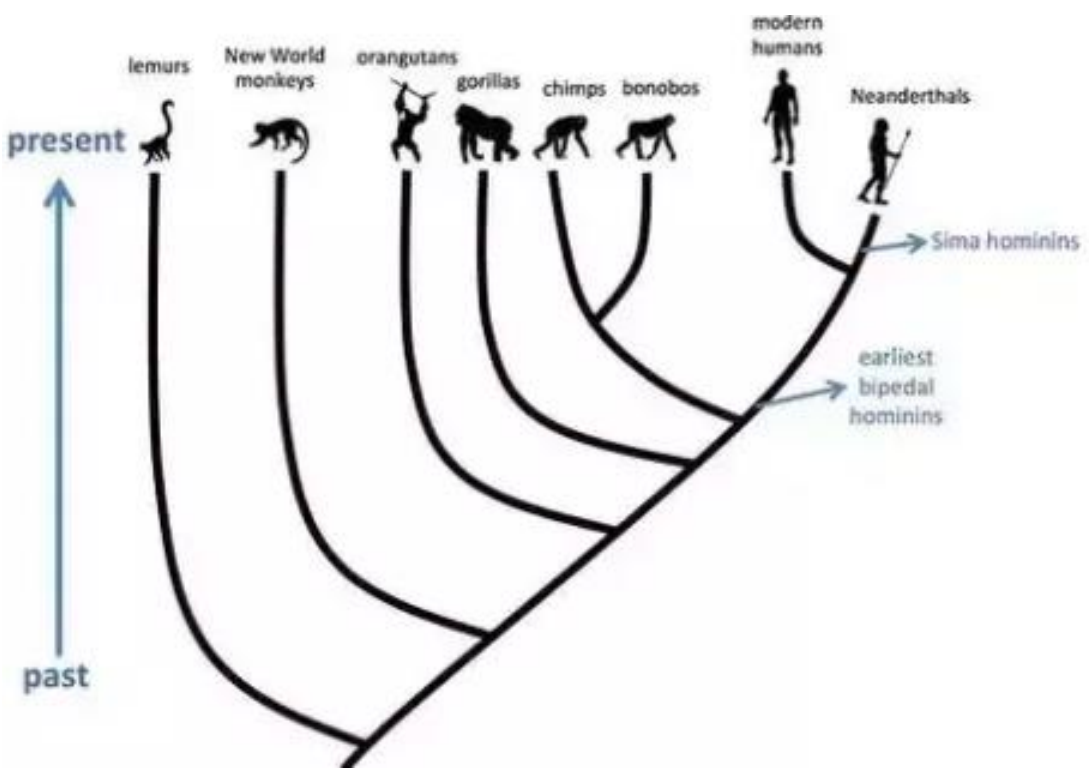
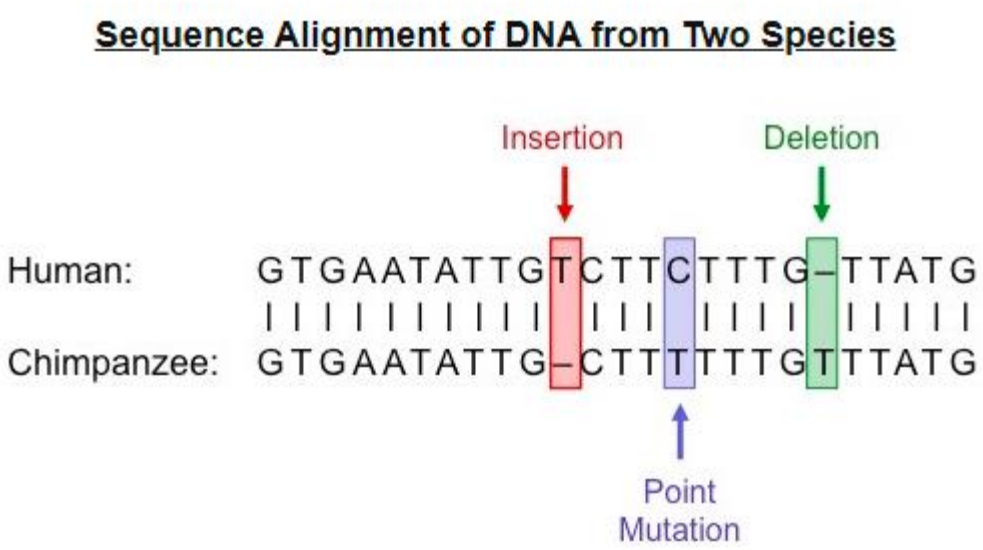
Genetic Variation





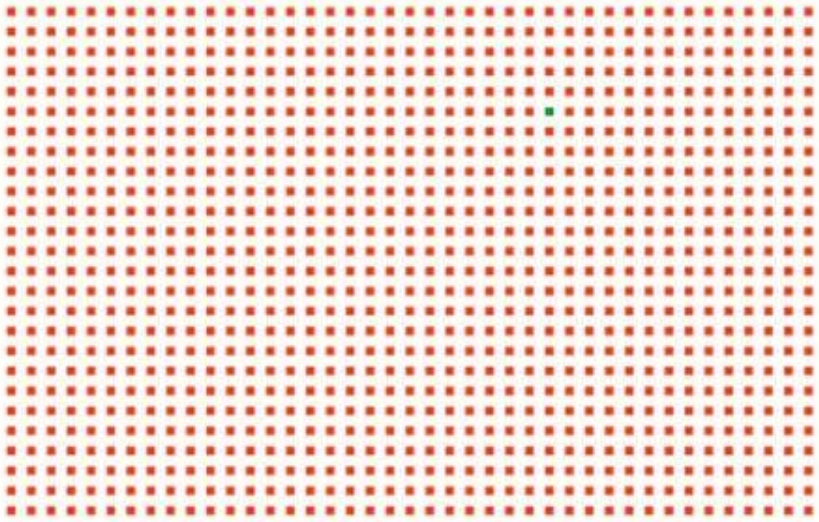
- **Sequence alignment** refers to comparing a DNA sequence from one individual with that of another individual or another species. This allows estimation of genetic relatedness between humans or between different species, and helps determine how closely related they are at the genetic level. In sequence alignment, mismatches observed between sequences can be considered as *mutations or genetic variants.*

	Gene Sequences That Codes for Proteins	Random DNA Segments Between Genes
Chimpanzee	100%	98%
Dog	99%	52%
Mouse	99%	40%
Chicken	75%	4%
Fruit Fly	60%	~0%
Roundworm	35%	~0%

Likelihood of Finding Similar DNA Sequences Between Human and Other Organisms

Similarity of DNA

- The human genome is over 3 billion base pairs long
- Two random people are 99.9% identical
- However, that still leaves 3 MILLION base pairs that can be different



If you align your DNA with that of another random individual, the sequences will match almost completely, except for about 0.1% differences, which corresponds to approximately 3 million base pairs.

- When we say "**mutation**," it may refer to a nucleotide substitution, insertion, or deletion. However, a mutation is always defined relative to a reference sequence. Therefore, the question arises: mutation relative to what reference? For this reason, the term "**variant**" is now often preferred, as it simply describes a difference from the reference sequence without implying pathogenicity.

- Both mutations and polymorphisms are **changes in the DNA sequence**. It is important to note that a ***mutation*** may be disease-causing or may have no clinical effect. Similarly, a ***polymorphism*** can be disease-causing or simply represent a benign DNA change with no clinical implications.

- Traditionally, if a specific DNA sequence change is present in **less than 1%** of the population, it is classified as a **mutation**, whereas if it is present in **more than 1%** of the population, it is classified as a **polymorphism**. However, the frequency alone does not determine pathogenicity. Some polymorphisms are pathogenic and disease-causing despite being present in more than 1% of the population. For example, the **Factor V Leiden variant**, which increases the risk of thrombosis, has an allele frequency of about 5% in some populations.

See the next slide

All DNA sequence variation arises via mutation of an ancestral sequence

	< 1%	≥ 1%
"Normal"	Rare variant or "private" polymorphism	polymorphism
"Disease"	Disease mutation	*Example: Factor V Leiden (thrombosis) 5% allele frequency*

Common but incorrect usage:

"a disease-causing mutation" **OR** *"a polymorphism"*

Genetic variation

- **Mutation:** A change in DNA sequence
  - Mutation ≠ deleterious change
  - Pathogenic mutation: DNA sequence changes responsible for causing disease or susceptibility to disease

- **Polymorphism:** Existence of two or more alleles of at least 1% frequency
  - Polymorphism ≠ neutral change
  - Alleles at a polymorphic locus can be pathogenic

    (e.g. GJB2 c.35delG - ~2% frequency)

GJB2:c.109G>A

chr13-20763612 C>T | p.Val37Ile | NM_004004.6 |

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**

See the next slide

- The table was obtained from the **gnomAD** (Genome Aggregation Database), a database developed through a collaboration between MIT, Harvard, and the Broad Institute in Boston. The goal of this initiative was to sequence the genomes of hundreds of thousands of supposedly healthy individuals. By analyzing these sequences, researchers can **determine the allele frequencies of genetic variants in the general population.**

- Understanding allele frequency is important for variant interpretation. For example, GJB2:c.109G>A | chr13:20763612 C>T | p.Val37Ile | NM_004004.6, **chromosome 13**, coordinate 20763612, there is a change from **C to T** at the **chromosomal level**. At the **coding sequence level**, c.109G>A refers to a change from **G to A** in the coding DNA sequence of the gene. *Because DNA is double-stranded, this C to T change corresponds to a G to A change on the complementary strand*. On the **protein level**, this results in p.Val37Ile, meaning a **substitution of valine at amino acid position 37 by isoleucine**. NM_004004.6 refers to the reference transcript used for this annotation.

- For example, in the East Asian population, the allele count for this variant is 1665 at the same genomic coordinate. If we divide 1665 by the total number of alleles (19952), we obtain the allele frequency shown in the last column, which is approximately 0.083 (about 8.3%). Therefore, in ***East Asians, this variant is considered a polymorphism***, since its frequency is **greater than 1%**. However, in the ***Ashkenazi Jewish population***, the same variant has an allele frequency of about 0.8% (0.008), which is **less than 1%,** and is therefore ***considered a mutation***. As a result, the same genetic variant can be classified differently depending on the population.

Mutation, polymorphism and variant

- "A mutation is defined as a permanent change in the nucleotide sequence with a frequency below 1%
- polymorphism is defined as a variant with a frequency above 1%
- The terms "mutation" and "polymorphism," however, which have been used widely, often lead to confusion because of incorrect assumptions of pathogenic and benign effects, respectively.
- **Thus, it is recommended that both terms be replaced by the term "<u>variant</u>""** ACMG 2015 guidelines American College of Medical Genetics

Categories of variation and their estimated frequencies

Table 9-1

Types of Mutation and Their Estimated Frequencies

Class of Mutation	Mechanism	Frequency (Approximate)	Examples
Genome mutation	Chromosome missegregation	$2\text{-}4 \times 10^{-2}$/cell division	Aneuploidy
Chromosome mutation	Chromosome rearrangement	6×10^{-3}/cell division	Translocations
Gene mutation	Base pair mutation	10^{-10}/base pair/cell division $10^{-5}\text{-}10^{-6}$/locus/generation	Point mutations

Based on Vogel F, Motulsky AG: Human Genetics, 3rd ed. Berlin, Springer-Verlag, 1997; and Crow JF: The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet 1:40-47, 2000.

- Genome mutations: affect the number of chromosomes in the cell, arising from errors (non disjunction) in chromosome segregation during meiosis or mitosis.
- A genome mutation that deletes or duplicates an entire chromosome alters the dosage and thus the expression levels of **hundreds or thousands of genes.** When a chromosome is deleted, it means that the expression level of those genes is impacted (decreased) . The same thing if it's duplicated (increased).
- Missegregation of a chromosome pair during meiosis causes genome mutations responsible for conditions such as trisomy 21 (Down syndrome).
- Genome mutations produce chromosomal aneuploidy and are the most common mutations seen in humans, with a rate of one missegregation event per 25 to 50 meiotic cell divisions.
- This estimate is clearly a minimal one because the developmental consequences of many such events may be so severe that the resulting aneuploid fetuses are spontaneously aborted shortly after conception without being detected.
- Genome mutations are also common in cancer cells. If you look at the chromosomal complement of tumor tissue, it is not uncommon for the number of chromosomes to deviate from 46. There are both gains and losses of chromosomes.

Table 9-1

Types of Mutation and Their Estimated Frequencies

Class of Mutation	Mechanism	Frequency (Approximate)	Examples
Genome mutation	Chromosome missegregation	$2\text{-}4 \times 10^{-2}$/cell division	Aneuploidy
Chromosome mutation	Chromosome rearrangement	6×10^{-4}/cell division	Translocations
Gene mutation	Base pair mutation	10^{-10}/base pair/cell division 10^{-5}-10^{-6}/locus/generation	Point mutations

Based on Vogel F, Motulsky AG: Human Genetics, 3rd ed. Berlin, Springer-Verlag, 1997; and Crow JF: The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet 1:40-47, 2000.

- Chromosome mutations: mutations that alter the structure (Not the number) of individual chromosomes. The changes involve only a part of a chromosome, such as **partial** duplications or triplications, deletions, inversions, and translocations, which can occur spontaneously or may result from abnormal segregation of translocated chromosomes during meiosis.

- Chromosome mutations, occurring at a rate of approximately one rearrangement per 1700 cell divisions, happen **much less frequently** than genome mutations.

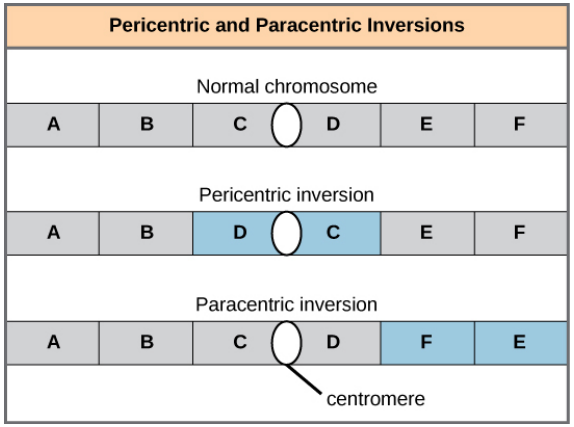
- Although the frequencies of genome and chromosome mutations may seem high, these mutations are rarely perpetuated from one generation to the next because they are usually incompatible with survival (lethal) or normal reproduction.

- Chromosome mutations are also frequently seen in cancer cells

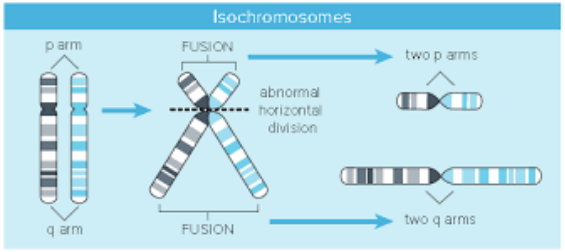
TYPES OF CHROMOSOME ANOMALIES

A chrmosomal anomaly can be:

- **Numerical**
  - Aneuploidy
    - autosomal
    - Sex chromosomal
    - Monosomy (loss of 1 chromosome)
    - Trisomy (gain of 1 chromosome)
    - Tetrasomy (gain of 2 chromosome)
  - Polyploidy
    - Triploidy
    - Tetraploidy
- **Structural**
  - gross
  - micro
  - Translocation (t):
    - Reciprocal
    - Robertonian
  - Inversions:
    - Paracentric
    - Pericentric
  - Deletions (del)
  - Insertions
  - Rings
  - Isochromosomes
- **Others**
  - e.g. Mosaicism



There are two types of **chromosomal inversions**: pericentric and paracentric. A *pericentric inversion* involves the centromere, meaning the breakpoints occur on both arms of the chromosome and the centromere is included in the inverted segment. For example, if a chromosome sequence ABCDEF undergoes a pericentric inversion around the centromere, the segment is flipped while including the centromere. In contrast, a *paracentric inversion* does not involve the centromere; both breakpoints occur within the same arm of the chromosome, so the centromere is not included. For example, a segment such as E–F can be inverted within one arm without involving the centromere. In addition, **deletions**, **insertions**, and **ring chromosomes** can occur; ring chromosomes form when telomeres are lost, leading to sticky chromosome ends that fuse together.



An **isochromosome** is an abnormal chromosome in which one arm is duplicated and the other arm is lost. Normally, each chromosome consists of a p arm and a q arm separated by a centromere. In an isochromosome formation, there is an abnormal fusion of either two p arms or two q arms. For example, isochromosome 5q means that chromosome 5 contains two q arms with loss of the p arm, while isochromosome 5p means that chromosome 5 contains two p arms with loss of the q arm.

Table 9-1

Types of Mutation and Their Estimated Frequencies

Class of Mutation	Mechanism	Frequency (Approximate)	Examples
Genome mutation	Chromosome missegregation	$2\text{-}4 \times 10^{-2}$/cell division	Aneuploidy
Chromosome mutation	Chromosome rearrangement	6×10^{-4}/cell division	Translocations
Gene mutation	Base pair mutation	10^{-10}/base pair/cell division 10^{-5}-10^{-6}/locus/generation	Point mutations

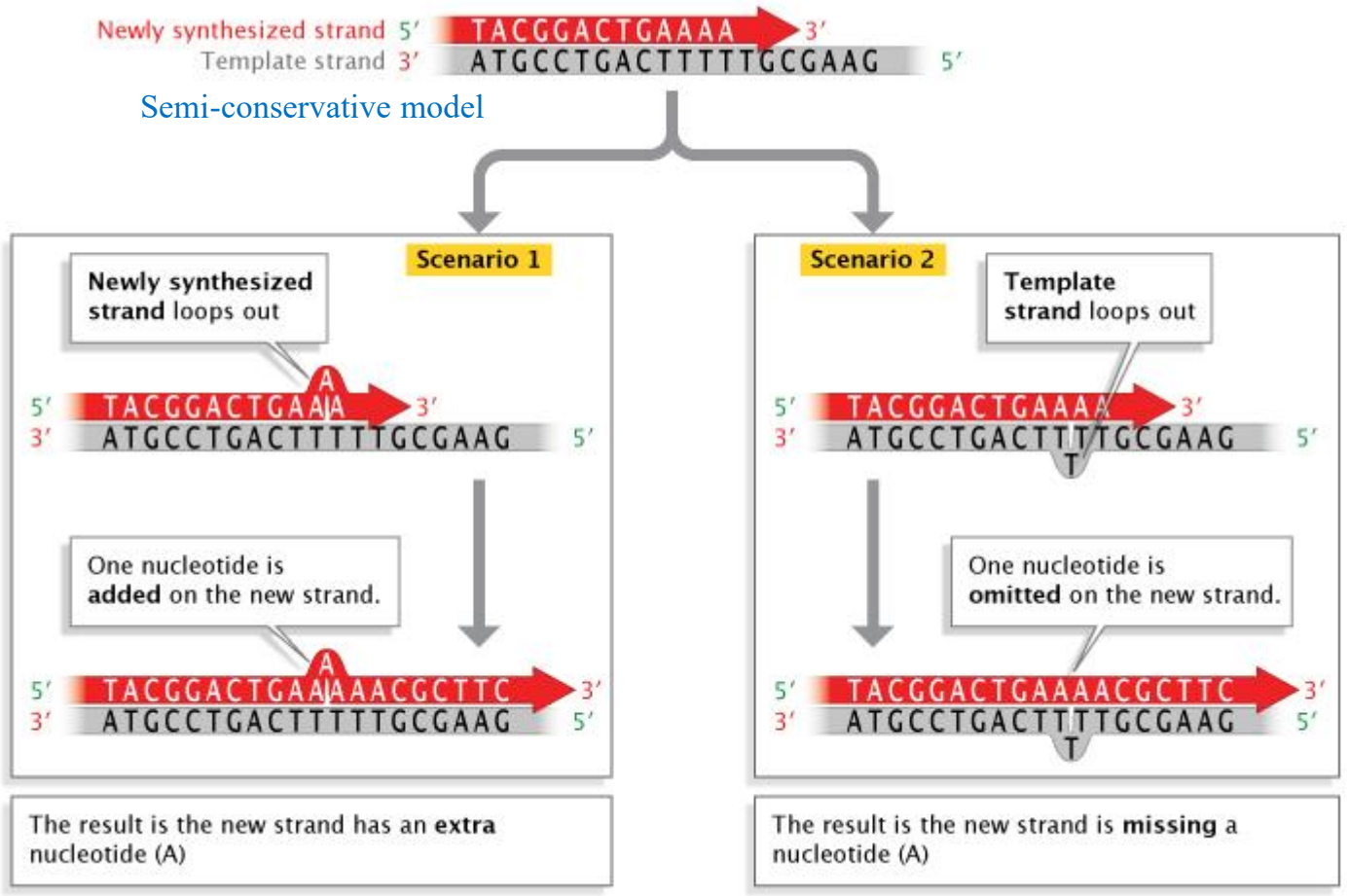
Based on Vogel F, Motulsky AG: Human Genetics, 3rd ed. Berlin, Springer-Verlag, 1997; and Crow JF: The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet 1:40-47, 2000.

- Gene mutations: mutations that alter individual genes.
- Gene mutations are changes in DNA sequence of the nuclear or mitochondrial genomes, ranging from a change in as little as a single nucleotide to changes that may affect many millions of base pairs.
- Gene mutations, including base pair substitutions, insertions, and deletions, can originate by either of two basic mechanisms:
  - ❑ errors introduced during the normal process of DNA replication, or
  - ❑ mutations arising from a failure to repair DNA after damage and to return its sequence to what it was before the damage.
- Some mutations are **spontaneous** (happening for no reason during the S phase), whereas others are induced by physical or chemical agents called **mutagens** (Anything that leads to a mutation, such as radiation, chemical agents), because they greatly enhance the frequency of mutations.

Replication Error

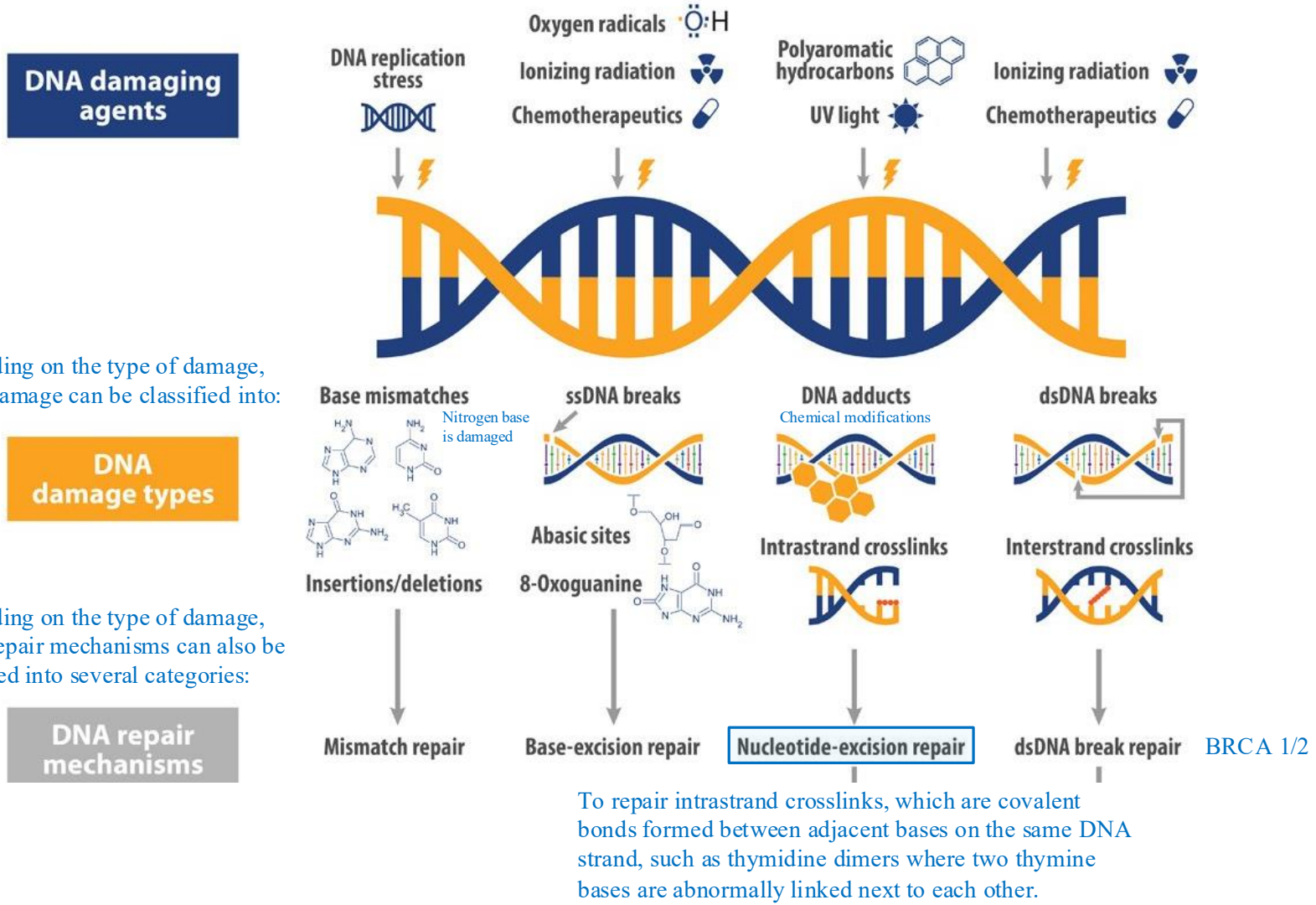
An incorrect nucleotide is introduced into one of the growing daughter strands only once every 10^{-10} million base pairs. (no repair)

Additional replication error checking corrects more than 99.9% of errors of DNA replication.

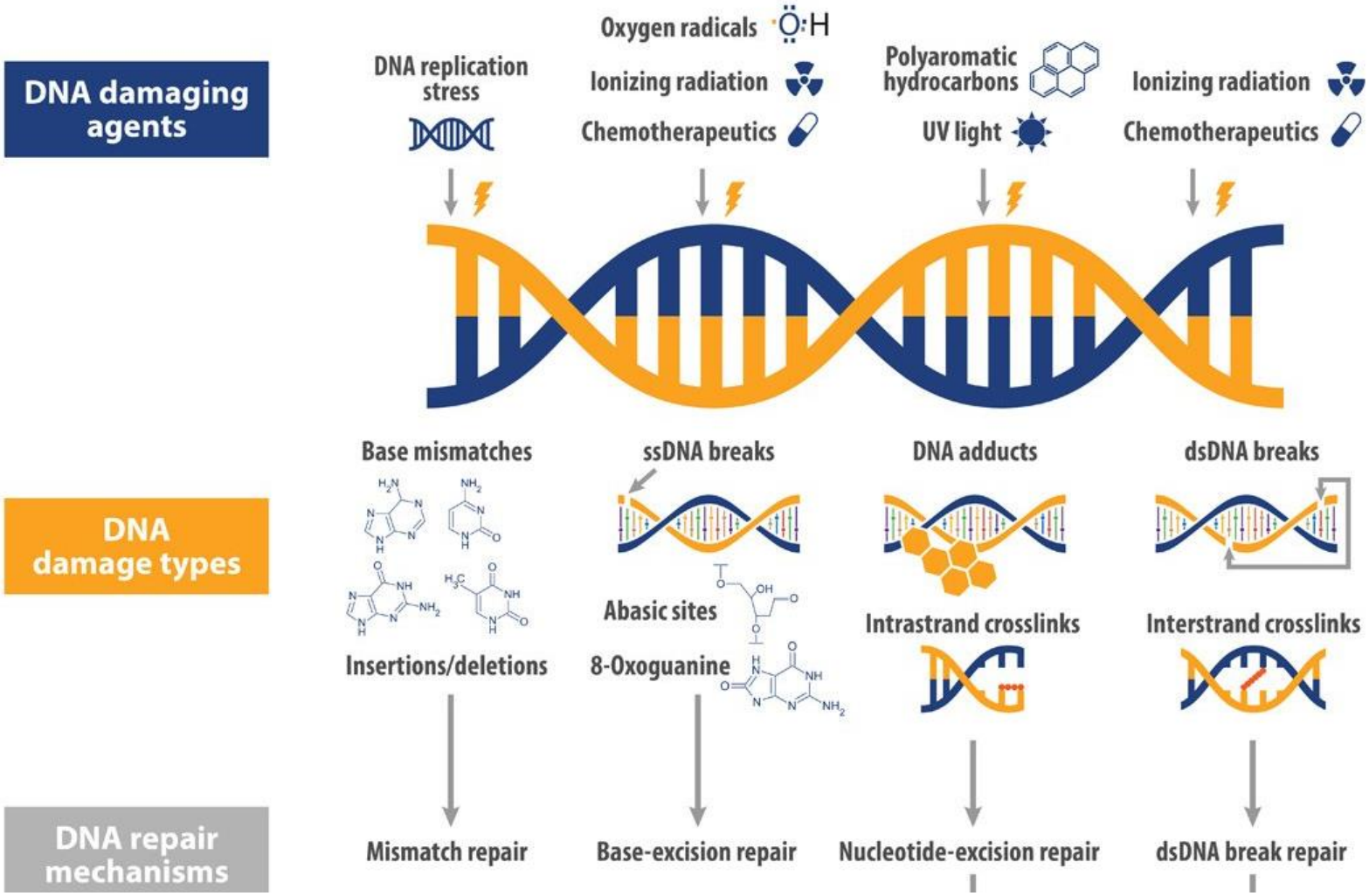


- Because the human diploid genome contains approximately 6×10^9 base pairs of DNA, replication errors introduce less than one new base pair mutation per cell division.

✓ A haploid genome contains about 3 billion nucleotides, while a diploid cell contains about 6 billion nucleotides. Even if only 0.1% of the genome differs, this corresponds to a significant number of mutations that are not corrected.



- 10,000 and 1,000,000 nucleotides are damaged per human cell per day by spontaneous chemical processes such as depurination, demethylation, or deamination; by reaction with chemical mutagens (natural or otherwise) in the environment; and by 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, as a consequence, will create mutations by introducing incorrect bases.

Thus, in contrast to replication-related DNA changes, which are usually corrected through proofreading mechanisms,

nucleotide changes introduced by DNA damage and repair often result in permanent mutations.

Factors influencing mutation rates

- Chromosomal abnormalities are more likely with increasing maternal age due to meiotic arrest
  Down's Syndrome

- Advanced paternal age is mainly at the DNA sequence level itself; therefore, Point mutation frequency increases with paternal age due to increased germ-cell divisions
  Achondroplasia: 80% de novo - fathers tend to be older

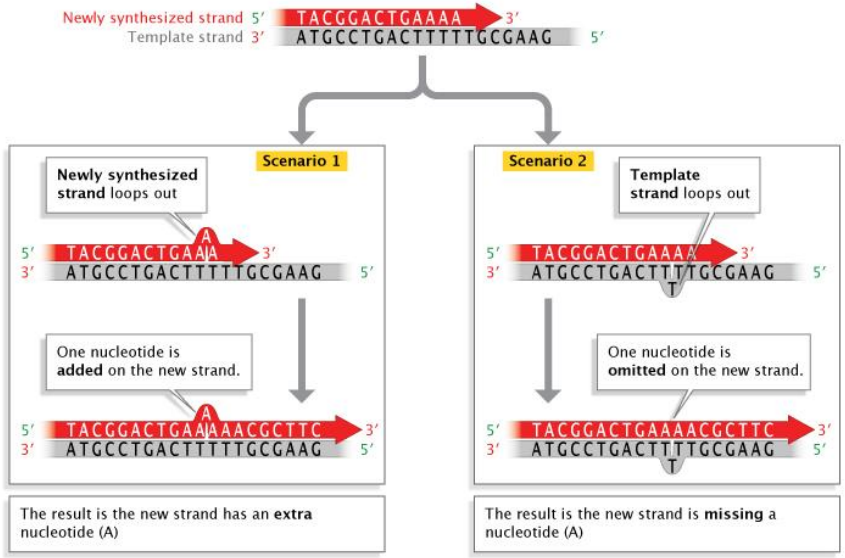
- Mitochondria have much increased mutation rates due to lack of repair systems because Mitochondria originated from endosymbiosis with bacteria.

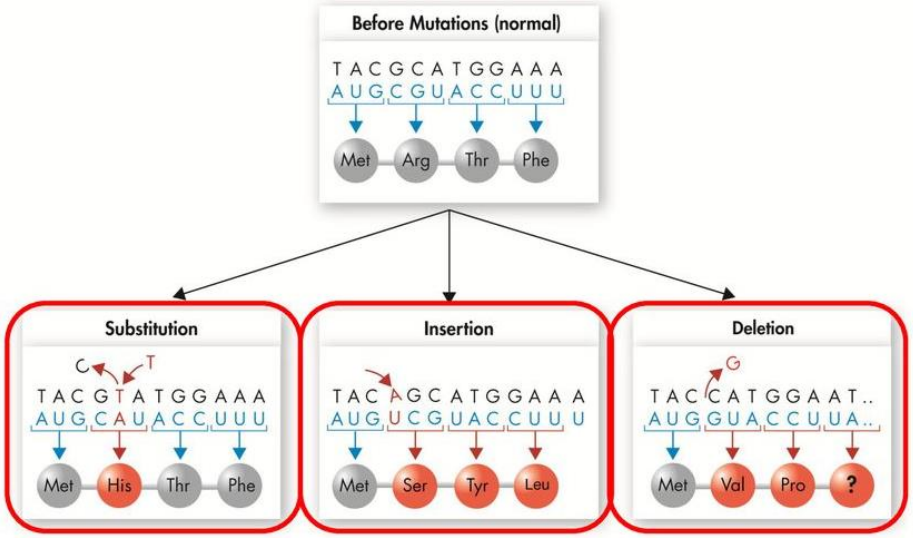
Gene Mutations: Point Mutations

- A point mutation is a change in a single nucleotide. There are three types of point mutations. SEE THE FIGURE

- There is a process called strand slippage or looping, through which small insertions or deletions can occur. Looping out may happen in the newly synthesized strand during DNA replication, leading to the insertion of extra nucleotides in the daughter strand, as shown in scenario 1 in the figure.

- In scenario 2, looping occurs in the template strand, meaning that the nucleotides that loop out during replication are not copied into the newly synthesized strand. As a result, this leads to a deletion in the daughter strand.





Gene and Variant nomenclature

- Similarly to organic chemistry, genetic variants and mutations follow a standardized nomenclature system, with databases such as HGVS used to ensure consistent naming, supported by dedicated websites.

Genes: https://www.genenames.org/

HGNC
HUGO Gene Nomenclature Committee

HGVS
HUMAN GENOME VARIATION SOCIETY

THE HUMAN VARIOME PROJECT

HUGO

Variant: https://varnomen.hgvs.org/

Sequence Variant Nomenclature

This site covers **HGVS-nomenclature**, the recommendations for the description of sequence variants. It is used to report and exchange information of variants found in DNA, RNA and protein sequences and serves as an international standard. When using the recommendations please cite: *Den Dunnen et al. 2016, Hum.Mutat. 37:564-569*. HGVS-nomenclature is authorised by the Human Genome Variation Society (HGVS), the Human Variome Project (HVP) and the Human Genome Organization (HUGO).

Reference Sequence Types

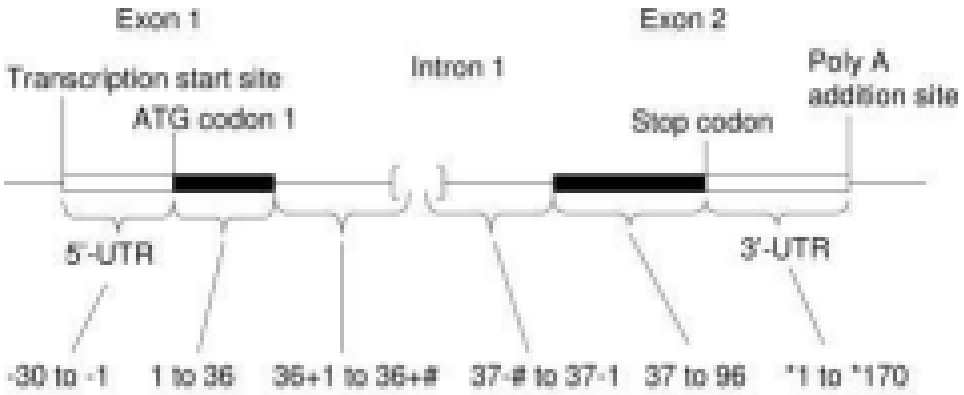
Depending on the variants to be reported, different reference sequence files are used at the DNA, RNA or protein level. It is mandatory to indicate the type of reference sequence file using a **prefix** preceding the variant description. Approved reference sequence types are **c.**, **g.**, **m.**, **n.**, **o.**, **p.** and **r.**:

- DNA
  - **g.** = linear genomic reference sequence
  - **o.** = circular genomic reference sequence
  - **m.** = mitochondrial reference (special case of a circular genomic reference sequence)
  - **c.** = coding DNA reference sequence (based on a protein coding transcript)
  - **n.** = non-coding DNA reference sequence (based on a transcript not coding for a protein)

Variant nomenclature: cDNA

- Nucleotide 1 is the A of the ATG initiation codon
- The nucleotide 5' of the ATG initiation codon is -1, the previous -2, etc.
- The nucleotide 3' of the stop codon is *1, the next *2, etc.
- Intronic nucleotides
  - **5' end of the intron**: the number of the last nucleotide of the preceding exon, a plus sign and the position in the intron, e.g., c.36+1G, c. 36+2T, etc.
  - **3' end of the intron**: the number of the first nucleotide of the following exon, a minus sign and the position upstream in the intron, e.g., c. 37-1G, c.37-2A

**Dividing the nucleotide number by 3 gives the number of the amino acid residue affected, in the example amino acid 26 (*36/3 = 12*)



# indicates any positive integer number

Ogino et al. (2007), JMD

Variant Nomenclature; cDNA

- Exons are sequences of DNA that remain present in the final mature RNA, while introns are spliced out during mRNA processing. After processing, any sequences removed are introns, and the remaining sequences are exons. However, parts of exons are non-coding.

- The number of introns in a gene is equal to the number of exons minus one.

- Coding DNA is related to mature mRNA (translated to protein). The symbol "c" refers to coding DNA, and c.1 represents the first nucleotide of the first exon, which is the start codon ATG. Accordingly, c.2 corresponds to T and c.3 corresponds to G.

- A minus notation such as c.-1 refers to the nucleotide immediately before (upstream) the first nucleotide of the start codon. For example, c.-2 indicates two nucleotides upstream of the A in the ATG start codon.

- After stop codons, the asterisk (*) is used in the nomenclature.

Variant Nomenclature; cDNA

- As shown in the figure in the previous slide, dark boxes represent exons, white boxes represent non-coding regions of exons, and thin lines represent introns.

- In exon 1 at the 5' end (upstream of ATG), there is a region that is part of the exon but does not code for protein, known as the 5' UTR. In exon 2, after the stop codon (downstream), there is also a non-coding region called the 3' UTR.
- A change at c.-10 indicates a mutation located 10 nucleotides upstream (towards the 5' direction) of the A in the start codon.

- Upstream of ATG, positions are numbered negatively, for example from -1 to -30 in this illustration, representing the 5' UTR.

- The last nucleotide of exon 1 here is c.36, and any nucleotide after (downstream) this becomes intronic and is assigned positions starting from +1. For example, c.36+1 refers to the first nucleotide immediately after exon 1; such intronic variants are important because they may affect splice donor sites and lead to disease.

- As you move toward exon 2, the first nucleotide of exon 2 would be c.37 (exon 1 ends with c.36), followed by c.38. Therefore, an intronic variant before exon 2 could be written as c.37-1, and c.37-5 indicates five nucleotides upstream relative to the first nucleotide of exon 2.

- After the stop codon, any sequence in the 3' UTR is indicated using the asterisk notation, for example *170, which refers to nucleotide 170 within the 3' UTR of exon 2.

Symbols for specific variation types

- ">" indicates a **substitution** at DNA level: c.76A>T
- "_" (underscore) indicates a **range** of affected residues, separating the first and last residue affected: c.76_78delACT
- "dup" indicates a **duplication**: c.90_92dupACC
- "del" indicates a **deletion**: c.127delA
- "ins" indicates a **insertion**: c.76_77insG
- "delins" indications a **deletion and insertion**: c.56_58delinsCATG
  - ***For all descriptions the **most 3' position** possible is arbitrarily assigned to have been changed

Typically, deleted nucleotides are not individually specified in the notation

```
          1    5         10        15        20        25
Normal    ATGATTAGCACGGGCCCCTGATACG
Mut       ATGATTAGCACGGGCCCTGATACG
```

We cannot know which C is deleted, so assign the most 3' position (c.18delC)

In homopolymer regions (where the same nucleotide is repeated), if a deletion occurs (for example four Cs becoming three Cs), it is not possible to determine which nucleotide is deleted. By convention, the most 3' position is reported as deleted in the nomenclature.

Variant nomenclature: Protein

In protein nomenclature, amino acids can be written using their full names, three-letter codes, or one-letter codes. Although one-letter codes are commonly used in molecular oncology and genetic testing reports, there is a general preference to avoid them in formal descriptions.

- 3-letter amino acid code is preferred to describe the amino acid residues (Lys vs. K for lysine)
- For all descriptions the **most C-terminal position possible** is arbitrarily assigned to have been changed
- Methionine encoded by the translation initiation site (*start codon*) is numbered as residue 1 ("**Met1**" or "**M1**")
- "**Ter**" or "*****" designating a translation termination codon

  It is recommended to use "Ter" instead of "*" to indicate a stop codon.

Variant nomenclature: Protein

- **Silent changes:** p.Leu54Leu or p.= (amino acids can be encoded by multiple codons)
- **Substitutions:** p.Trp26Cys
- **Nonsense variant:** p.Trp26Ter or p.Trp26*

A nonsense mutation refers to a change that creates a stop codon

- **No-stop change:** p.Ter110GlnextTer17 or p.*110Glnext*17
- **In-frame deletions:** p.Gln8del or p.Cys28_Met30del
- **Duplications:** p.Gly4_Gln6dup
- **Insertions:** p.Lys2_Met3insGlnSerLys
- **Frameshifts:** short description: p.Arg97fs
  long description: p.Arg97Profs*23

Frameshift mutations usually lead to the introduction of a premature stop codon.

where the "Arg97Pro" describes the substitution of Arg for Pro at position 97, "fs" indicating the frameshift and the "*23" describes the position of the translational termination (stop) codon in the new reading frame (starting with proline as amino acid #1)

*The "no-stop" mutation occurs when a normal stop codon is changed into an amino acid, causing translation to continue until a new stop codon is reached. In such cases, the protein becomes longer by 17 amino acids in total.

Variant Nomenclature: Protein

1. Silent Changes
Examples:
•p.Leu54Leu
•p.=
A DNA sequence change occurs, but the amino acid remains the same because multiple codons can encode the same amino acid.

2. Substitutions (Missense Variants)
Example:
•p.Trp26Cys**:**
•Trp = Tryptophan
•Cys = Cysteine
•26 = amino acid position
This means that tryptophan at position 26 has been replaced by cysteine.

3. Nonsense Variant
Examples:
•p.Trp26Ter
•p.Trp26*
Tryptophan at position 26 is replaced by a stop codon.

4. No-Stop Change (Stop-Loss Variant)
Examples:
•p.Ter110GlnextTer17
•p.*110Glnext*17
•Ter110 = the original stop codon was at position 110.
•Gln = the stop codon is changed to glutamine.
•extTer17 = translation continues for 17 additional amino acids before reaching a new stop codon.

5. In-Frame Deletions
Examples:
•p.Gln8del
•p.Cys28_Met30del
•p.Gln8del → glutamine at position 8 is deleted.
•p.Cys28_Met30del → amino acids from positions 28 to 30 are deleted.

6. Duplications
Example:
•p.Gly4_Gln6dup
The amino acids from Gly4 to Gln6 are repeated once.

7. Insertions
Example:
•p.Lys2_Met3insGlnSerLys
Between Lys2 and Met3, the amino acids: Gln (Glutamine) Ser (Serine) Lys (Lysine) have been inserted.

8. Frameshifts
Short Description Example: p.Arg97fs
Long Description Example: p.Arg97Profs*23
Interpretation of p.Arg97Profs*23:
•Arg97 = original amino acid at position 97 is arginine.
•Pro = arginine is replaced by proline.
•fs = frameshift.
•*23 = a new stop codon appears 23 amino acids downstream in the new reading frame.

Major types of gene mutations: definitions

Silent (synonymous) – does not result in amino acid change

Missense (nonsynonymous) – changes a codon specific for one amino acid to specify another amino acid

Deletion – loss of DNA, single bp to kb

Duplication – gain of DNA, single bp to kb

Nonsense – single base substitution resulting in a stop codon

Frameshift – involves a deletion, insertion, or indel that changes the reading frame (and usually leads to a premature stop codon)

Splice site – typically affect splice donor or acceptor

Regulatory mutations – affect promoter, enhancer or UTR

Regulatory mutations occur in non-coding regions that control gene expression levels rather than protein sequence.

Dynamic mutations –amplification of repeat sequences

(Fragile X, Huntington's)

Dynamic mutations are those in which the size of the mutation changes across generations.

Copy number variants refer to large genomic deletions or duplications, typically involving sequences of about 1 kilobase or more. Smaller changes involving only a few bases are classified as insertions or deletions (indels).

V0 → V1

Slide 19 → scenario 1 should be 2 and vice verse

Slide 22 → box in the lower left corner moved to slide 25

Slide 28 → added