Medical Genetics

Sheet: 14 - Genetic Variation II

Writer: Leen Attar

Scientific: Mothana Al-Olimat

Final: Omar Ismail

Doctor: Bilal Al-Azab

Genetic Variation II

Frequency of mutations in human disease

1• Nucleotide substitutions Type of mutation

(Missense Nonsense - mutation that affect RNA processing such as Splice variant /mutation that influencegene expression).
- 50% of disease causing variants are ⇒ Missense
- There is a rare contribution of a variant on regulatory sequence towards a disease.

This data (on the table) is changing: at this moment of time we know that 50% for instance of nucleotide substitutions exists as missense in the coding region, yet this number can change.

Type of mutation	% disease causing mutations
Nucleotide substitutions	
Missense (amino acid substitution)	50%
Nonsense (premature termination codon)	10%
RNA processing (splice, polyadenylation, etc)	20%
Gene expression regulation (TF binding site, etc)	rare
Deletions & insertions	
Small indels	25%
Large rearrangements (deletion, duplication, inversion, etc)	5%
Insertion of LINE or Alu (interrupting regulation or coding)	rare
Repeat expansion	rare

Note: These data are changing!

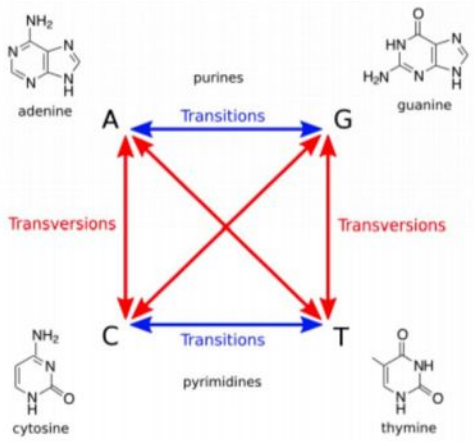
Genetics in Medicine, 8th ed, Thompson&Thompsc

Identifying a disease causing variant in a regulatory sequence such as enhancer or promoter region, is harder than identifying a variant in a coding region as we can predict the impact of this coding variant based on the change in protein's sequence or structure, we can also predict, based on the codons, if a variant leads to a missense amino acid change.
On the other hand, to identify a variant in a regulatory sequence such as the promotor region we need to interpret the change in DNA in a regulatory element not based on amino acid change rather than other factors which are harder to catch for geneticists.

2• "Indels" or "delins"

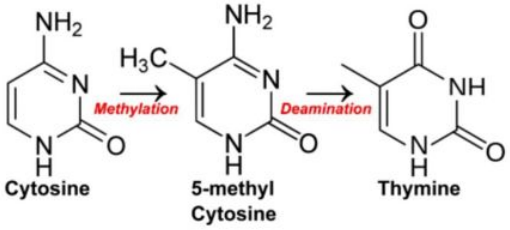
(abbreviation of deletions & insertions) such as small indels, large indels, huge insertions such as LINE or Alu repeats (Transposons), also repeat expansions like fragile X mental retardation or Huntington disease.

Point mutations



In the context of substitution point mutation:
- Transition: from purine to purine or from pyrimidine to pyrimidine (A↔G / C↔T).
- Transversion: from purine to pyrimidine or vice versa.

○ The most common variant in nucleotide substitution is (C→T) transition, due to spontaneous deamination of 5-methylcytosine to become thymine.



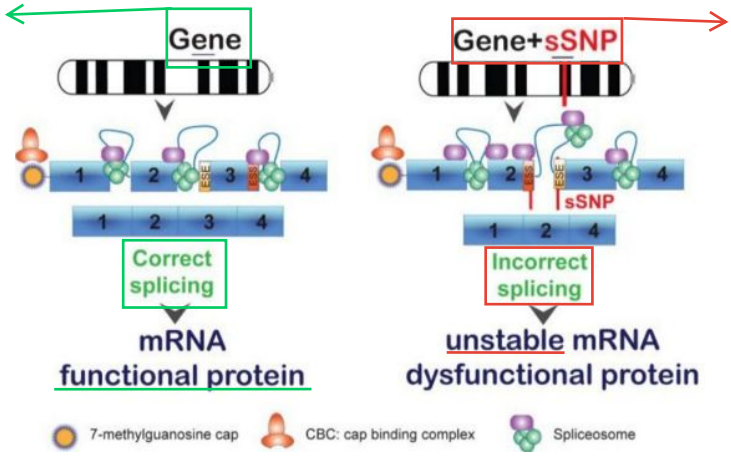
1 ▪ Silent variants (synonymous variant):

》 They don't change the amino acid sequence. E.g. (P. Ala123Ala) where the change in 3rd nucleotide of alanine codon's mRNA (C→ U) still codes for alanine.

》 Most synonymous variants are benign, but not all of them (if they impact splicing or RNA secondary structure).

Silent variants could be disease causing if they alter pre-RNA splicing.

Actual gene encoding for a certain protein with its pre–mRNA (exons & introns) and mature mRNA (exons only).
The spliceosome (ribonucleoprotein) is responsible for removing the introns from the pre–mRNA to give the mature mRNA.



The spliceosome recognises the boundaries (edges) of exons and introns based on their sequence, so if synonymous variant (sSNP) occurs for a nucleotide that happens to exist at the boundaries between the exons and introns, it will influence the pre–mRNA splicing, consequently we will have unstable mRNA that could or could not produce a protein, and if it produces a protein, it would be unfunctional/ dysfunctional.

Another situation is if it alters the secondary mRNA structure so the stability of RNA is influenced, therefore rate of translation or end product (protein itself) could be defective.

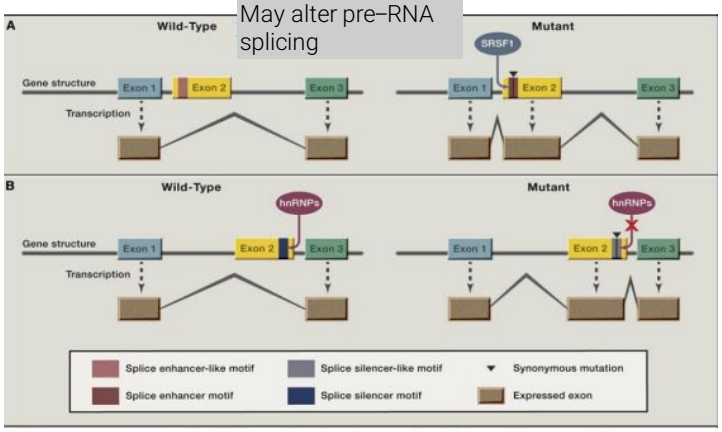
E.g., looking at the CFTR gene (responsible for cystic fibrosis), there is an isoleucine at position Ile507 which will remain Isoleucine if a substitution mutation changes the codon from (ATC→ ATT), this synonymous variant is known to be a disease causing (CF).
Although it's a synonymous variant and it's not an issue of exon– intron boundaries, **its location influences the half-life stability of mRNA thus causing the disease.**

- Notice that such finding is much challenging than finding the impact of a straight forward missense variant.

- The synonymous variant of Ile507 exists next to 508 amino acid which happens to be phenylalanine. This codon's deletion which is known as delta phenylalanine (ΔF508) is the most common disease–causing mutation in CFTR gene.

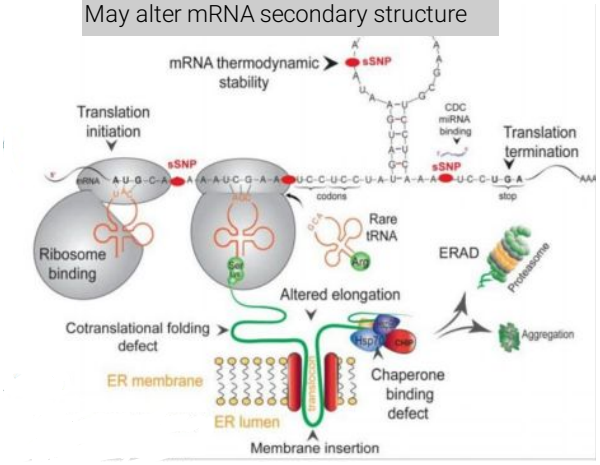
-Take home message: synonymous variants may alter translation initiation efficiency/ translation elongation rate/ ribosomal pause rhythm / cotranslational folding or the overall fate of the protein (amount of protein could be low) which causes disease.

May alter pre–RNA splicing



(A) A synonymous mutation leads to the gain of an exonic splicing enhancer motif. Consequently, binding of the splicing regulator SRSF1 is enhanced, resulting in the inclusion of an otherwise skipped exon.
(B) A synonymous mutation deactivates an exonic splicing silencer motif, thereby abolishing the binding of hnRNP splicing regulators

May alter mRNA secondary structure



2 ▪ Missense mutations (Non-synonymous variants):

Change in the amino acid due to nucleotide substitution, could be benign or pathogenic.
Two types:

● Conservative ⇒ New amino acid has similar properties to the original (polar to polar change, hydrophobic to hydrophobic, etc.)

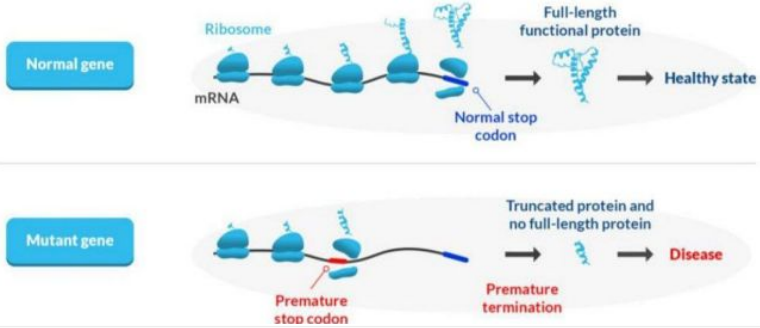
● Non–conservative ⇒ New amino acid has different properties than the original (polar to nonpolar, hydrophobic to hydrophilic, etc.)

- Example: Hemoglobin Beta–subunit is encoded by a gene known as HBB. A nucleotide substitution at the coordinate (c.17A>T) leads to non–synonymous change of glutamic acid at position 6 to valine (p.Glu6Val) which is considered a non–conservative missense mutation because the properties of Glu (acidic, -ve) differ from Val (hydrophobic) thus causing β–globin protein aggregates ⇒ Sickle cell anemia.

	1	2	3	4	5	6	7	8	9
	Val	His	Leu	Thr	Pro	Glu	Glu	Lys	Ser
NORMAL	GTG	CAT	CTG	ACT	CCT	GAG	GAG	AAG	TCT
SICKLE	GTG	CAT	CTG	ACT	CCT	GTG	GAG	AAG	TCT
	Val	His	Leu	Thr	Pro	Val	Glu	Lys	Ser

3 ▪ Non-sense variants:

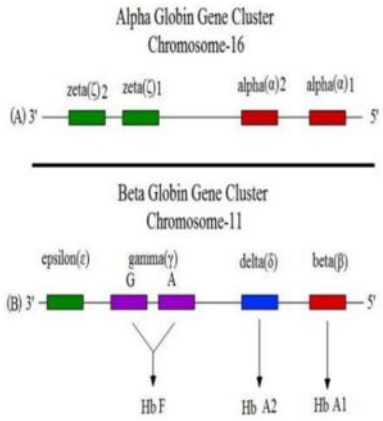
Change of a codon into a termination codon (stop signals: UAA, UAG, UGA), which causes errors in translation and may result in nonsense mediated decay (NMD), truncated protein, or splicing impact.
- Nonsense variant are usually more deleterious then non–synonymous & synonymous variants yet not all of the nonsense variant are pathogenic.



▪ Normal→ This is the actual stop codon for normal gene.

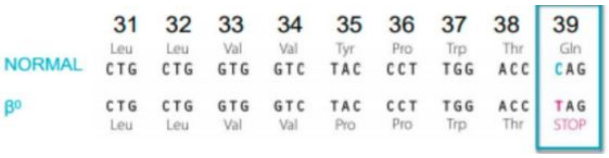
▪ Mutant→ Premature stop codon, therefore translation termination occurs at this early site leading to multiple implications (truncate protein, nonsense mediated decay).

Blood Disease Thalassemia



》 Type 1 (alpha thalassemia): Mutant Alpha (α) globin genes.
》 Type 2 (Beta thalassemia): Mutant Beta (β) globin genes.

○ Beta thalassemia→ In the codon CAG encoding for the amino acid 39 (glutamine), if we substitute C by T, it becomes TAG which means on RNA level UAG which is a stop signal and this variant creates a premature termination codon.

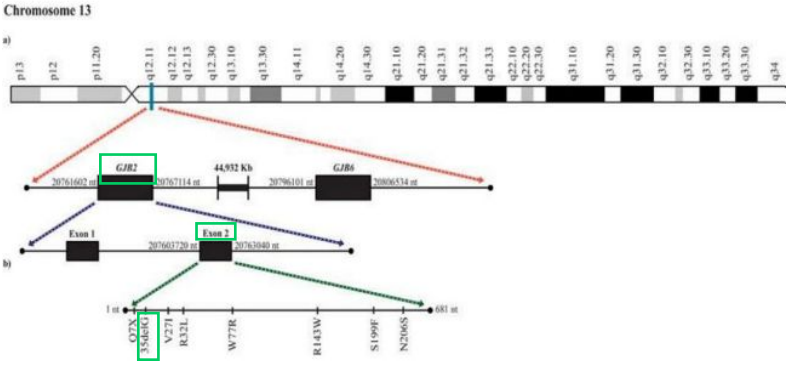


- Heterozygous variants have thalassemia minor.
- Homozygotes have no β–globin protein thus they develop β–thalassemia major.
  ↳ So, nonsense variants could either lead to end–product or truncated protein (protein was supposed to exist after premature termination signal does not exist anymore) or a phenomenon known as nonsense mediated decay.

4 ▪ Frameshift variant:

Change in the mRNA reading frame which causes error in translation. Often leads to a premature termination codon downstream the variant itself (frameshift causes nonsense).
-Although more deleterious than synonymous and non–synonymous, it's not always pathogenic.

Example: GJB2 is a gene responsible for bilateral prelingual sensorineural hearing loss {prelingual=before the baby start to speak}, a mutation at which is mostly inherited in an autosomal recessive fashion.
↳ The most common mutation of GJB2 is deletion of G–nucleotide on the position of c.35 (GJB2 c.35delG), this leads to (p.Gly12fs) which means that a frameshift started at the amino acid no.12 (Glycine) and anything after the glycine also changes. If you compare normal sequence with 35 del G, you will see that glycine no.12 changed to valine and a frameshift due to this variant leads to truncation signal or premature stop codon.

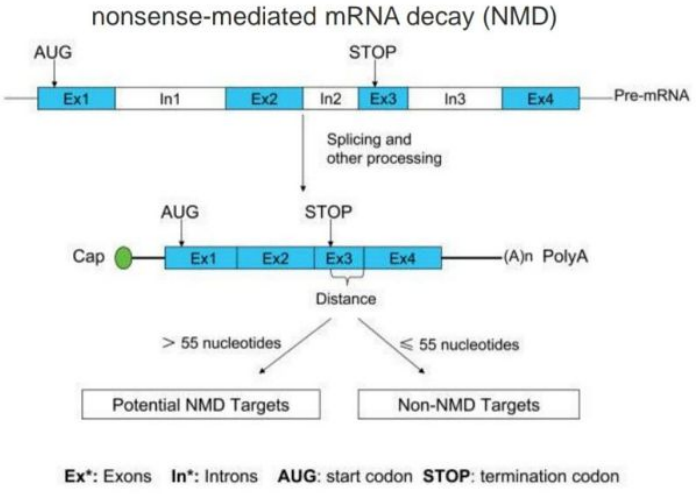


This figure shows a certain locus on chromosome 13 (q12.11) where you can see GJB2.
Zooming on GJB2 you can see two exons, 35 del G is found at the beginning of exon 2, amino acids after this deletion will be lost (due to the premature stop codon).

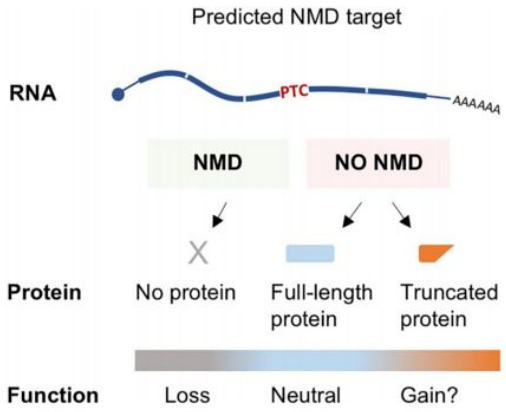
Nonsense mediated decay

This is a pre–mRNA with its exons & introns.
If a stop signal appears by a nonsense variant in one of the exons (exon 3 in this example), that mRNA is destined to one of two fates.

○ The first is that the mRNA upstream of the stop signal will be translated and the mRNA downstream won't be translated which yields a truncated protein in the cell (incomplete protein) which could be unfunctional or even a harmful protein.



○ The other is that the cell recognizes the mRNA carrying the premature termination and degrades it before being translated (Non–sense mediated decay).
↳ What determines which one of these two fates will take place is how early the premature stop codon is in that exon.



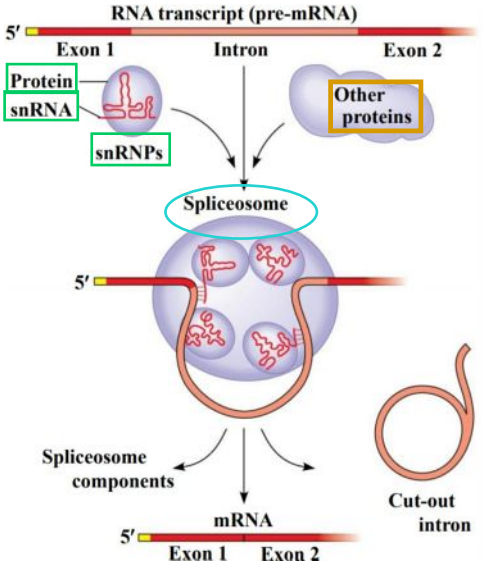
Rule of thumb 👍
- If the number of nucleotides between premature termination signal and 3' prime end of exon (3' exon–intron junction) is 55 or less that means the mRNA will not go NMD rather it will be translated. If the number is more than 55 the mRNA will go NMD.

- If NMD occurred there will be no protein as the mRNA didn't get translated, but if NMD didn't occur there will be a truncated protein which either lost its function or gained a negative effect known as dominant negative effect.

5 ▪ In-frame deletions and insertions:

Indels that occur to nucleotides in multiples of three (3,6,9,12, ...) which leads to deletions and insertions of amino acids without altering the reading frame.
- It may be benign or pathogenic.
- Example: CFTR phenylalanine deletion (amino acid no.508) → c. (p.Phe508del / ΔF508) which is caused by a deletion of the whole codon (TTT) or (UUU) on the mRNA level, and that causes a block in protein processing leading to cystic fibrosis.

Normal	ATC	ATC	TTT	GGT	GTT
	Ile	Ile	Phe	Gly	Val
ΔF508	ATC	ATC	GGT	GTT	
	Ile	Ile	Gly	Val	

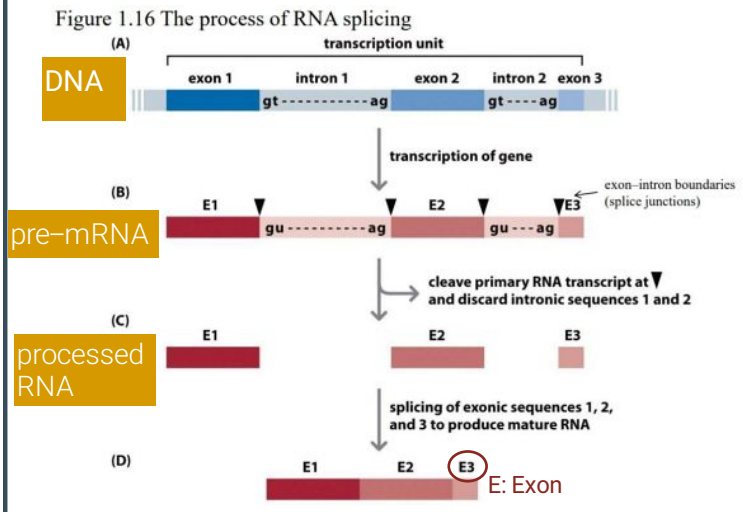


Exons and introns in the pre–mRNA are processed by colossal (large) enzymes known as spliceosomes.

This enzyme is complex (made of multiple subunits) ⇒ small nuclear RNA (snRNA) mixed with proteins are together known as (snRNPs) along with other proteins compromise the actual enzyme that catalyses removal of introns.

↳ This enzyme recognises the boundaries between the exons and the introns.

Now if you zoom on the sequence of those boundaries

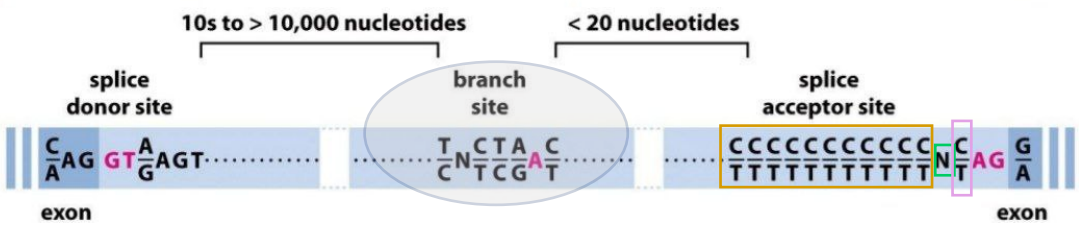


Inside the intron itself at its boundaries there is something known as exon–intron boundaries or splice junctions. Splice donor site is at the 5' end of the intron (GT in DNA, GU in mRNA).
Splice acceptor site is at the 3' end of the intron (AG).

↳ The spliceosome recognises these splice junctions and removes them during the processing thus removing the introns.

There is a consensus sequence in **splice donor site** that is bigger than GT but GT itself is invariant which means that in almost all genes you'll find GT in the splice donor site.
The sequence surrounding GT is the most commonly found in splice donor sites with A/G meaning that we can either find A or G at that site as they are equally common.

Moving on to the **splice acceptor site**, within the intron AG is almost invariant, then the remaining part of it could be a polymer of Cs or a polymer Ts, N means any nucleotide can be at that site, then could be C or T (C/T).
⁎ Notice that part of the splicing junctions is in the exons and part of it is in the introns.



Now if you move upstream of the later exon (backward within the intron) there is another region that plays a critical role in splicing known as branch site (A).

↳ The branch site, which includes an almost invariant A, exists around less than 20 nucleotides upstream from the splice acceptor site.

Summary

◯ Most introns in eukaryotic genes contain conserved sequences that correspond to three functionally important regions:
•Two of the regions, the splice donor site and the splice acceptor site, span the 5' and 3' boundaries of the intron.
•The branch site is an additional important region that typically occurs less than 20 nts upstream of the splice acceptor site.

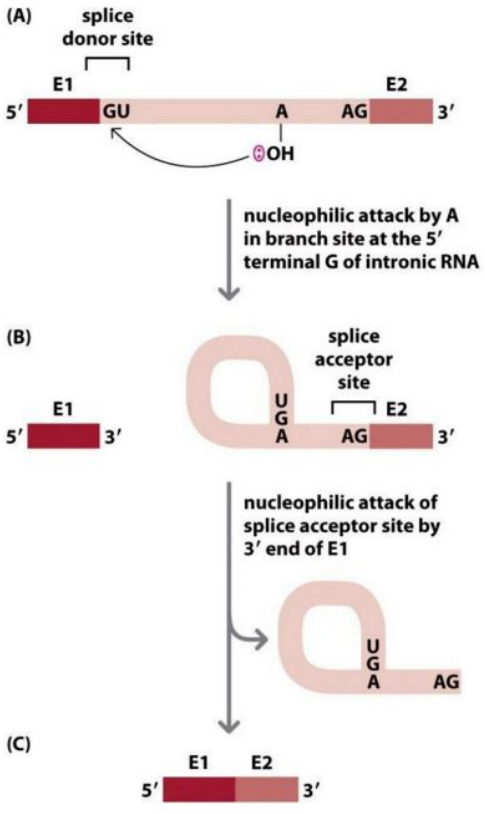
- The nucleotides shown in pink in these three consensus sequences are almost invariant.
The other nucleotides detailed in both the intron and the exons are those most commonly found at each position.

In some instances, two nucleotides may be equally common, as in the case of C and T near the 3' end of the intron.
Where N appears, any of the four nucleotides may occur.

The mechanism of RNA splicing

》 On the chemical level, the 2' OH–group of the invariant Adenine within the branch site has pair of electrons that perform a nucleophilic attack on the G of the invariant GU at the splice donor site (5' of the intron) resulting in a covalent bond between them which will break the phospho–diester bond between guanine and nucleotide before it (the 3' end of exon 1) ⇒ this will release the 5' end of the intron forming a shape called lariat structure.

》 Then, the 3' OH of the 3' nucleotide in exon 1 will perform a nucleophilic attack on the first nucleotide at the 5' end of the following exon ⇒ this will release lariat structure & the intron will be lost & the two exons will be attached together through a covalent phosphodiester bond.



(A)The unprocessed primary RNA transcript with intronic RNA separating sequences E1 and E2 that correspond to exons in DNA.

(B) The splicing mechanism involves a nucleophilic attack on the G of the 5' GU dinucleotide. This is carried out by the 2' OH group on the conserved A of the branch site and results in the formation of a lariat structure and cleavage of the splice donor site.

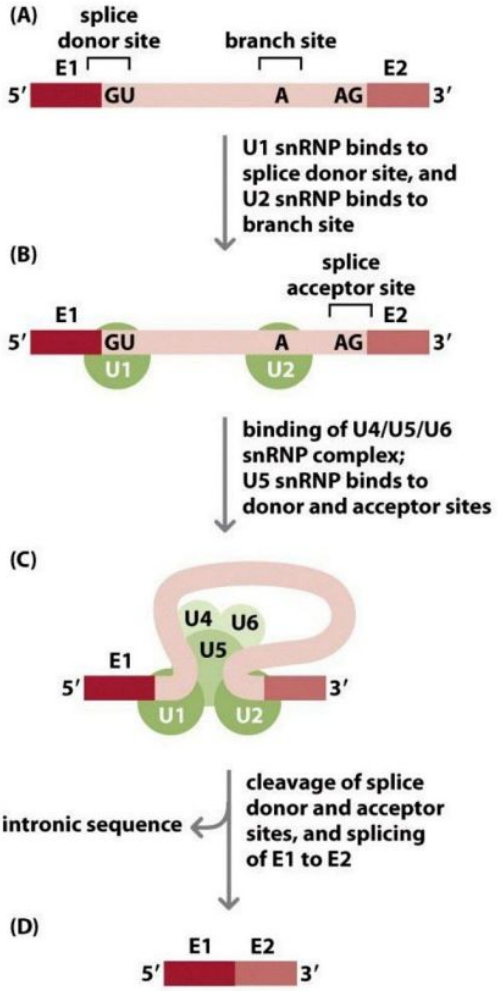
(C) The 3' OH at the 3' end of the E1 sequence performs a nucleophilic attack on the splice acceptor site, causing release of the intronic RNA (as a lariat- shaped structure) and fusion (splicing) of E1 and E2.

◇ **Regarding the role of spliceosome (large ribonucleoprotein with multiple subunits)**

↳ **one subunit is U1 (snRNP) which includes snRNA complementary in sequence to the splice donor site so it will recognise it and bind to it.**

↳ **Another subunit is U2 which will recognise the branch site and will bind to it similarly (by RNA- RNA base pairing).**

☆ U1 & U2 then will recruit other subunits of the spliceosome (U4, U5, U6) which stabilize the interaction between splice donor and splice acceptor site as U5 snRNP binds to both simultaneously. This catalysis the nucleophilic attack of the branch site on the splice donor site followed by catalyses of the nucleophilic attack of the 3' end of exon 1 on the 5' end of the following exon.



A) The unprocessed primary RNA transcript.

B) Within the spliceosome, part of the U1 snRNA is complementary in sequence to the splice donor site consensus sequence. As a result, the U1 snRNA-protein complex (U1 snRNP) binds to the splice junction by RNA–RNA base pairing. The U2 snRNP complex similarly binds to the branch site by RNA–RNA base pairing.

C) Interaction between the splice donor and splice acceptor sites is stabilized by the binding of a multi-snRNP particle that contains the U4, U5, and U6 snRNAs.
- The U5 snRNP binds simultaneously to both the splice donor and splice acceptor sites.
- Their cleavage releases the intronic sequence.

D) E1 and E2 to be spliced together.

6 ▪ Splice site variants:

A genetic alteration in the DNA sequence that occurs at the boundary of an exon and an intron (splice site). This change can disrupt RNA splicing resulting in the loss of exons or the inclusion of introns and an altered protein-coding sequence.

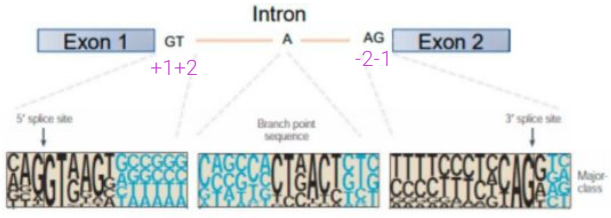
Impact the interaction of the spliceosome subunits with their target sequence which is the consensus splice donor or splice acceptor sequence and that leads to exon skipping (the exon may or may not be included in the mature mRNA).

Splice site variants most likely affect these ±1,2 nucleotides of the exon–intron boundary (+1 nucleotide is G & +2 nucleotide is T, both of which are at the splice donor site. -1 & -2 nucleotides are G & A respectively and both are at the splice acceptor site).

Other positions in the splice consensus sequence that may be affected by a splice site variant are:

》 ±15 of the boundary, the branch site sequence (20 nucleotides upstream of splice acceptor site)

》 And variants affecting the first and last three nucleotides of an exon.



7 ▪ Regulatory variants:

Occur in regulatory locations that do not encode for protein but they regulate the level of expression for a specific gene such as the promoter, enhancer, silencer or UTRs (untranslated regions – part of an exon which isn't encoding for a protein).
These will result in an altered level of protein expression.

Example: HBB (hemoglobin beta gene) in which a variant in the promoter of β–globin gene leads to decreased expression of β–globin.
(c.–101C>T) means that there is a nucleotide substitution C to T 101 nucleotides upstream the first exon.

Note: compound heterozygotes with a severe mutation have mild β–thalassemia mostly when sequencing the gene searching for mutations, we sequence the exons only, so any variants in the introns or promoters can go undetected.
(pic) shows the exons of the gene, mRNA and protein.