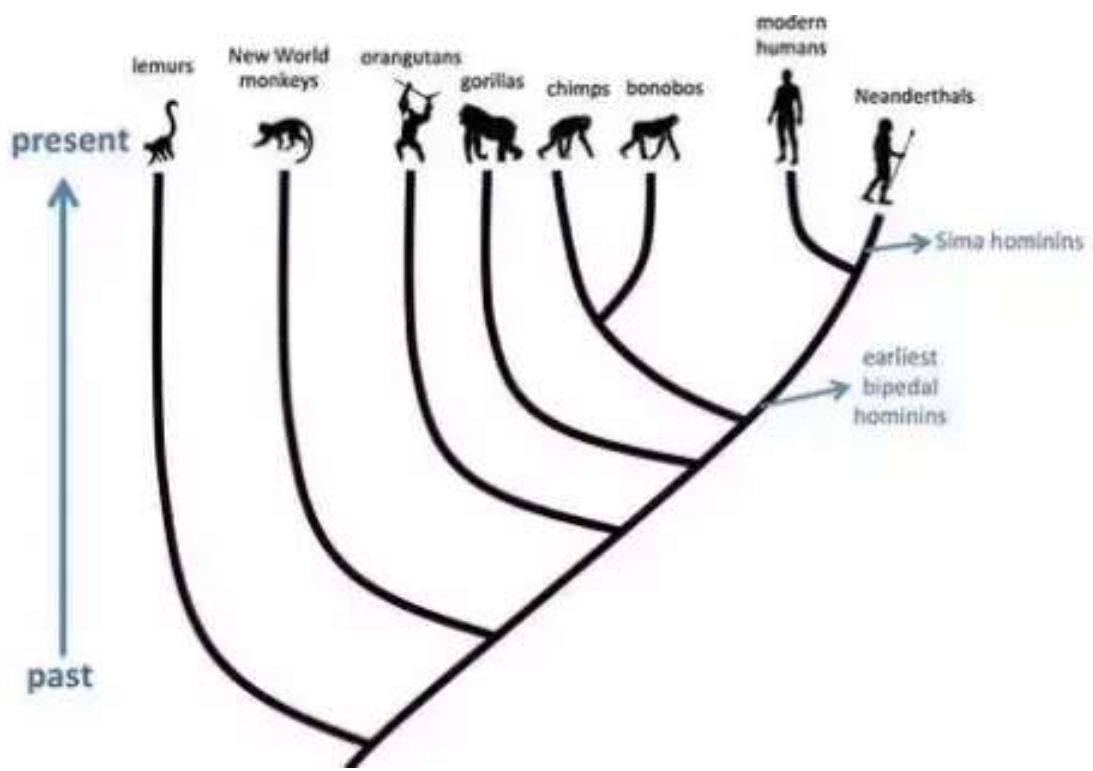
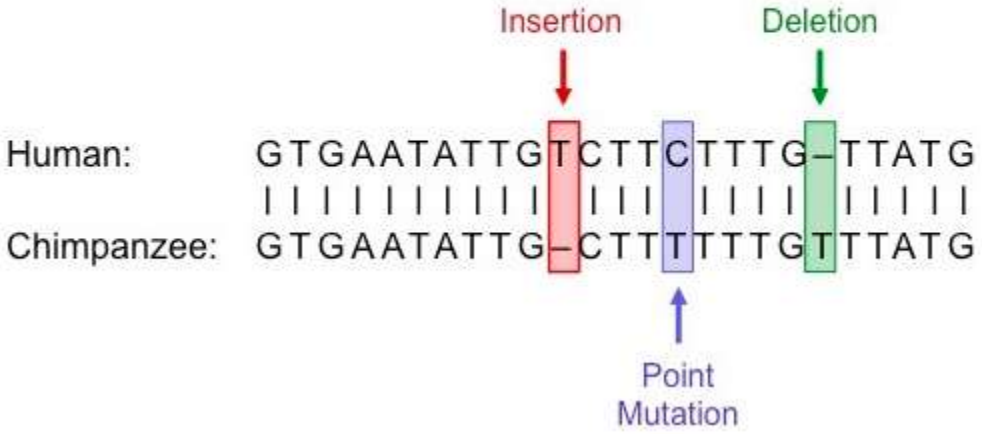
Genetic Variation



comparing DNA sequence of an individual with another individual or another species. → calculate relatedness between individuals & how genetically related are humans to other species. → any mismatches can be considered a mutation/variant.

Sequence Alignment of DNA from Two Species



* Sequence alignment between Human DNA & other mammals & vertebrates DNA shows striking similarity in DNA sequence	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%

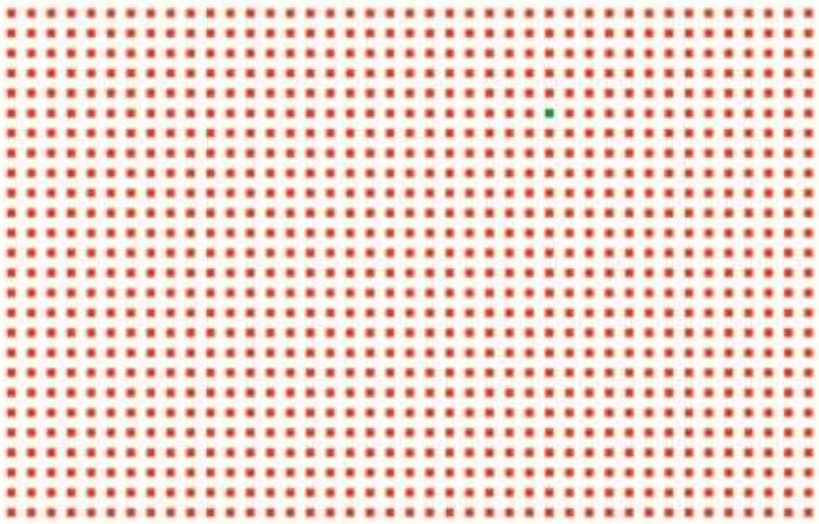
Likelihood of Finding Similar DNA Sequences Between Human and Other Organisms

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

↳represents the remaining 0.1%.

↳ if you align your DNA with someone else randomly, it will match except for 0.1%. which translates into 3 million base pairs.



All DNA sequence variation arises via mutation of an ancestral sequence

(there is a common misconception that a mutation is disease-causing while polymorphism is not disease-causing & this is WRONG)

* BOTH polymorphisms & mutations could be → normal (x disease)
  → disease-causing.

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

this disease-causing variant exists in 5% of the population "polymorphism"

*BOTH "mutation" & "Polymorphism" are changes in DNA sequence BUT

*the difference between "mutation" & "polymorphism" has to do with the frequency in the population
  → <1% → mutation
  → ≥1% → polymorphism.

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)

* this table is from a database "gnomAD" which is an initiative between MIT & Harvard
* the idea was to sequence hundreds of thousands of supposedly healthy individuals.

coding sequence of the specific gene (GJB2)

GJB2:c.109G>A

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

→ the same as the change from C→T on the chromosomal level since they are complementary. (depends on which strand we are looking at)

chromosomal level → coordinate 20763612 on ch.13, a change from C to T is seen

a.a # 37 is changed from Val to Ile

↑ protein level

* We can read the DNA change by
  - looking at the chromosome coordinate
  - looking at the (c.) which has to do with the coding sequence of the gene.

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**

Allele Count: total # of times this specific mutant allele (T in this example) was found.

Allele Number: total # of scanned alleles.

1665 out of 19952 alleles were found to have this variant ⇒ 1665 ÷ 19952 = 8.3% → >1% therefore it is considered as polymorphism in the East Asian population.

0.008026 = 0.8% → the same variant is considered a mutation in Ashkenazi Jews since it is <1%

Conclusion:

"Polymorphism" & "Mutation" terms could be used for the same variant in 2 different ethnic populations.

Therefore the term "Variant" is more preferably used especially in genetic testing reports of hereditary diseases.

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 "variant"" ACMG 2015 guidelines

American College of Medical Genetics

* a variant could be a mutation or polymorphism

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^{-4}/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 in chromosome segregation during meiosis or mitosis. (Non-disjunction)
- A genome mutation that deletes or duplicates an entire chromosome alters the dosage and thus the expression levels of hundreds or thousands of genes.
- 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. (high)
- 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. ⇒ underestimated bcz sometimes this event occurs early on & the embryo is lost before it is known that there is a pregnancy in the first place.
- Genome mutations are also common in cancer cells
  - ↳ if we look at the chromosomal complement in the tumor tissue, the # of chromosomes is commonly abnormal (NOT 46) there are gains or losses.

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 $10^{-5}\text{-}10^{-6}$/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.

part of a chromosome is deleted or duplicated or inverted or translocated, it could spontaneous or inherited

- Chromosome mutations: mutations that alter the structure 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.

structural NOT numerical.

- 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 or normal reproduction.

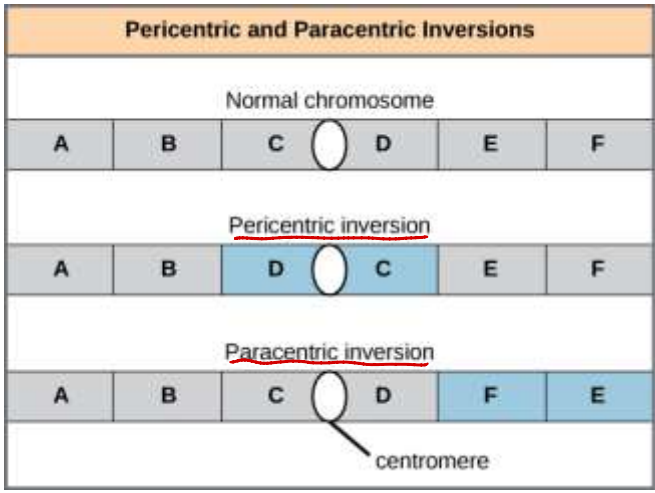
→ they happen frequently BUT NOT frequently inherited bcz they are lethal.

- 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:**
    - does NOT involve the centromere ← **Paracentric**
    - includes the centromere ← **Pericentric**
  - **Deletions (del)**
  - **Insertions**
  - **Rings** → occurs when telomeres are deleted & the ends become sticky & fuse together forming a "ring chromosome"
  - **Isochromosomes** → a chromosome with either 2 p arms OR 2 q arms
    - i.e: isochromosome 5q, i(5q) means that ch.5 has 2 q arms & NO p arms. OR i(5p) means that ch.5 has 2 p arms & NO q arms.
- **Others**
  - e.g. Mosaicism



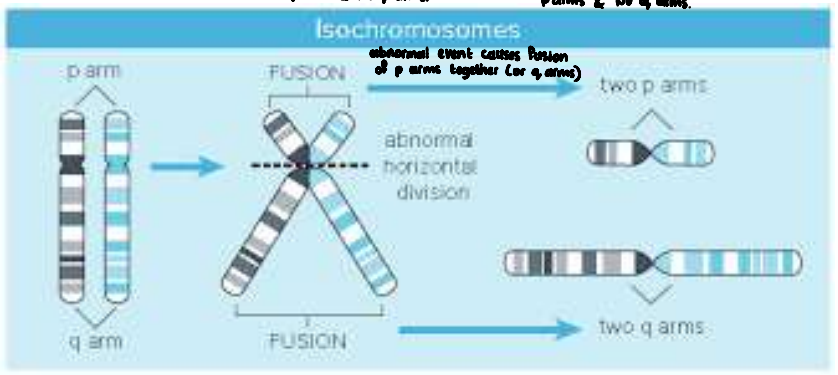


Table 9-1

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.

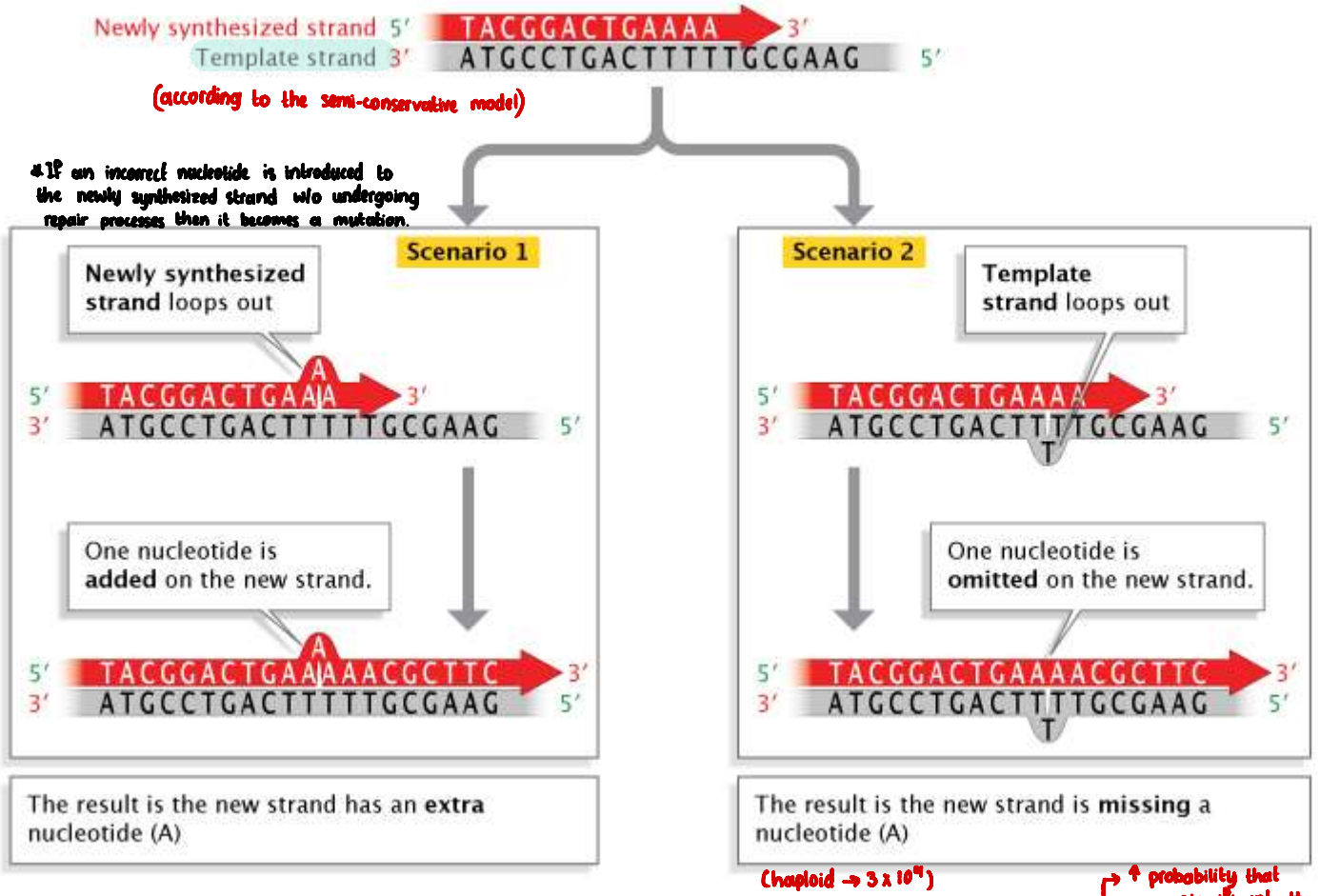
- Gene mutations, including base pair substitutions, insertions, and deletions, can originate by either of two basic mechanisms: When could it happen?
  - errors introduced during the normal process of DNA replication, or
  - mutations arising from a failure to repair DNA after damage and to return its sequence to what it was before the damage.

- Some mutations are spontaneous, occuring with no reason during S phase (DNA replication) whereas others are induced by physical or chemical agents called mutagens, because they greatly enhance the frequency of mutations.
  → Anything that leads to a mutation such as radiation, chemical agents, etc...

Replication Error

An incorrect nucleotide is introduced into one of the growing daughter strands only once every 10^{-10} million base pairs.

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

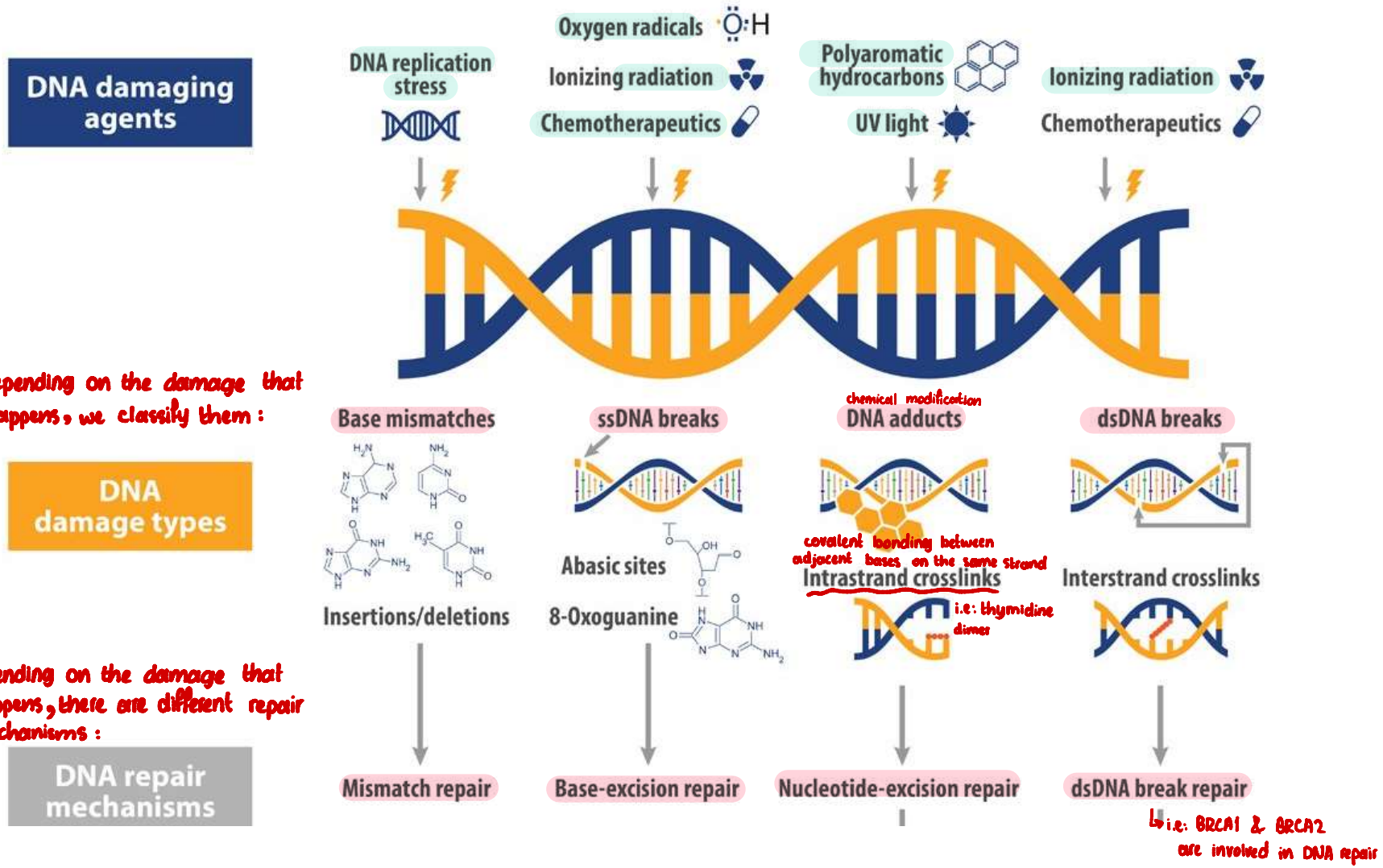


↳ 0.1% NOT corrected

(haploid → 3×10^9)

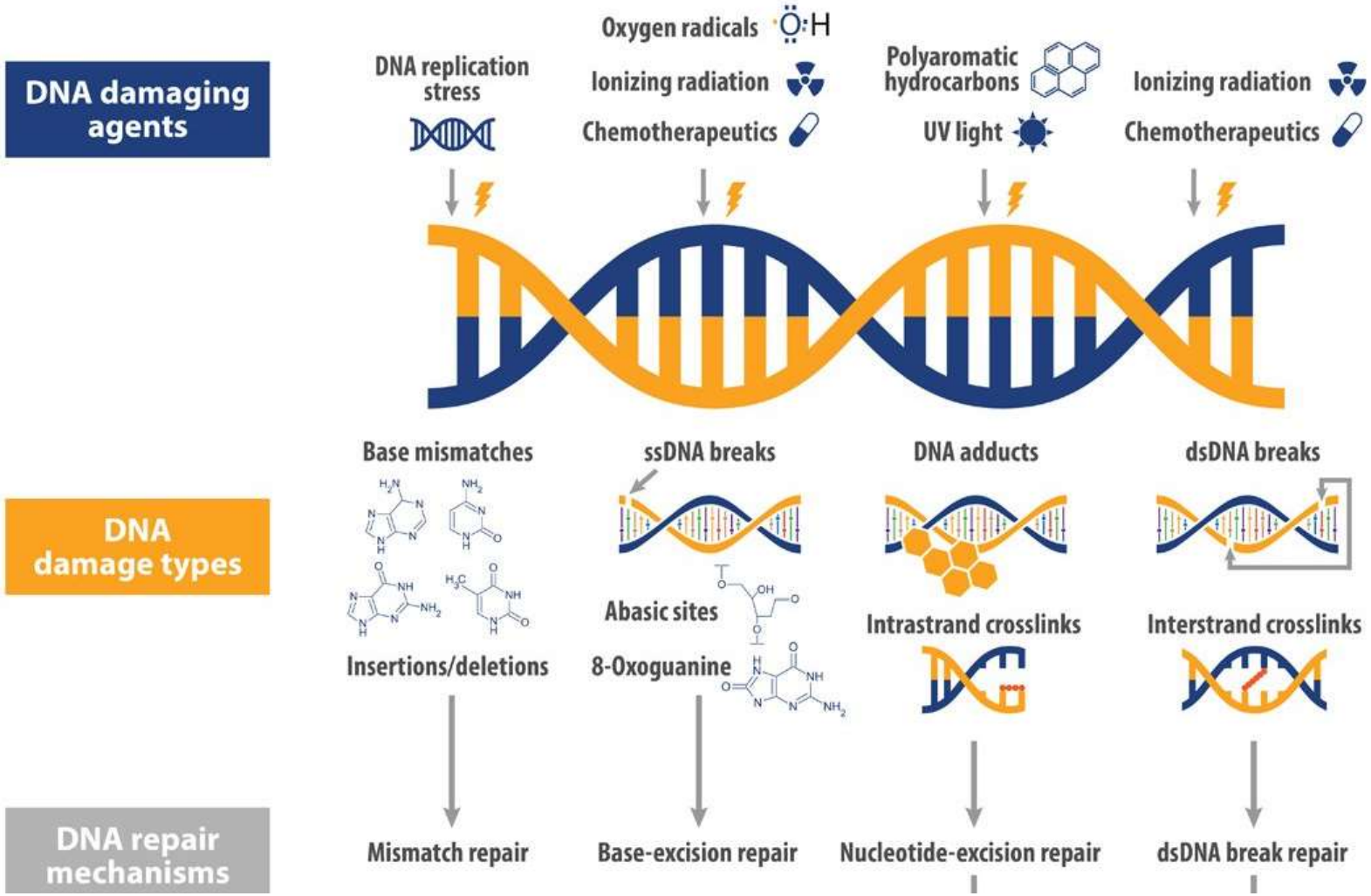
↑ probability that significant # of mutations are NOT corrected

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.



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. → if the damage is NOT repaired it often leads to a permanent mutation.



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

- 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

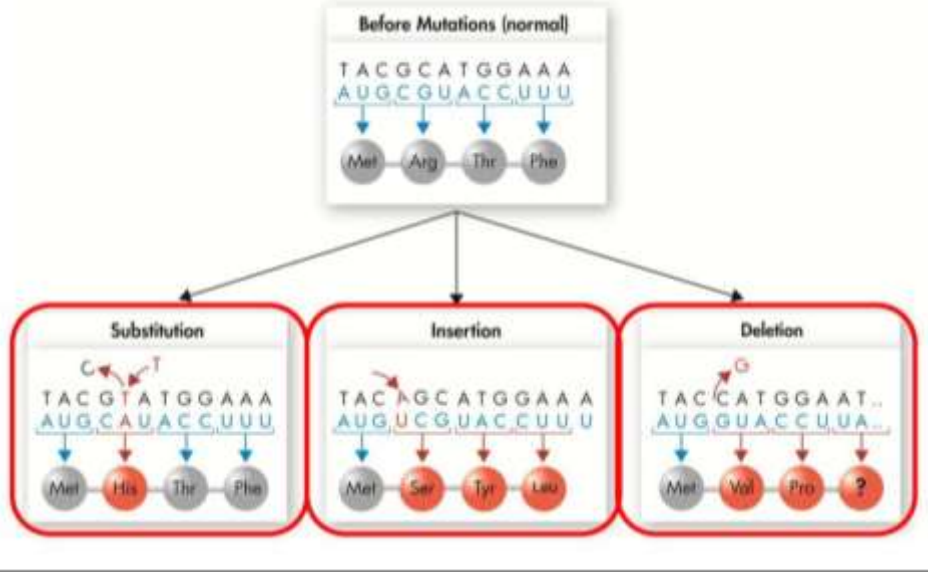
* there are more point mutations in the sperms of older males than younger males.

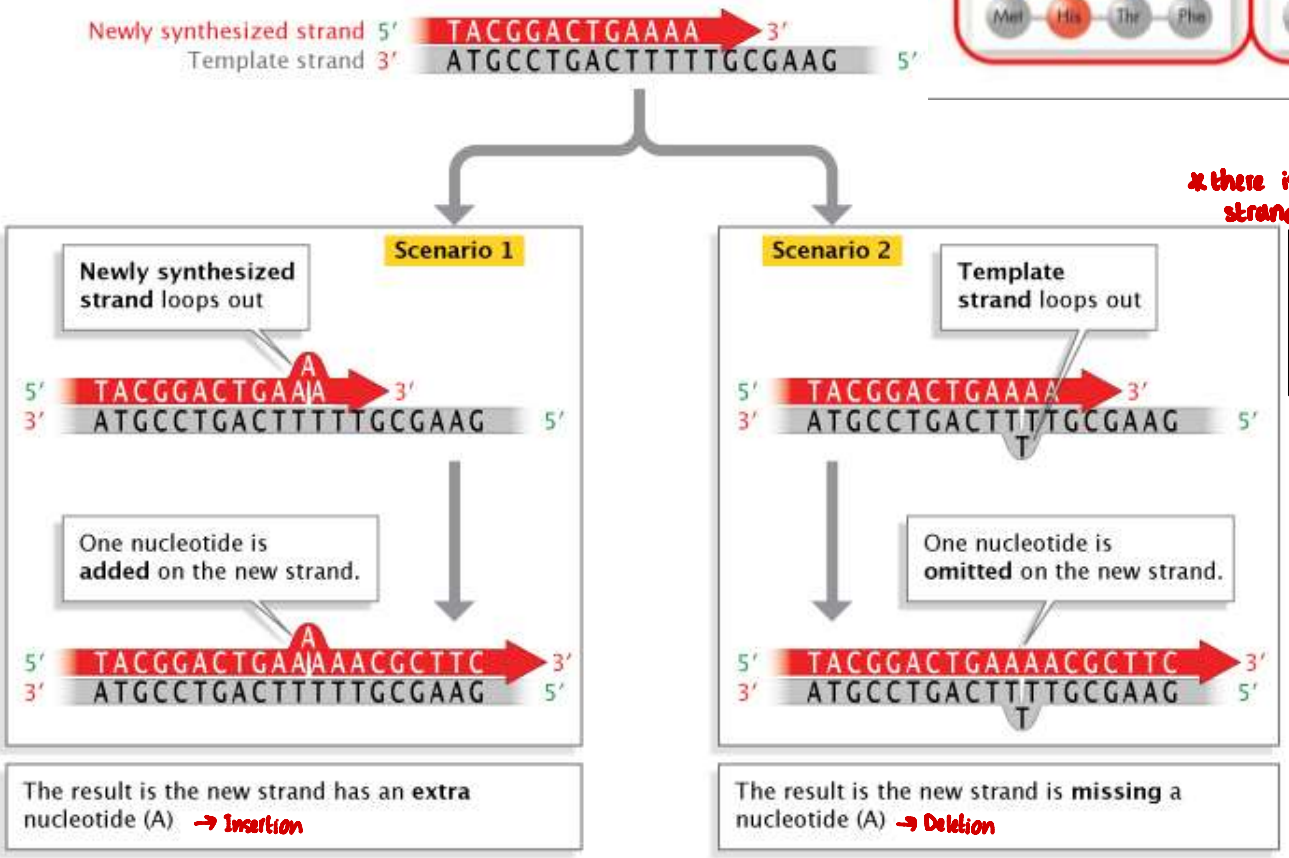
* the mitochondria arises from endosymbiosis from bacteria, it does NOT have repair mechanisms.

Gene Mutations: Point Mutations

A point mutation is a change in a single nucleotide.

There are three types of point mutations:





* there is something known as strand slippage / looping

→ if "looping out" occurs in the newly synthesized strand → introduce extra nucleotides. → Insertion.

→ if "looping out" occurs in the template strand → Deletion

(nucleotides that looped out in the template strand during DNA replication will NOT be part of the replicated sequence)

Gene and Variant nomenclature

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

related to the mature mRNA

* c.1 refers to the 1st nucleotide of the first codon which is the initiation codon ATG.

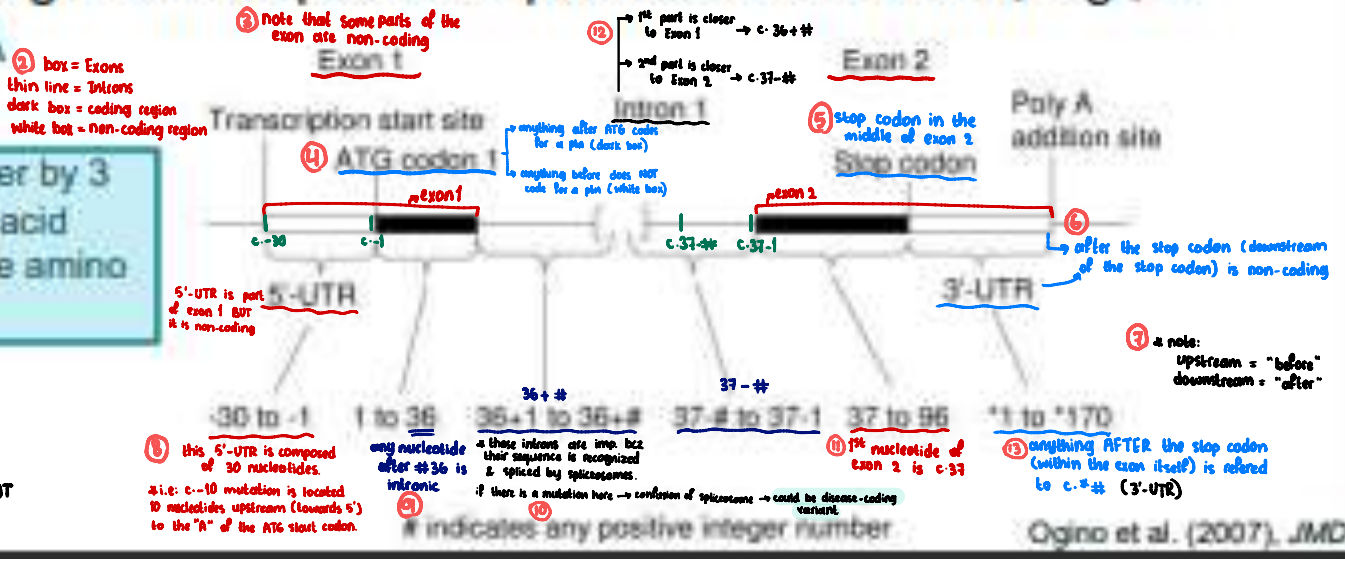
- Nucleotide 1 is the A of the ATG initiation codon (c.1, c.2, c.3)
- The nucleotide 5' of the ATG initiation codon is -1, the previous -2, etc.
- The nucleotide 3' of the stop codon is *1, the next *2, etc.
- Intronic nucleotides

"*" is used to refer to the nucleotides located AFTER the stop codon.

* c.-1 refers to the nucleotide before the 1st nucleotide in the start codon.
* c.-2 is the nucleotide located at 2 nucleotides before the "A" in the start codon.

  - **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)



Ogino et al. (2007), JMD

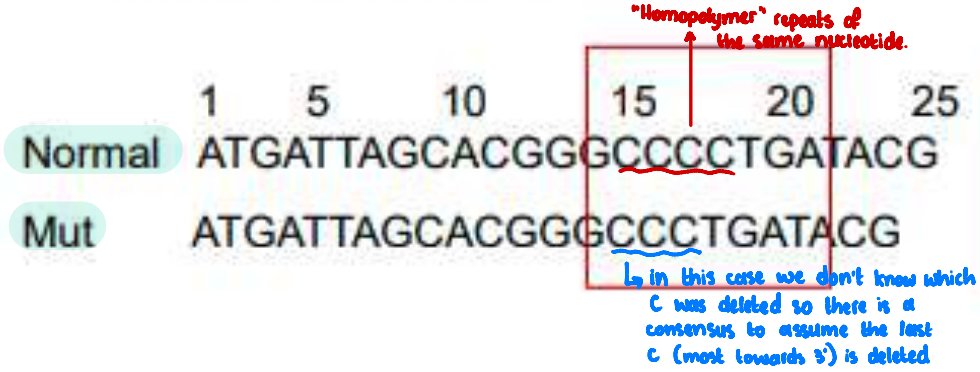
① **Exon** vs. **Intron**

Exon: a sequence of DNA that exists in the mature mRNA

Intron: sequence of DNA that is spliced out (by a spliceosome) during mRNA processing & removed. (NOT found in mature mRNA.

Symbols for specific variation types

- ">" indicates a **substitution** at DNA level: c.76A>T (from A→T)
- "_" (underscore) indicates a **range** of affected residues, separating the first and last residue affected: c.76_78delACT ⇒ it is typically expressed as: c.76_78del which means that 76, 77, & 78 are deleted.
- "dup" indicates a **duplication**: c.90_92dupACC
- "del" indicates a **deletion**: c.127delA ⇒ deletion of only 1 nucleotide. c.127del is also correct
- "ins" indicates a **insertion**: c.76_77insG ⇒ between nucleotides 76 & 77 there's a "G" insertion.
- "delins" indications a **deletion and insertion**: c.56_58delinsCATG → nucleotides 56, 57, & 58 are deleted & CATG are inserted in their place.
  - ***For all descriptions the **most 3' position** possible is arbitrarily assigned to have been changed



We cannot know which C is deleted, so assign the most 3' position

(c.18delC)

Variant nomenclature: Protein

- 3-letter amino acid code is preferred to describe the amino acid residues (Lys vs. K for lysine) → 1-letter a.a code is still used especially in molecular oncology
- 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

more preferred than "*"

Variant nomenclature: Protein

* Remember:
① an a.a could be encoded by >1 different codons
② if the DNA change still encodes for the same a.a then the variant is silent / synonymous.
③ As DNA nucleotides counting starts from 5' to 3', plns a.a$_s$ counting is from N-terminus to C-terminus.

- **Silent changes:** p.Leu54Leu or p.= → caused by a synonymous variant.
- **Substitutions:** p.Trp26Cys → a.a # 26 changed from Trp → Cys
- **Nonsense variant:** p.Trp26Ter or p.Trp26* → variant that led to a stop codon.
- **No-stop change:** p.Ter110GlnextTer17 or p.*110Glnext*17 → variant in a stop codon that changed it & caused a.a translation (Gln in this case) & the termination shifted to another 17 a.a$_s$ further until another stop codon appeared downstream. (the pln became 17 a.a$_s$ longer)
- **In-frame deletions:** p.Gln8del or p.Cys28_Met30del
  a.a Gln at position #8 is deleted & the a.a$_s$ before & after it are NOT impacted
  ↳ a.a$_s$ # 28, 29, & 30 are deleted.
- **Duplications:** p.Gly4_Gln6dup → a.a$_s$ # 4, 5, & 6 are duplicated
- **Insertions:** p.Lys2_Met3insGlnSerLys → those 3 a.a$_s$ are inserted between Lys2 & Met3
- **Frameshifts:** short description: p.Arg97fs
  ↳ due to an insertion or deletion mutation on the DNA level.
  a.a #97 Arg became Pro
  This frameshift mutation met a termination codon
  long description: p.Arg97Profs*23
  23 codons downstream to the frameshift.

* usually frameshifts lead to a premature stop codon. (early 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)

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

Dynamic mutations – amplification of repeat sequences (Fragile X, Huntington's)

* there is something known as "Copy Number Variant or CNV" which means a large deletion or duplication if it is ≥1kb then it is considered as CNV.
* if only a few bases are deleted or duplicated then it is considered "Indel" deletion or insertion.

↳ a sequence that is non-coding BUT it regulates the level of expression

↳ mutations that change in size across generations (previous lectures)