8 are called co-receptors.

- **cytotoxic T cell (CD8+)**- a T lymphocyte that kills cancer cells, cells that are infected (particularly with viruses), or cells that are damaged in other ways.

- **Macrophages**- a type of white blood cell, of the immune system, that engulfs and digests cellular debris, foreign substances, microbes, cancer cells, and anything else that does not have the type of proteins specific to healthy body cells on its surface

- **immunoglobulins**- Antibody molecules as a class are known as immunoglobulins (Ig). Five different classes of immunoglobulins-IgM, IgD, IgG, IgA, and IgE. Their heavy chains are denoted by the corresponding lower-case Greek letter (μ, δ, γ, α and ϵ, respectively). IgG is by far the most abundant immunoglobulin and has several subclasses (lgG1, 2, 3, and 4 in humans).

- Immunoglobulins can be made as both a transmembrane receptor and a secreted antibody, and the C domains of antibodies are crucial to their diverse effector functions.

- Each antibody molecule has a two-fold axis of symmetry and is composed to such self antigens. When a foreign of two identical heavy chains and two identical light chains. Heavy and light chains each have variable and constant regions. the variable regions of a heavy chain and a light chain combine to form an antigen-binding site, so that both chains contribute to the antigen-binding specificity of the antibody molecule.

- **TCR**- a molecule found on the surface of T cells that is responsible for recognizing fragments of antigen as peptides bound to MHC molecules. The binding between TCR and antigen peptides is of relatively low affinity and is degenerate: that is, many TCRs recognize the same antigen peptide and many antigen peptides are recognized by the same TCR. The T-cell receptor is composed of two chains of roughly equal size, called the T-cell receptor α and β chains, each of which spans the T-cell membrane.

- BCR and TCR are both highly variable antigen receptors diversified by somatic V(D)J recombination. Both T cells and B cells are cellular components of adaptive immunity.

- **BCR**- Membrane-bound immunoglobulin on the B-cell surface serves as the cell's receptor for antigen.

  The main difference between B cell receptor and antibody is that the B cell receptor is a transmembrane receptor of the B cells whereas the antibody is a protein molecule that the B cells produce. Furthermore, B cell receptor has a specific antigen binding site, which can bind to an antigen while B cells produce antibodies specifically for the neutralization of a particular pathogen.

- The antigen receptors of B cells and T cells recognize antigen in fundamentally different ways. B cells directly recognize the native antigen that either has been secreted by a pathogen or is expressed on its surface. B cells eventually differentiate into effector plasma cells that secrete antibodies that will bind to and neutralize these antigens and pathogens. In contrast,T-cell receptors do not directly recognize native antigens. Rather, they recognize antigens that have been processed, partly degraded, and displayed as peptides bound to proteins on the surface of antigen-presenting cells.

- **memory cells**- a significant number of activated antigen-specific B cells and T cells persist after antigen has been eliminated. These cells

are known as memory cells and form the basis of immunological memory. They can be reactivated much more quickly than naive lymphocytes, which ensures a more rapid and effective response on a second encounter with a pathogen and thereby usually provides lasting protective immunity.

- **MHC** - a set of genes that code for cell surface proteins essential for the acquired immune system to recognize foreign molecules in vertebrates. The main function of MHC molecules is to bind to antigens derived from pathogens and display them on the cell surface for recognition by the appropriate T-cells.

- **MHC class I**- there are two main types of MHC molecules, called MHC class I and MHC class II. CD8 T cells selectively recognize peptides that are bound to MHC class I molecules, and CD4 T cells recognize peptides presented by MHC class II. In most cells, MHC class I molecules collect peptides derived from proteins synthesized in the cytosol and are thus able to display fragments of viral proteins on the cell surface. Because MHC class I molecules are expressed on most cells of the body, they serve as an important mechanisms to defend against viral infections. MHC class I molecules bearing viral peptides are recognized by CD8-bearing cytotoxic T cells, which then kill the infected cell

- **MHC class II**- MHC class II molecules are expressed predominantly by antigen-presenting cells (dendritic cells, macrophages, and B cells) and bind peptides derived largely from proteins. These encompass proteins taken up by phagocytosis, proteins derived from pathogens living within macrophage vesicles, or proteins internalized by B cells by endocytosis. This group of cells must either activate, or be activated by, CD4 T cells.

- **HLA**- The human leukocyte antigen (HLA) system or complex is a gene complex encoding the MHC proteins in humans. These cell-surface proteins are responsible for the regulation of the immune system in humans. The HLA gene complex resides on a 3 Mbp stretch within chromosome 6p21. HLA genes are highly polymorphic, which means that they have many different alleles, allowing them to fine-tune the adaptive immune system. The proteins encoded by certain genes are also known as antigens, as a result of their historic discovery as factors in organ transplants. Different classes have different functions:

HLAs corresponding to MHC class I (A, B, and C) which all are the HLA Class1 group present peptides from inside the cell. For example, if the cell is infected by a virus, the HLA system brings fragments of the virus to the surface of the cell so that the cell can be destroyed by the immune system. These peptides are produced from digested proteins that are broken down in the proteasomes. In general, these particular peptides are small polymers, about 9 amino acids in length. Foreign antigens presented by MHC class I attract killer T-cells (also called CD8 positive- or cytotoxic T-cells) that destroy cells. MHC class I proteins associate with 2-microglobulin, which unlike the HLA proteins is encoded by a gene on chromosome 15.

HLAs corresponding to MHC class II (DP, DM, DO, DQ, and DR) present antigens from outside of the cell to T-lymphocytes. These particular antigens stimulate the multiplication of T-helper cells (also called CD4 positive T cells), which in turn stimulate antibody-producing B-cells to produce antibodies to that specific antigen. Self-antigens are suppressed by regulatory T cells.

- for example: HLA-A is a group of HLA that are coded for by the HLA-A locus, which is located at human chromosome 6p21.3. HLA is a MHC antigen specific to humans. HLA-A is one of three major types of human MHC class I cell surface receptors. The others are HLA-B and HLA-C.

- **affinity**- The strength of the interaction between a single antigen-binding site and its antigen.

- **somatic hypermutation**- introduces point mutations into the V regions of rearranged immunoglobulin genes in activated B cells, producing some variants that bind more strongly to the antigen. this leads to the phenomenon of:

- **affinity maturation**- the affinity of antibodies for the antigen increases as the immune response progresses. the process by which Tfh cell-activated B cells produce antibodies with increased affinity for antigen during the course of an immune response. With repeated exposures to the same antigen, a host will produce antibodies of successively greater affinities.

- each person possesses billions of lymphocytes, these cells collectively enable a response to a great variety of antigens. The wide range of

antigen specificities in the antigen-receptor repertoire is due to variation in the amino acid sequence at the antigen-binding site, which is made up from the variable (V) regions of the receptor protein chains. In each chain the V region is linked to an invariant constant (C) region, which provides effector or signaling functions.

- **class switching**- a process which enables antibodies with the same antigen specificity but different functional properties to be produced

- **antibody/immunoglobulin repertoire**- the total number of antibody specificities available to an individual. in humansit is at least 10^{11} and probably several orders of magnitude greater.

- **RSS**- recombination signal sequences (RSSs). DNA rearrangements are guided by conserved noncoding DNA sequences that are found adjacent to the points at which recombination takes place. A RSS consists of a conserved block of seven nucleotides-the heptamer 5'CACAGTG3' -which is always contiguous with the coding sequence, followed by a nonconserved region known as the spacer, which is either 12 or 23 base pairs (bp) long, followed by a second conserved block of nine nucleotides-the nonamer 5'ACAAAAACC3' (this sequences are the consensus and can vary slightly from individual to individual).

- **Cloning**- the process of producing genetically identical individuals of an organism either naturally or artificially.

- **Clone**-The process of immunological B-cell maturation involves transformation from an undifferentiated B cell to one that secretes antibodies with particular specificity.

- **Autoimmunity**- the system of immune responses of an organism against its own healthy cells and tissues.

- **V(D)J recombination**- a unique mechanism of genetic recombination that occurs only in developing lymphocytes during the early stages of T and B cell maturation. It involves somatic recombination, and results in the highly diverse repertoire of antibodies/immunoglobulins (Igs) and T cell receptors (TCRs) found on B cells and T cells, respectively. The process is a defining feature of the adaptive immune system.

- In the vertebrate immune system, each antibody is customized to attack one particular antigen (foreign proteins and carbohydrates) without attacking the body itself. The human genome has at most 30,000

genes, and yet it generates millions of different antibodies, which allows it to be able to respond to invasion from millions of different antigens. The immune system generates this diversity of antibodies by shuffling, cutting and recombining a few hundred genes (the VDJ genes) to create millions of permutations, in a process called V(D)J recombination. RAG-1 and RAG-2 are proteins at the ends of VDJ genes that separate, shuffle, and rejoin the VDJ genes. This shuffling takes place inside B cells and T cells during their maturation.

- **the 12/23 rule**- a gene segment flanked by a RSS with a 12 bp spacer can be joined only to the gene segment flanked by a RSS with 23 bp spacer. (joining of gene segments almost always involves a 12-bp (base pairs) and a 23-bp RSS)

- in the K gene the RSS: v=12 bs, j=23 bs. just 23 and 12 can combine together.

- **Germline configuration**- is the term used to refer to immunoglobulin and T-cell receptor genes before any rearrangement.

- a complex of **RAG-1** and **RAG-2** proteins, together with high-mobility group chromatin proteins, recognize and align the two RSSs that are the target of the cleavage reaction. RAG-1 is thought to specifically recognize the nonamer of the RSS.
  the RAG proteins generate DNA hairpins at the coding ends of the V, D, or J segments, after which Artemis catalyzes a single-stranded cleavage at a random point within the coding sequence but near the original point at which the hairpin was first formed.

- Because both the heavy and the light-chain V regions contribute to antibody specificity, each of the 320 different light chains could be combined with each of the approximately 6000 heavy chains to give around 1.9×10^6 different antibody specificities.

- The most variable parts of the T-cell receptor interact with the peptide bound to an MHC molecule.

- **FC region** (fragment crystallizable region)- the tail region of an antibody that interacts with cell surface receptors called Fc receptors and some proteins of the complement system. This property allows antibodies to activate the immune system.

- **Fc receptor**- a protein found on the surface of certain cells including, among others, B lymphocytes, natural killer cells, macrophages,...

that contribute to the protective functions of the immune system. Its name is derived from its binding specificity for a part of an antibody known as the Fc region. Fc receptors bind to antibodies that are attached to infected cells or invading pathogens. Their activity stimulates phagocytic or cytotoxic cells to destroy microbes, or infected cells by antibody-mediated phagocytosis or antibody-dependent cell-mediated cytotoxicity.

- The **Fc portion** can deliver antibodies to places they would not reach without active transport.

- The primary antibody repertoire is diversified by three processes that modify the rearranged immunoglobulin gene: **(1) Somatic hypermutation** affects the V region and diversifies the antibody repertoire by introducing point mutations into the V regions of both chains,which alters the affinity of the antibody for antigen. **(2) Class switch recombination** involves the C region only: it replaces the original Cμ heavy-chain C region with an alternative C region, thereby increasing the functional diversity of the immunoglobulin repertoire. **(3) Gene conversion** diversifies the primary antibody repertoire in some animals, replacing blocks of sequence in the V regions with sequences derived from the V regions of pseudogenes.
  All these processes result in irreversible somatic mutation of the immunoglobulin genes, but unlike V(D)J recombination they are initiated by an enzyme called activation-induced cytidine deaminase (AID), which is expressed specifically in activated B cells.

- **AID**- (activation-induced (cytidine) deaminase) an enzyme which in humans is encoded by the AICDA gene.It creates mutations in DNA-it changes a C:G base pair into a U:G mismatch. The protein is involved in somatic hypermutation, gene conversion, and class-switch recombination of immunoglobulin genes in B cells of the immune system.

- **Somatic hypermutation** (SHM), in which the antibody genes are minimally mutated to generate a library of antibody variants, some of which with higher affinity for a particular antigen than any of its close variants. Somatic hypermutation of rearranged V regions does not occur in T cells. This means that variability in the CDRl and CDR2 regions is limited to that of the germline V gene segments, and that most diversity is focused on the CDR3 regions.

- **Class switch recombination** (CSR), in which B cells change their expression from IgM to IgG or other immune types.

- **Gene conversion** (GC) a process that causes mutations in antibody genes of chickens, pigs and some other vertebrates.
- **affinity maturation**- favorable mutations make changes that increase the affinity of the B-cell receptor for its antigen, and B-cell clones producing receptors with the highest affinity for antigen are favored for survival. Some of the mutant immunoglobulins bind antigen better than the original B-cell receptors, and B cells expressing them are preferentially selected to mature into antibody-secreting cells. This gives rise to a phenomenon called affinity maturation of the antibody population.
- B cells whose V regions have accumulated deleterious mutations and can no longer bind antigen die. B cells whose V regions have acquired mutations that improve the affinity of the B-cell receptor for antigen are able to compete more effectively for antigen, and receive signals that drive their proliferation and expansion. The antibodies they produce also have this improved affinity.
- In contrast to B cells, all the diversity in T-cell receptors is generated during gene rearrangement, and somatic hypermutation of rearranged V regions does not occur in T cells.
- The λ light-chain locus, located on human chromosome 22. The k light-chain locus, on chromosome 2. The heavy-chain locus, on chromosome 14.
- The TCRα locus, located on chromosome 14 and TCRβ locus, located on chromosome 7.
- CDRs (Complementarity-determining regions)- are part of the variable chains in immunoglobulins and T cell receptors, generated by B-cells and T-cells respectively, where these molecules bind to their specific antigen.
- **Germline configuration** is the term used to refer to immunoglobulin and T-cell receptor genes before any rearrangement. This is because prior to rearrangement, the configuration is the sae as in germ cells, eggs and sperm.
- **Germinal centers** (GCs)- are sites within secondary lymphoid organs lymph nodes and the spleen where mature B cells proliferate, differentiate, and mutate their antibody genes (through somatic hypermutation aimed at achieving higher affinity), and switch the class of

their antibodies (for example from IgM to IgG) during a normal immune response to an infection. These develop dynamically after the activation of follicular B cells by T-dependent antigen.

- **TdT**- a specialized DNA polymerase expressed in immature, pre-B, pre-T lymphoid cells, and acute lymphoblastic leukemia/lymphoma cells. TdT adds N-nucleotides to the V, D, and J exons of the TCR and BCR genes during antibody gene recombination, enabling the phenomenon of junctional diversity.

- An important difference between immunoglobulins and T-cell receptors is that immunoglobulins exist in both membrane-bound forms (B-cell receptors) and secreted forms (antibodies).

- **self-reactive**- capable of participating in an autoimmune response.

- The destruction of intra cellular invaders is the function of the T lymphocytes, which are responsible for the **cell-mediated immune responses** of adaptive immunity.

- **RNA-Seq** uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment.

- **Combinatorial libraries** are collections of chemical compounds, small molecules or macromolecules such as proteins, synthesized by combinatorial chemistry, in which multiple different combinations of related chemical species are reacted together in similar chemical reactions. Chemical synthesis methods are used to generate large groups of compounds that can themselves be elaborated in a similar combinatorial fashion.

- **amplicon**- a piece of DNA or RNA that is the source and/or product of amplification or replication events. It can be formed artificially, using various methods including polymerase chain reactions (PCR)

- **indel**- an insertion or deletion of bases in the genome of an organism (it's a type of mutation). In coding regions of the genome, unless the length of an indel is a multiple of 3, it will produce a frameshift mutation.

- **hot spot of mutation**- A region of DNA that exhibits an unusually high propensity to mutate.

- **Species richness** is the number of different species represented in an ecological community, landscape or region. Species richness is simply a count of species, and it does not take into account the abundances of the species or their relative abundance distributions. Species diversity takes into account both species richness and species evenness.

- **The original Simpson** index λ equals the probability that two entities taken at random from the dataset of interest (with replacement) represent the same type. Its transformation 1 therefore equals the probability that the two entities represent different types. This measure is also known as GiniSimpson index.

$$1 - \lambda = 1 - \sum_{i=1}^{R} p_i^2$$

- local **abundance** is the relative representation of a species in a particular ecosystem. It is usually measured as the number of individuals found per sample. The ratio of abundance of one species to one or multiple other species living in an ecosystem is referred to as relative species abundances.

- **junction** $\sim$ CDR3

- A **haplotype** (haploid genotype) is a group of alleles in an organism that are inherited together from a single parent.

- **conscount**: consensus count- the number of the raw sequences.

- **dupcount**: duplicate count- the number of molecules that came from the same sequence.

- **Synonymous mutations** are point mutations, meaning they are just a miscopied DNA nucleotide that only changes one base pair in the RNA copy of the DNA. A codon in RNA is a set of three nucleotides that encode a specific amino acid. Most amino acids have several RNA codons that translate into that particular amino acid. Most of the time, if the third nucleotide is the one with the mutation, it will result in coding for the same amino acid. This is called a synonymous mutation because, like a synonym in grammar, the mutated codon has the same meaning as the original codon and therefore does not change the amino acid. If the amino acid does not change, then the protein is also unaffected.

- **Nonsynonymous mutations** have a much greater effect on an individual than a synonymous mutation. In a nonsynonymous mutation, there is usually an insertion or deletion of a single nucleotide in the sequence during transcription when the messenger RNA is copying the DNA. This single missing or added nucleotide causes a frameshift mutation which throws off the entire reading frame of the amino acid sequence and mixes up the codons. This usually does affect the amino acids that are coded for and change the resulting protein that is expressed. The severity of this kind of mutation depends on how early in the amino acid sequence it happens. If it happens near the beginning and the entire protein is changed, this could become a lethal mutation.

- **untranslated region (or UTR)** refers to either of two sections, one on each side of a coding sequence on a strand of mRNA. If it is found on the 5' side, it is called the 5' UTR, or if it is found on the 3' side, it is called the 3' UTR. mRNA is RNA that carries information from DNA to the ribosome, the site of protein synthesis (translation) within a cell. The mRNA is initially transcribed from the corresponding DNA sequence and then translated into protein. However, several regions of the mRNA are usually not translated into protein, including the 5' and 3' UTRs.

- **Copy number variation (CNV)** is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals in the human population. Copy number variation is a type of structural variation: specifically, it is a type of duplication or deletion event that affects a considerable number of base pairs.

- **Polymorphism** is when there are two or more possibilities of a trait on a gene.