Medical Genetics

Sheet: Human Genetic Variation- 13

Writer: Hiba Abu Hayyeh

Scientific: Mothana Olimat

Final: Tala Khamis

Doctor: Bilal Azab

Human Genetic Variation

Overview

- ✓ The sequence of nuclear DNA is approximately 99.9% identical between any two unrelated humans.
- ✓ The remaining 0.1% -which is responsible for the genetic diversity-is translated into 3 million different nucleotides.
  - ⇨ Total number of nucleotides in haploid cells = 3 billions
  - ⇨ Total number of nucleotides in diploid cells = 6 billions
- ✓ There are similar DNA Sequences Between Human and Other Organisms
  - ⇨ Chimpanzee and human genomes are very similar.

The Concept of Mutation and Genetic Polymorphism

- ✓ A mutation is defined as a permanent change in the nucleotide sequence with a frequency below 1%.
- ✓ Polymorphism: existence of two or more alleles at a specific gene's locus within a population of at least 1% frequency.
  - ⇨ A variant with a frequency above 1%.

However, the terms "**mutation**" and "**polymorphism**" often lead to confusion due to **incorrect** assumptions of **pathogenic** and **benign** effects respectively.

- ⇨ INCORRECT: mutations are pathogenic while polymorphisms are not.
- ⇨ CORRECT: mutations could be pathogenic or non-pathogenic, same for polymorphisms.

For example, consider GJB2 gene:

- ✓ A deletion of G nucleotide at position 35 in the coding DNA sequence (c.35delG) causes bilateral prelingual sensorineural hearing loss
- ✓ It is found in 2% of the general population worldwide.
- ✓ A pathological polymorphism.

- ✓ So, whether a variant is formally considered a polymorphism or not depends **entirely** on whether its frequency in a population exceeds 1% of the alleles in that population.
  - ⇨ Frequency $< 1\% \mapsto$ mutation.
  - ⇨ Frequency $\geq 1\% \mapsto$ polymorphism.

- ✓ It is recommended that both terms (mutation and polymorphism) be replaced by the term "variant" ➔ permanent change in the DNA.

Allele Frequency and Ethnic Background

The adjacent table represents GJB2:c.109G>A / p.Val37Ile frequencies among different populations .

GJB2: c.109G>A indicates that the nucleotide "G" at position 109 is substituted for "A"(new) in GJB2 gene p.Val37Ile indicates that amino acid No.37 "Valine" changed to "Isoleucine" (new).

Allele count = Counts of each alternate allele for each site across all samples.

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
East Asian	1665	19952	96	0.08345
Ashkenazi Jewish	83	10342	0	0.008026
Other	31	7212	0	0.004298
Latino/Admixed American	95	35428	1	0.002681
European (Finnish)	42	25104	0	0.001673
European (non-Finnish)	179	128578	1	0.001392
African/African-American	25	24964	1	0.001001
South Asian	12	30584	0	0.0003924
XX	1083	129104	53	0.008389
XY	1049	153060	46	0.006854
Total	**2132**	**282164**	**99**	**0.007556**

- In our example, it equals to the number of alleles that contain "A" at position 109 in a certain population.

Allele number= Total number of observed alleles in called genotypes

- Number of alleles that have "A" at position 109 +"G" at position 109 in a certain population.

Allele Frequency= The **Allele Counts** divided by the total number of observed alleles (**Allele number**).

$$\text{Allele frequency} = \frac{\text{allele count}}{\text{Allele number}}$$

Regarding the table above, for East Asian population:

✓ Allele frequency for that population = 1665 ÷19952 = 0.08345 ≈ 8%
✓ So, it is considered a polymorphism (8% > 1%)

For South Asian population:

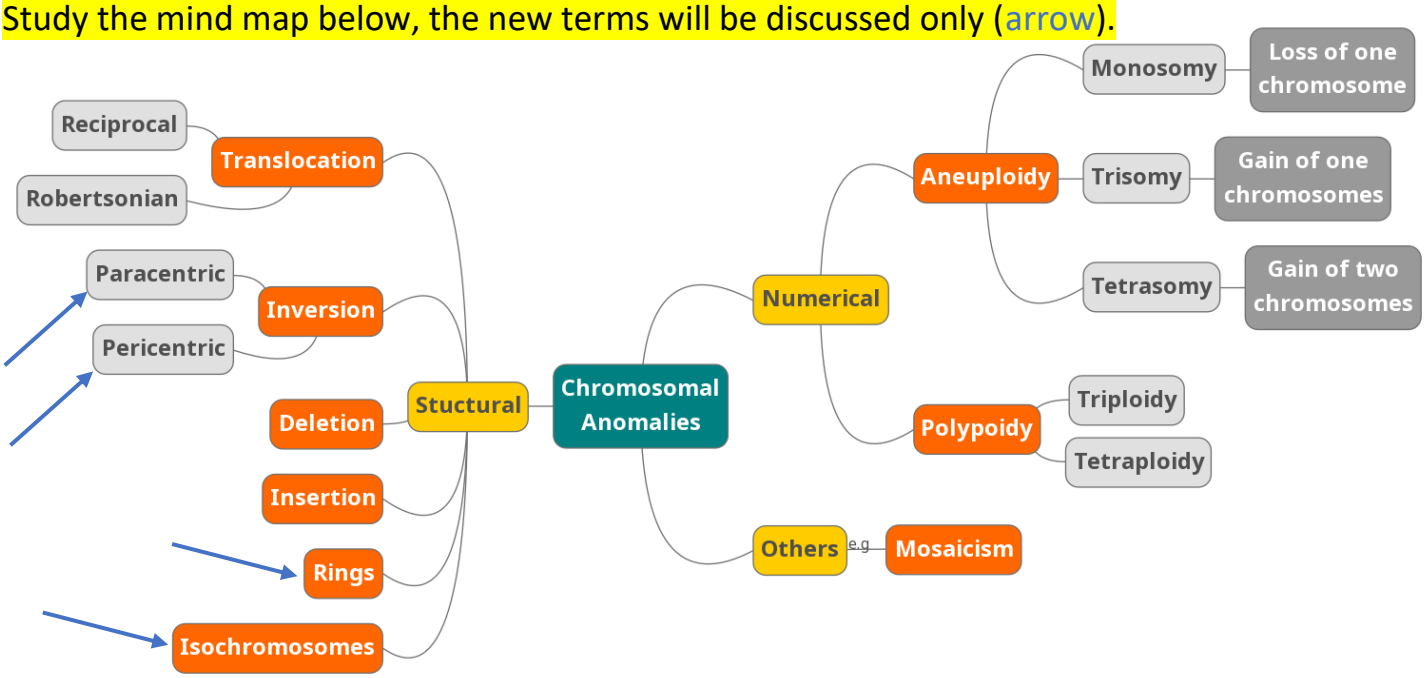
✓ Allele frequency for that population = 12 ÷30584 = 0.0003924 ≈ 0.04%
✓ So, it is considered a mutation (0.04% < 1%)

Note that the same variant could be considered a mutation in one population and a polymorphism in other.

Chromosomal Anomalies

- ✓ Chromosomal anomaly is a missing, extra, or irregular portion of chromosomal DNA.
- ✓ These can occur in the form of **numerical** abnormalities, where there is an atypical number of chromosomes, or as **structural** abnormalities, where one or more individual chromosomes are altered.

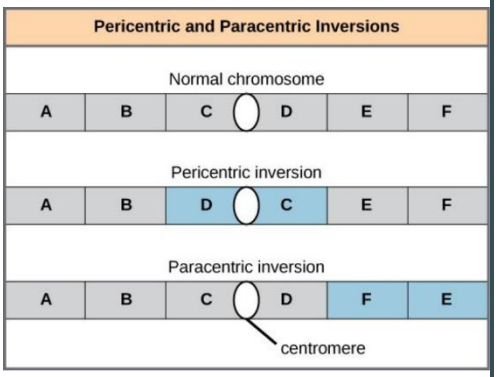
Study the mind map below, the new terms will be discussed only (arrow).



Regarding the following types of **structural chromosomal** abnormalities:

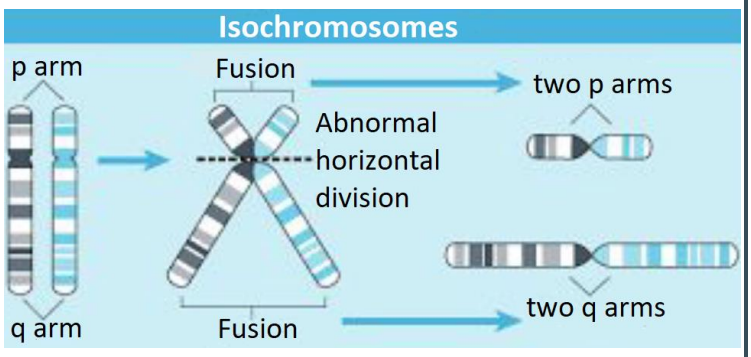
1- Inversion

- ✓ Chromosome rearrangement in which a segment of a chromosome is reversed end to end.
- ✓ Inversions are divided into two types depending on the location of the inversion with respect to the centromere:
  - a- **Paracentric inversion**: inversions that **do not** involve the centromere.
  - b- **Pericentric inversion**: inversions that **do** involve the centromere.



2- Isochromosomes

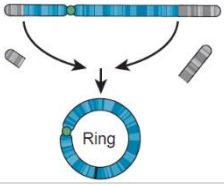
Chromosomes in which one arm is missing and the other duplicate in a mirror-image fashion. (Two p arms or two q arms in one chromosome).



3- Rings

Aberrant chromosome whose ends have fused together to form a ring.

Losing of telomerase →shortening of chromosome → sticky ends → ends fused, and ring is formed.



Types of mutations

Mutations occur at three different levels: Genome, chromosome and gene.

Genome mutations

- ✓ Mutations that affect the **number** of chromosomes in a cell.
- ✓ Arise from errors in chromosome segregation during meiosis or mitosis OR deletion or duplication of an **entire** chromosome that alter the dosage level of hundreds or thousands of expressed genes.
- ✓ Genome mutations produce chromosomal **aneuploidy**.
    - Down syndrome (trisomy 21) →aneuploidy →genome mutation
      Due to missegregation of a chromosome pair during meiosis.

- ✓ Genome mutations are the most common mutations seen in humans, with a rate of one missegregation event per 25 to 50 meiotic cell divisions
    - Every 25-50 cells divide, one out of them will be missegregated
    - This estimate is clearly a minimal one because the developmental consequences of many such events are likely so severe that the resulting fetuses are aborted spontaneously shortly after conception without being detected.
- ✓ Genome mutations are also common in cancer cells.

Chromosome mutations

- ✓ Mutations that alter the **structure** of chromosomes.
- ✓ Arise from chromosomal rearrangement:
    - ⇨ Duplications or triplications, deletions, inversions, and translocations of **only a part** of a chromosome **(not the entire chromosome).**
    - ⇨ Could occur spontaneously or due to abnormal segregation of translocated chromosomes during meiosis.
- ✓ Happen at much **less** frequency than genome mutations.
- ✓ Chromosome mutations are also frequently seen in cancer cells.
- ✓ Karyotyping is the standard procedure to examine chromosomal structures.

Although the frequencies of genome and chromosome mutations may seem high, these mutations are rarely perpetuated from one generation to the next because they are usually incompatible with survival or normal reproduction.

Gene mutations

- ✓ Mutations that alter DNA sequence of the nuclear or mitochondrial genomes.
- ✓ Arise from base pair **substitutions**, **insertions**, and **deletions**. Can originate by either of two basic mechanisms (discussed in the next pages):
    - Errors introduced during the normal process of DNA **replication**.
    - Failure to repair and return the DNA to its original sequence after **damage**.

- ✓ Rage from a change in as little as a single nucleotide (point mutation) to changes that may affect many millions of base pairs.

✓ Some mutations are spontaneous, whereas others are induced by physical or chemical agents called mutagens, because they greatly enhance the frequency of mutations.

Replication Errors

- ➢ The enzyme DNA polymerase faithfully duplicates the two strands of the double helix based on strict base-pairing rules (A pairs with T, C with G) but introduces one error every 10^{10} nucleotides.
- ➢ Additional proofreading (repair) then corrects more than 99.9% of these errors of DNA replication.
- ➢ Since human diploid cells have 6×10^9 base pairs, each whole cell DNA replication introduces less than 1 new base pair mutation in a cell division.

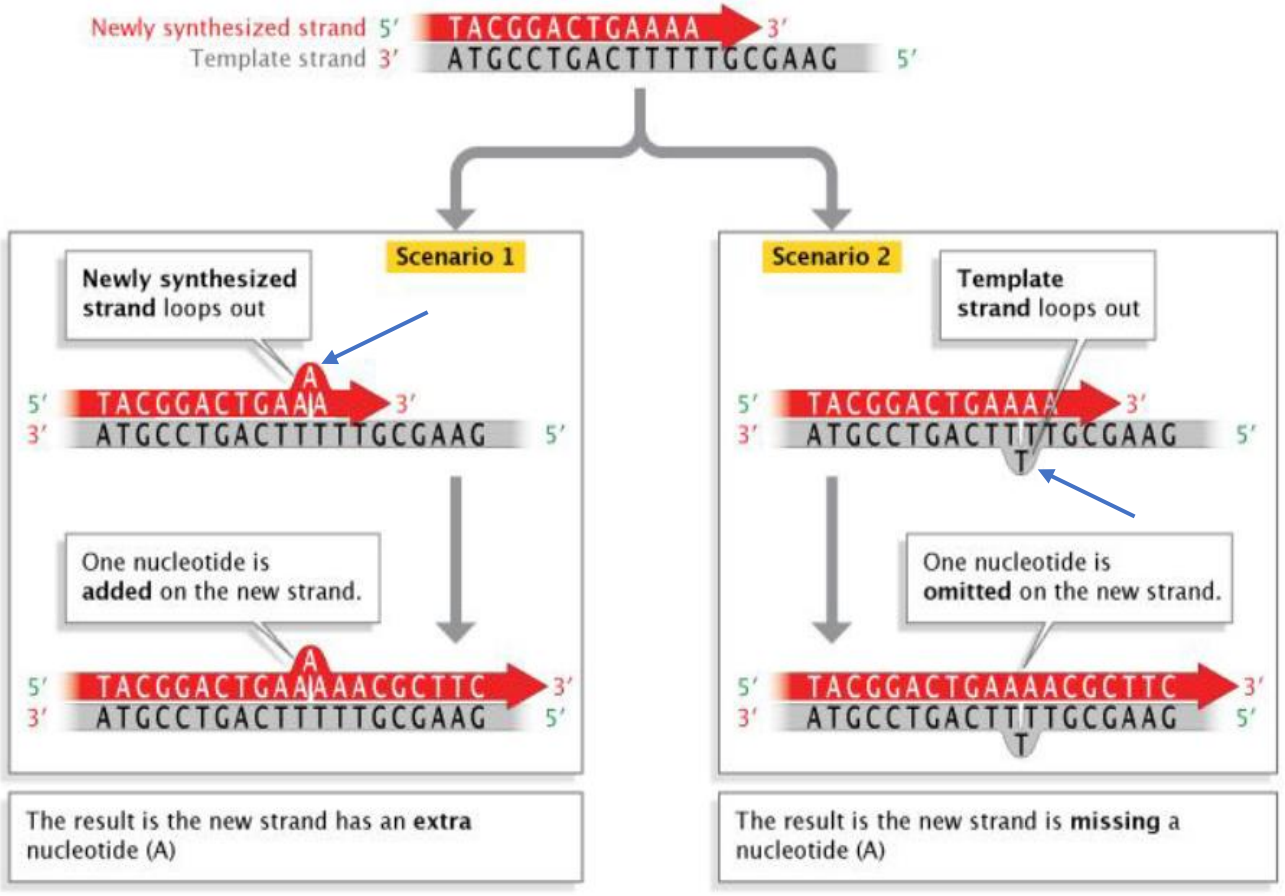
✓ Replication errors (Indel) occur due to strand slippage through 2 scenarios (see figure below):

Scenario 1

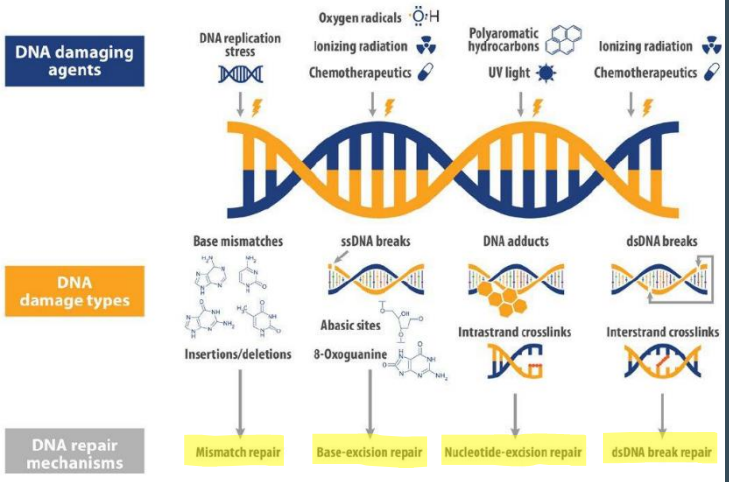
a) Slippage and looping out of the **strand being synthesized**.
b) The nucleotide at the top of the loop won't be recognized by DNA polymerase.
c) So, DNA polymerase will **add** the same unrecognized nucleotide again (arrow)
d) The result is an **insertion** gene mutation.

Scenario 2

a) Slippage and looping out of the **template strand**.
b) The nucleotide at the top of the loop won't be recognized by DNA polymerase.
c) So, DNA polymerase will **skip** the unrecognized nucleotide (arrow).
d) The result is a **deletion** gene mutation.



Mutagens



- ✓ A mutagen is a physical or chemical agent that permanently changes genetic material thus increases the frequency of mutations above the natural background level.
- ✓ 10,000 and 1,000,000 nucleotides are damaged per human cell per day by:
  - Spontaneous chemical processes such as depurination, demethylation, or deamination.
  - Reaction with chemical mutagens (natural or otherwise) in the environment.
  - Exposure to ultraviolet or ionizing radiation.
- ✓ Some but not all of this damage is repaired.
- ✓ Even if the damage is recognized and excised, the repair machinery may not read the complementary strand accurately and, consequently, will create mutations by introducing incorrect bases.
- ✓ Thus, in contrast to replication errors, which are usually corrected through proofreading mechanisms, nucleotide changes introduced by DNA damage and repair often result in permanent mutations. (It is easier to correct replication errors than damage-caused errors).
- ✓ DNA change could be beneficial if it leads to adapting to certain environment (natural selection).

Factors Influencing Mutation Rates

1- Chromosomal abnormalities are more likely with increased maternal age due to meiotic arrest
   - ⇨ During oogenesis, meiosis II is arrested in metaphase II until fertilization "An egg met a sperm."
   - ⇨ This prolonged arrest predisposes oocytes to missegregate their chromosomes.
   - ⇨ e.g. Down syndrome
2- Point mutation frequency increases with paternal age due to increased germ-cell divisions.
   - ⇨ Stem cells that give rise to sperm are continuously dividing. Thus, the stem cells of an older man are more likely than those of a younger man to have sustained a mutation arising from an error during DNA replication.
   - ⇨ e.g. Achondroplasia, 80% of de novo cases- fathers tend to be older.
3- Mitochondria have much increased mutation rates due to lack of repair systems.

Assume two fathers, father 1 is 50 years old and had a baby with achondroplasia, father 2 is 50 years old and had a normal baby. You sequenced their DNAs, and no related mutations were found. Which father has a higher probability for having a second affected baby?

Both of them don't have any related mutation in their DNAs. for them to have an affected baby, they have to develop a **spontaneous** mutation which is associated with age.

Both are old (same age) = Same probability.

Sequence Variant Nomenclature

Overview

✓ We'll cover HGVS-nomenclature which serves as international standard.
✓ It is used to report and exchange information of variant found in DNA, RNA, and protein sequences.

Reference Sequences

✓ variants can occur at different levels (DNA, RNA or protein level).
  ⇨ Each level has different types (coding vs noncoding DNA) (mRNA vs tRNA) etc...
  ⇨ So, each level has different reference sequence files.
  ⇨ Reference sequence is described using a prefix preceding the variant description.

We'll discuss DNA reference sequences only:

- c. = Coding DNA reference sequence (Codes for proteins).
- n. = Non-coding DNA reference sequence (does not code for proteins).
- g. = Linear genomic reference sequence.
- o. = Circular genomic reference sequence.
- m. = Mitochondrial reference (special case of a circular genomic reference sequence).

Note: p. and r. indicate a variant at protein level and RNA level respectively.

We'll discuss "variant nomenclature" for variants that occur in **coding DNA** and **proteins**

Abbreviations in variant descriptions

- "**>**" (greater than) is used to describe **substitution** variants (not used in protein level).
  c.76A>T means that**: A is substituted for T(new) at position 76 in a coding DNA.**

- "**_**" (underscore) is used to indicate a range, separating the first and the last residue:
  c.76_78delACT = nucleotides number 76, 77 and 78 which are (A, C and T) are deleted in a coding DNA.

- "**dup**" indicates a **duplication:**
  c.90_92dupACC = nucleotides number 90, 91 and 92 which are (A, C and C) are duplicated in a coding DNA.

- "**del**" indicates a **deletion:**
  c.127delA = the nucleotide A at position 127 is deleted in a coding DNA

- "**ins**" indicates an **insertion:**
  c.77insG = the nucleotide G at position 77 is inserted in a coding DNA

- "**Delins**" indicates **deletion and insertion:**
  c.56_58delinsCATG = nucleotides number 56, 57 and 58 are deleted, and nucleotides C, A, T and G are inserted instead.

***Note: Indle = Delins, but Indle is not used in HGVS nomenclature.**

Variant Nomenclature: cDNA

Coding DNA can be divided to 3 major regions:

1- Protein coding region (exons)

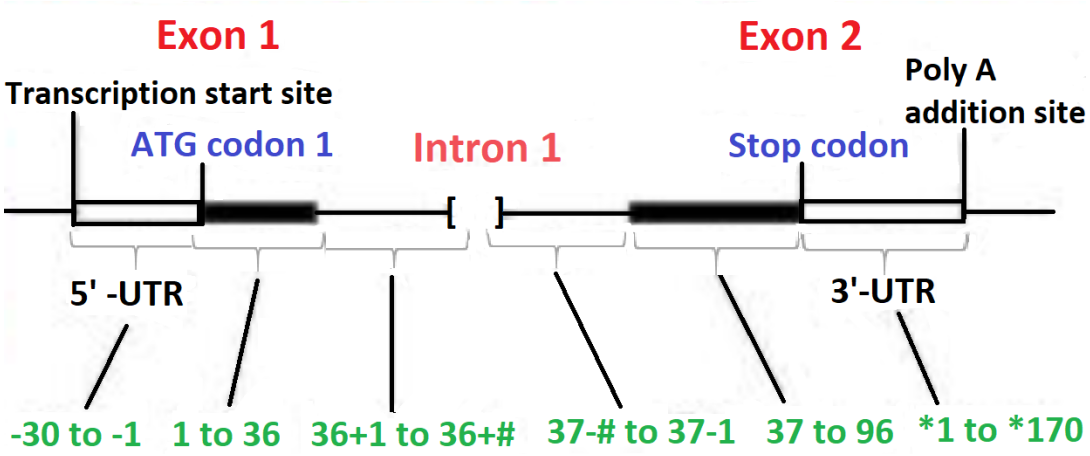
- ✓ Numbering starts with "c.1" at the **A** of the ATG translation initiation (start) codon and ends with the last nucleotide of the translation termination (stop) codon, i.e. TA**A**, TA**G**,

2- Untranslated region (UTR)

- ✓ Nucleotides upstream (5') of the ATG-translation initiation codon (start) are marked with a "**-**" (minus) ……see figure below **-30 to -1**
- ✓ nucleotides downstream (3') of the translation termination codon (stop) are marked with a "*****" (asterisk) ***1 to * 170**

3- Introns

- ✓ Nucleotides at the 5' end of an intron are numbered relative to the last nucleotide of the directly upstream exon, followed by a "**+**" (plus) and their position in the intron. **36+1**
- ✓ Nucleotides at the 3' end of an intron are numbered relative to the first nucleotide of the directly downstream exon, followed by a "**-**" (minus) and their position out of the intron. **37-1**



Variant Nomenclature: Protein

General rules:

- 3-letter amino acid code is preferred to describe the amino acid residue (Lys vs K for lysine)
- Methionine is always the first translated amino acid (start codon) so it is numbered as residue 1 (Met1 or M1).
- Ter or "*" designating a translation termination codon.

Some protein variants and their nomenclature:

- **Silent change**: do not result in amino acid change
  p.Leu54Leu or p.=
  Amino acid Leu at position 54 is not changed.
- **Substitution**: one amino acid is replaced by one other amino acid.
  p.Trp26Cys
  Amino acid Trp at position 26 is replaced by Cys(new)

- **Nonsense variant**: single base substitution resulting in a stop codon
  p.Trp26Ter or p.Trp26*
  Amino acid Trp at position 26 is changed to stop codon.

- **No-stop change (extension)**: a sequence change extending the reference amino acid Sequence at the N- or C-terminal end with one or more amino acids.
  p.Ter110GlnextTer17 or p.*110Glnext*17
  a variant in the stop codon (Ter/*) at position 110, changing it to a Gln-codon (a no-stop variant) and adding a tail of new amino acids to the protein's C-terminus, ending at a new stop codon (Ter/*) at position 17.

- **In -frame deletions**: deletion of 3 or multiple of 3 nucleotides (no frame shift occurs).
  p.Gln8del = amino acid Gln at position 8 is deleted.
  p.Cys28_Met30del =deletion of three amino acids, from Cys at position 28 to Met at position 30.
- **Duplications**: a copy of one or more amino acids are repeated.
  p.Gly4_Gln6dup
  Duplication of the amino acids from "Gly" at position to "Gln" at position 6.
- **Insertions**: one or more amino acids are added.
  p.Lys2_Gly3insGlnSerLys
  the insertion of amino acids GlnSerLys between amino acids Lys at position 2 and Gly at position 3.
  MetLysGlyHisGlnGlnCys →MetLys**GlnSerLys**GlyHisGlnGlnCys
- **Frameshifts**: deletion, insertion or Indle that change the reading frame (usually leads to premature stop codon).
  p.Arg97fs or p.Arg97Profs*23
  a variant with Arg at position 97 as the first amino acid changed, shifting the reading frame, replacing it for a Pro and terminating at position Ter23 in reference to the new reading frame (starting with Proline as amino acid 1).

Other protein variants:

- ✓ **Silent (Synonymous)** = does not result in amino acid change
- ✓ **Missense (nonsynonymous)** = changes a codon specific for one amino acid to specify another amino acid
- ✓ **Nonsense** = single base substitution resulting in a stop codon
- ✓ **Frameshift** = involves a deletion, insertion or indel that changes the reading frame (usually results in a premature stop codon)
- ✓ **Regulatory mutation**= affect promotor, enhancer or UTR.
- ✓ **Dynamic mutation**=amplification of repeat sequences (Fragile X, Huntington)
- ✓ **Splice site mutation**= inserts, deletes or changes a number of nucleotides in the specific site at which splicing takes place during the processing of pre-mRNA into mature mRNA
  ⇨ Affect splice donor (5' end of the intron) or acceptor (3' end of the intron).

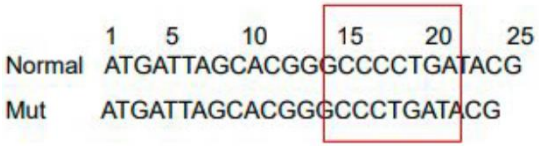
Note: When you can't tell which nucleotide is changed:

- For DNA and RNA → assign the most 3' position.
- For proteins → assign the most C-terminal position.

For example, here you can't tell which "c" has been deleted:

⇨ So, you arbitrary assign the most 3' one.

⇨ c.18delC



To Sum Up

- Allele frequency < 1% ↦ mutation.
- Allele frequency ≥ 1% ↦ polymorphism.

$$\text{Allele frequency} = \frac{\text{allele count}}{\text{Allele number}}$$

- **Structural** chromosomal abnormalities:
  - **1- Inversions**
    - a- Paracentric → **Do not** involve the centromere. ✗
    - b- Pericentric → Involve the centromere. ✓
  - **2- Isochromosomes** → 2 p arms OR 2 q arms in the same chromosome.
  - **3- Rings**→ Chromosomes with fused ends.

Types of Mutation and Their Frequencies			
Class of Mutation	**Mechanism**	**Examples**	**Frequency**
Genome Mutations	Chromosome missegregation	Aneuploidy	Most frequent
Chromosome Mutations	Chromosome rearrangement	Translocations	Intermediate frequency
Gene Mutations	Base pair mutation	Point mutations	Least frequent

- Base pair mutations in gene mutations occur due to:
  - 1- Replication errors (Indel) → Slippage of either the template strand or the new one.
  - 2- Mutagens→ Greatly enhance the frequency of mutations.

- Factors influence mutation rate: 📈
  - 1- Maternal age → Chromosomal abnormality → Down syndrome.
  - 2- Paternal age →Point mutation →Achondroplasia.
  - 3- Lack of repair system in mitochondria.

good luck