Eukaryotic Gene Structure and Its Role in Genetic Disease

William H. Majoros, Ph.D.

Assistant Professor

Center for Statistical Genetics and Genomics

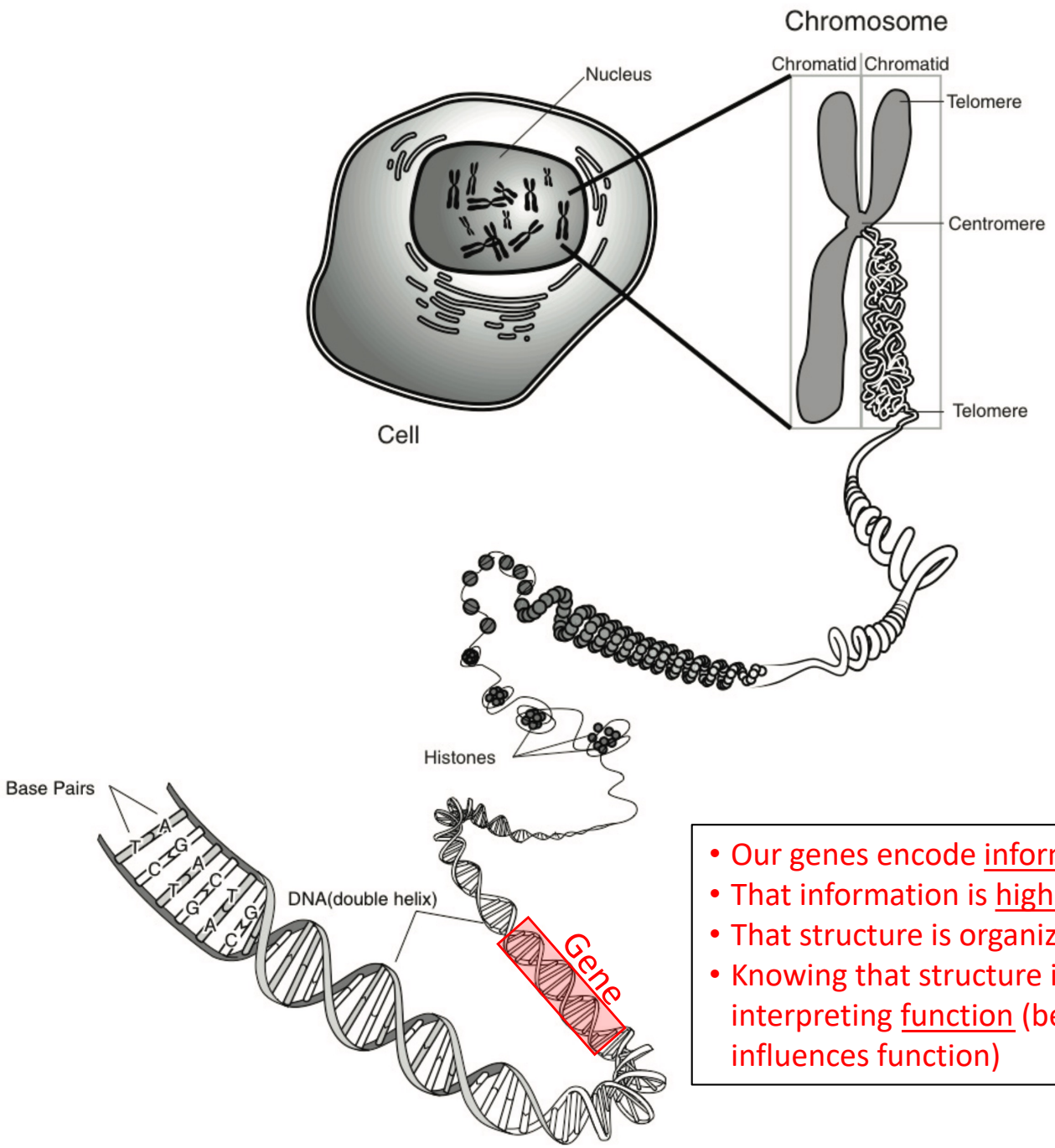
Center for Genomic and Computational Biology

Center for Advanced Genomic Technologies

Duke University School of Medicine

bmajoros@duke.edu

The Eukaryotic Cell

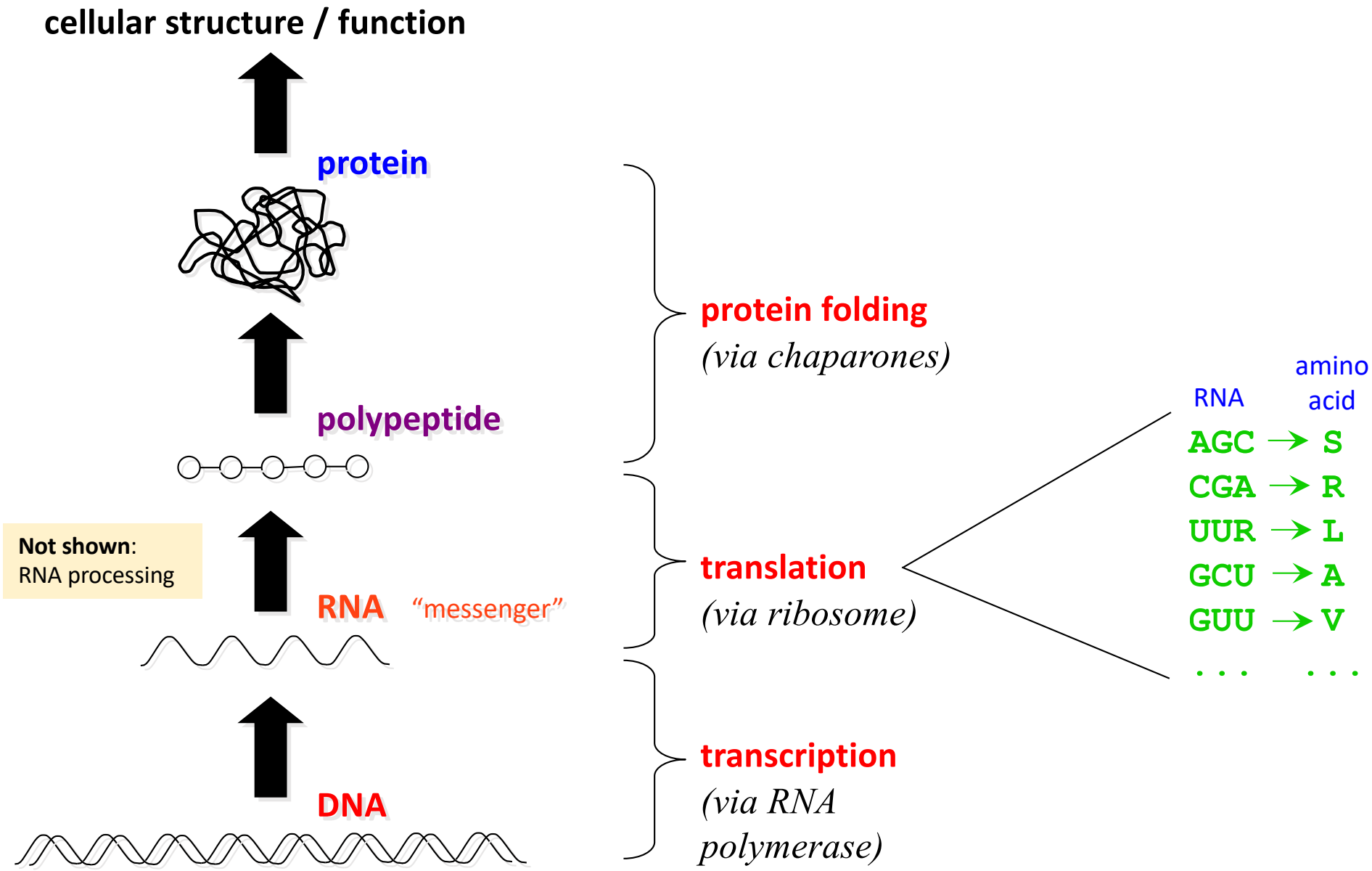


Sister chromatids after replication

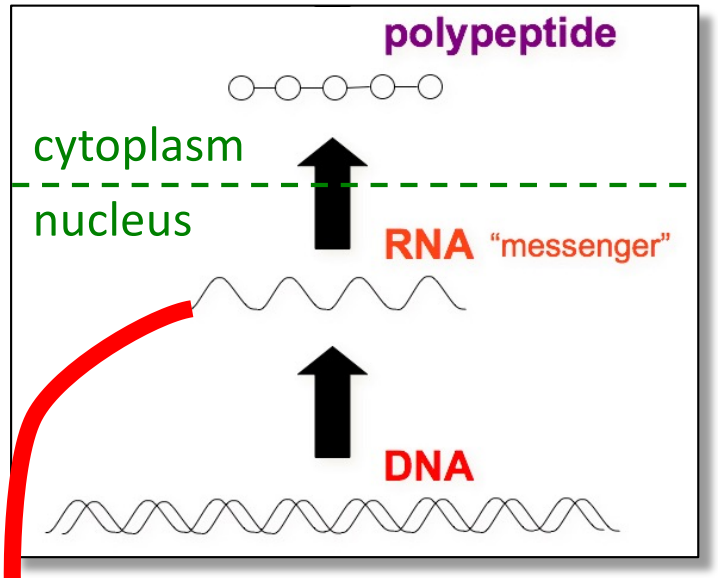
- Our genes encode information.
- That information is highly structured.
- That structure is organized hierarchically.
- Knowing that structure is important for interpreting function (because structure influences function)

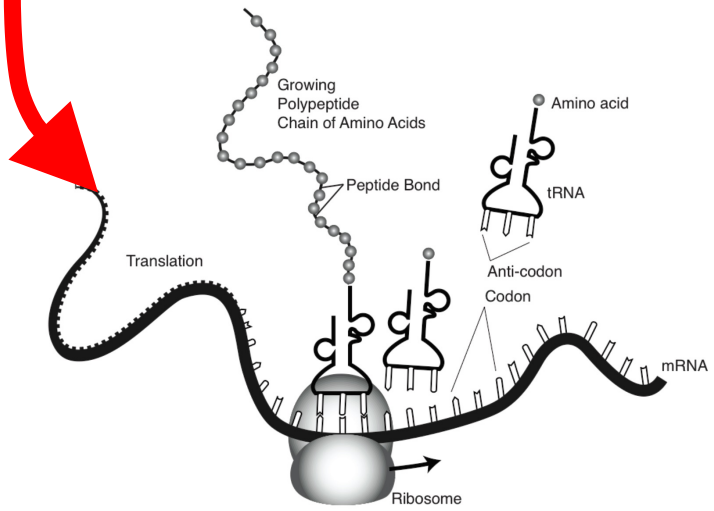
NHGRI

The Central Dogma

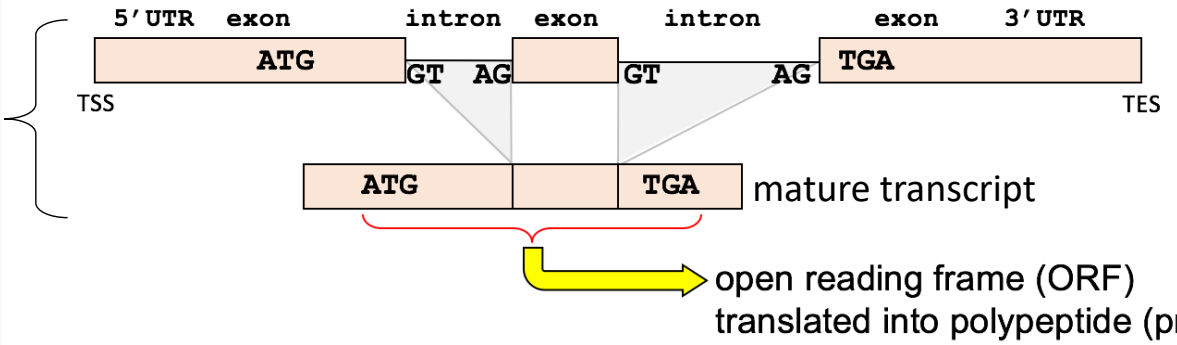


Gene Structure





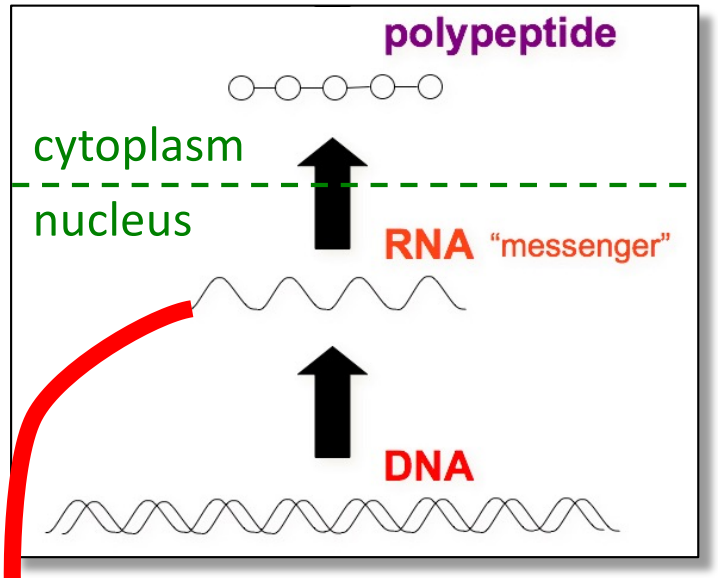
95% of human genes contain introns.

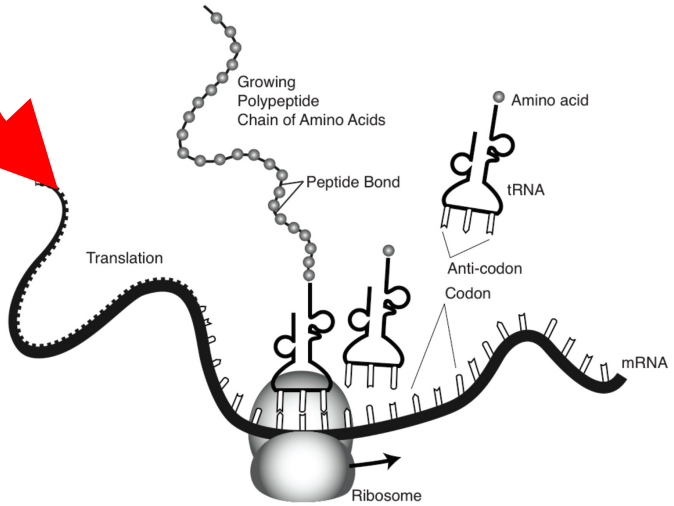


The way that a gene is spliced can impact how it is later translated.

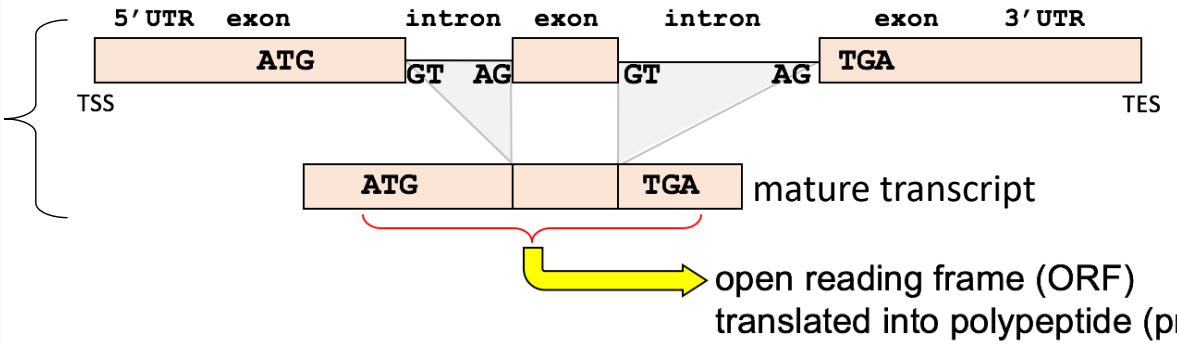
- Failing to remove an intron within the ORF would result in additional nucleotides being translated into amino acids
- Changes to individual splice sites can result in frameshifts

Gene Structure





95% of human genes contain introns.



The way that a gene is spliced can impact how it is later translated.

Translation reading frames:

phase 0:

```
ATG GAC CAC CCA ATT GTG GTT GAG CAG CCA GAT GCC TGG ACA GAG GAC AAT GGC TTC TGA
met asp his pro ile val val glu gln pro asp ala trp thr glu asp asn gly phe ***
```

phase 2:

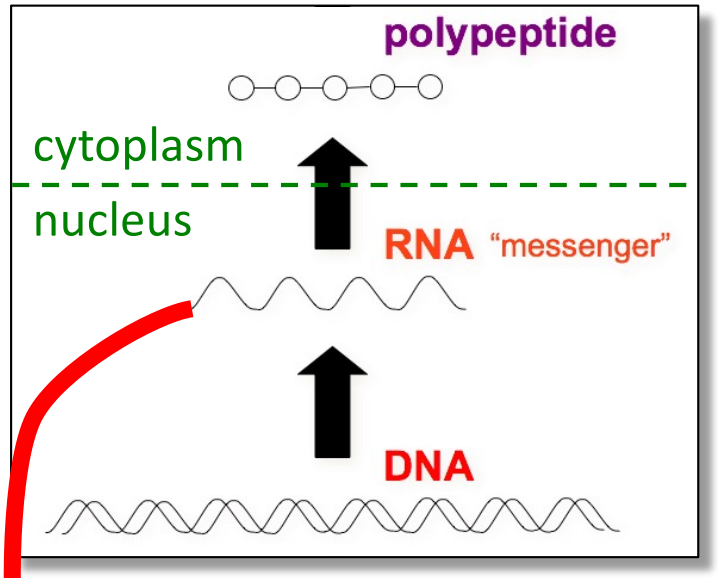
```
A TGG ACC ACC CAA TTG TGG TTG AGC AGC CAG ATG CCT GGA CAG AGG ACA ATG GCT TCT GA
  trp thr thr gln leu trp leu ser ser gln met pro gly gln arg thr met ala ser => no stop
```

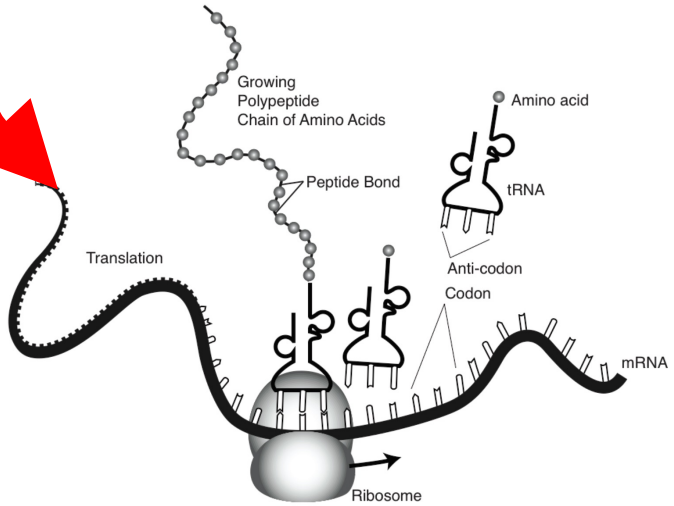
phase 1:

```
AT GGA CCA CCC AAT TGT GGT TGA GCA GCC AGA TGC CTG GAC AGA GGA CAA TGG CTT CCA TG
   gly pro pro asn cys gly ***
```

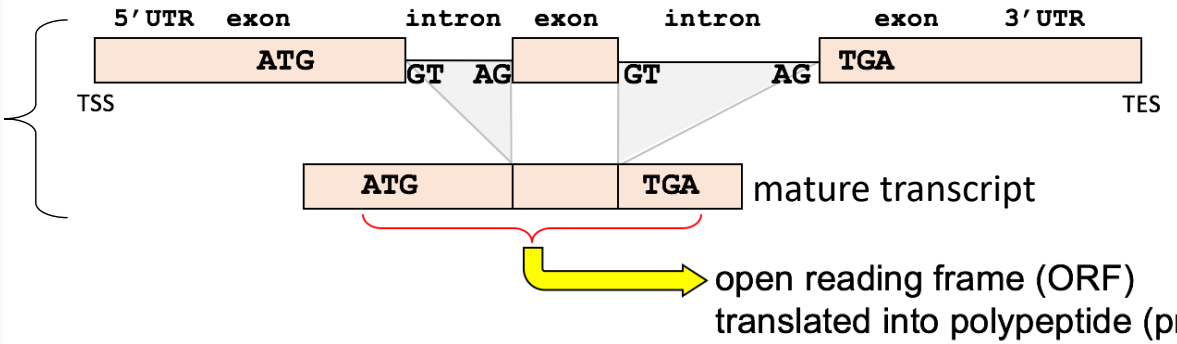
Changes in splicing can change the reading frame, and that can be highly disruptive.

Gene Structure





95% of human genes contain introns.



The way that a gene is spliced can impact how it is later translated.

Translation reading frames:

phase 0:

```
ATG GAC CAC CCA ATT GTG GTT GAG CAG CCA GAT GCC TGG ACA GAG GAC AAT GGC TTC TGA
met asp his pro ile val val glu gln pro asp ala trp thr glu asp asn gly phe ***
```

phase 2:

```
A TGG ACC ACC CAA TTG TGG TTG AGC AGC CAG ATG CCT GGA CAG AGG ACA ATG GCT TCT GA
  trp thr thr gln leu trp leu ser ser gln met pro gly gln arg thr met ala ser => no stop
```

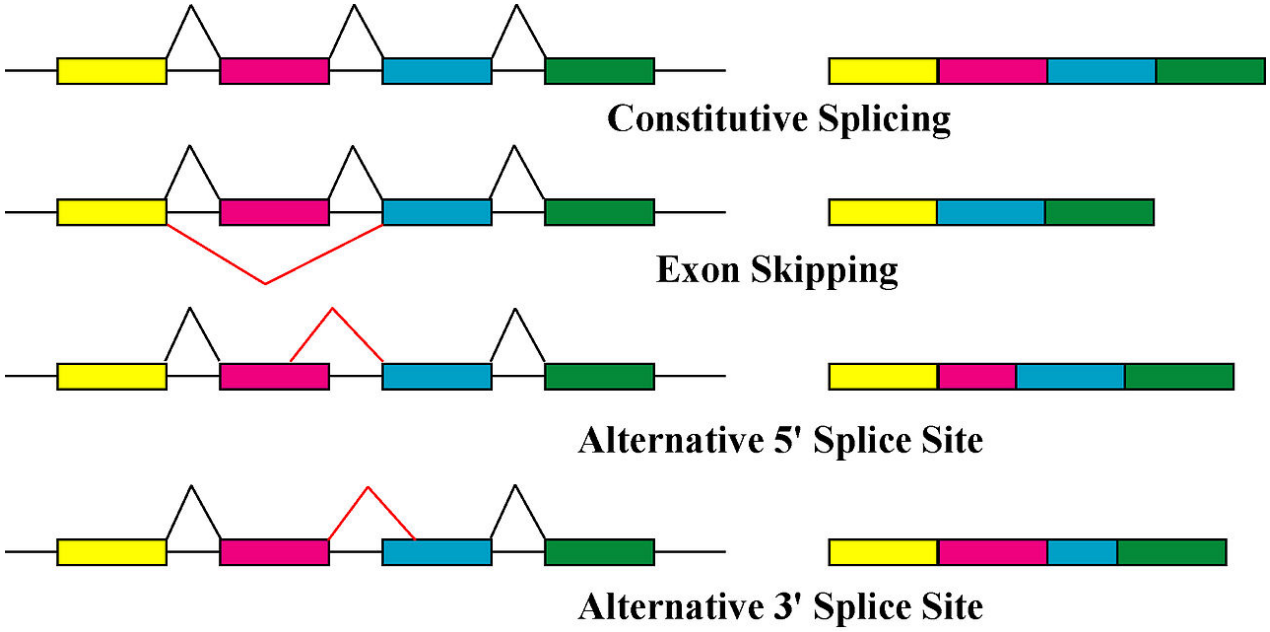
phase 1:

```
AT GGA CCA CCC AAT TGT GGT TGA GCA GCC AGA TGC CTG GAC AGA GGA CAA TGG CTT CCA TG
   gly pro pro asn cys gly ***
```

In order to know what protein is produced by a gene, we need to know the exact <u>splicing pattern</u> + <u>reading frame</u> = <u>gene structure</u>.

Splice Isoforms

95% of human genes are spliced, and 95% of those have multiple isoforms.

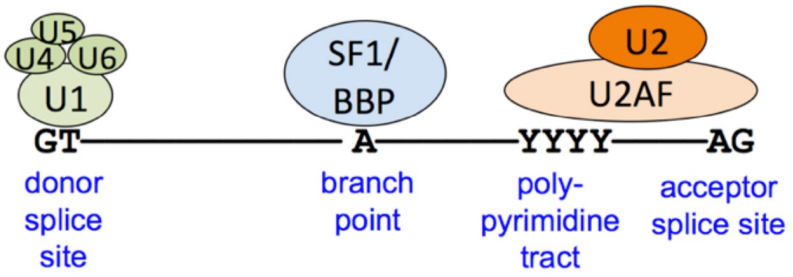


There is known to be stochasticity in the production of isoforms (including "spurious" isoforms that appear to have no function)

https://en.wikipedia.org/wiki/Protein_isoform

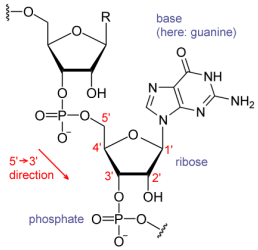
Splicing is a Complex & Highly Regulated Process

An intron:

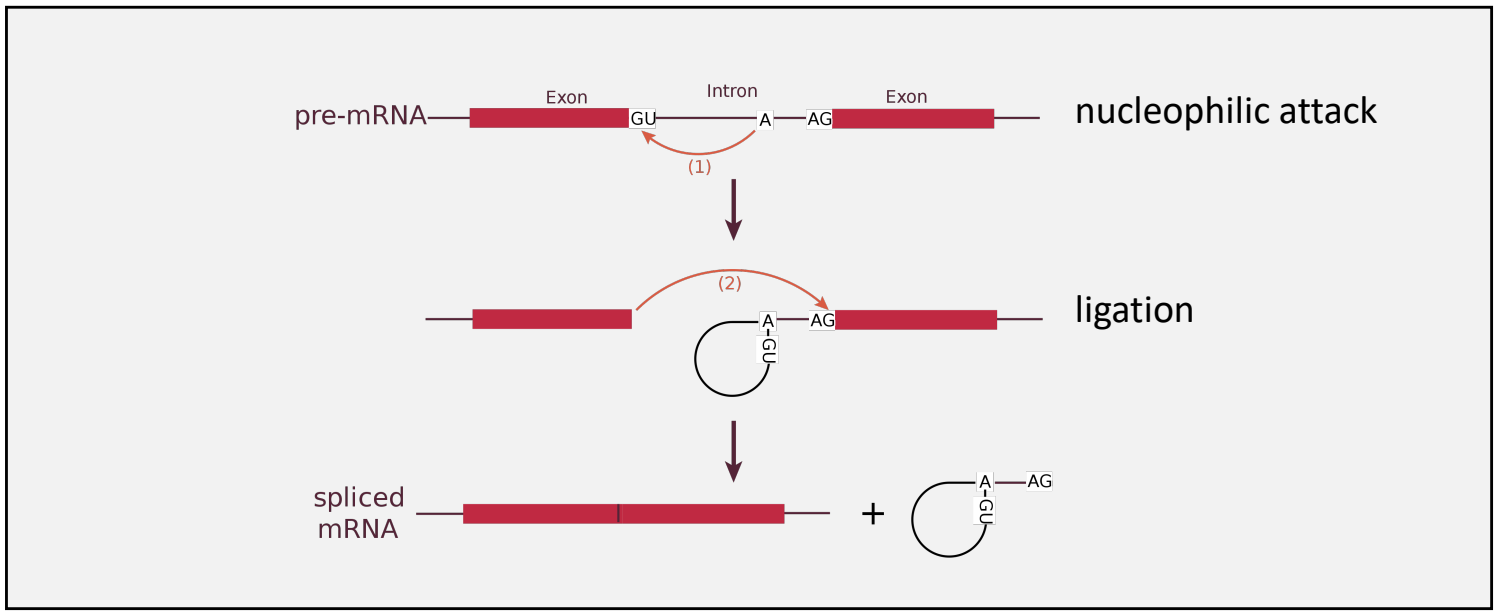


purines: A,G (R)
pyrimidines: C, T (Y)

U1 – U6 snRPs (small nuclear ribonucleoprotein)
U2AF (U2 auxiliary factor)
SF1 = splicing factor 1
BBP = branch-binding protein

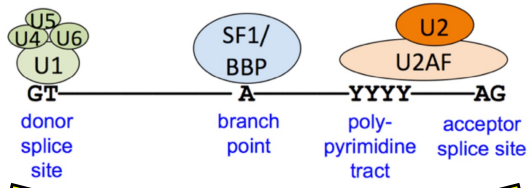


RNA nucleotides have an additional reactive 2' OH that facilitates this!



https://en.wikipedia.org/wiki/RNA_splicing#/media/File:RNA_splicing_reaction.svg

Splicing is a Complex & Highly Regulated Process



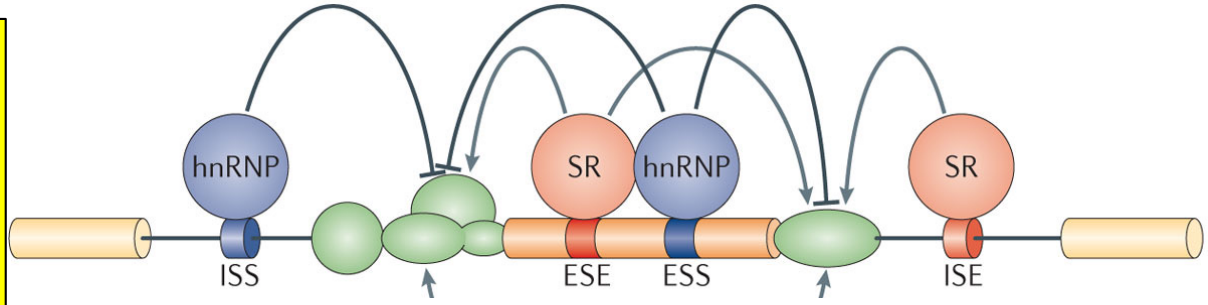
Donor site: usually **GT** (major spliceosome), but sometimes AT (minor spliceosome)

donor splice site

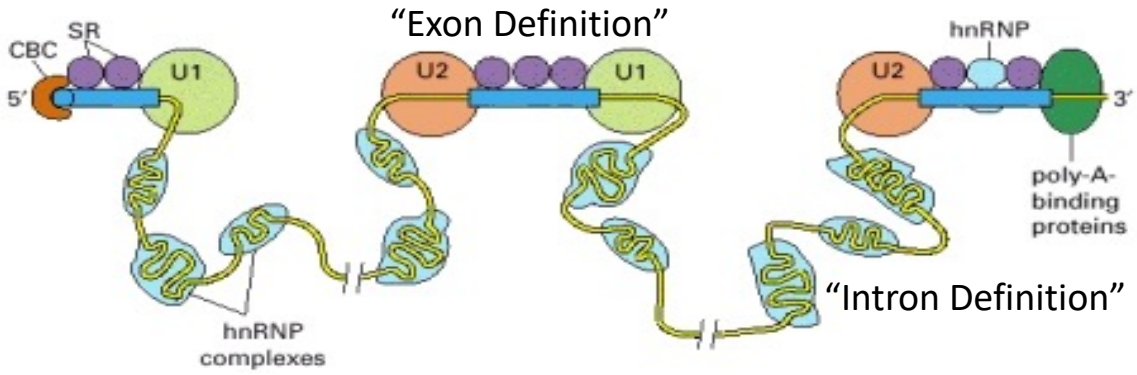
acceptor splice site

Acceptor site: usually **AG** (major spliceosome), but sometimes AC (minor spliceosome)

The splicing code and the protein code reside in the same space, and jointly influence the selective landscape for gene sequences.



Kornblihtt *et al.*, 2013

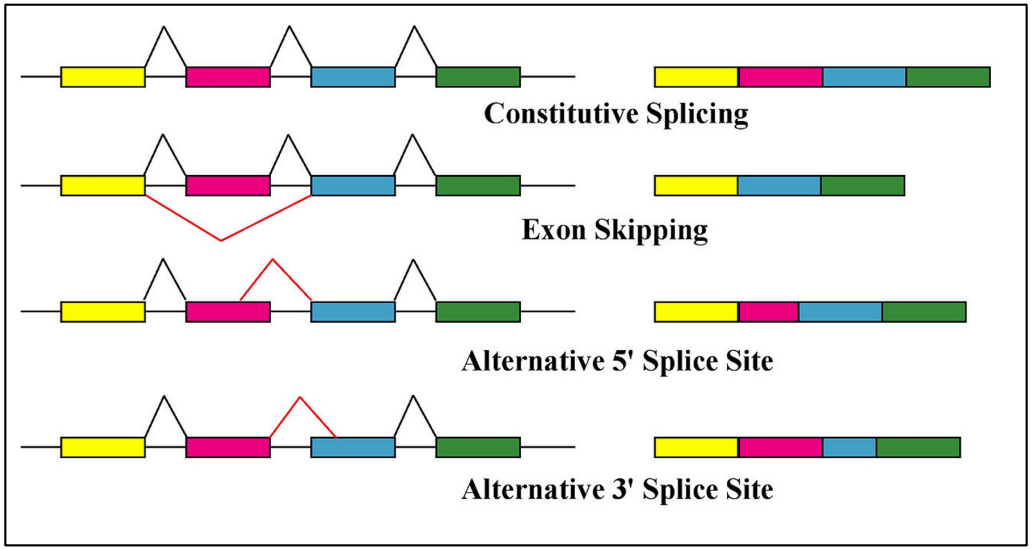


Alberts *et al.*, 2002

Splice sites alone are not sufficient to determine splicing patterns!

Splicing Regulation Can be Cell-type Specific

Because some splice isoforms are specific to individual cell types or conditions, splicing must be differentially regulated.

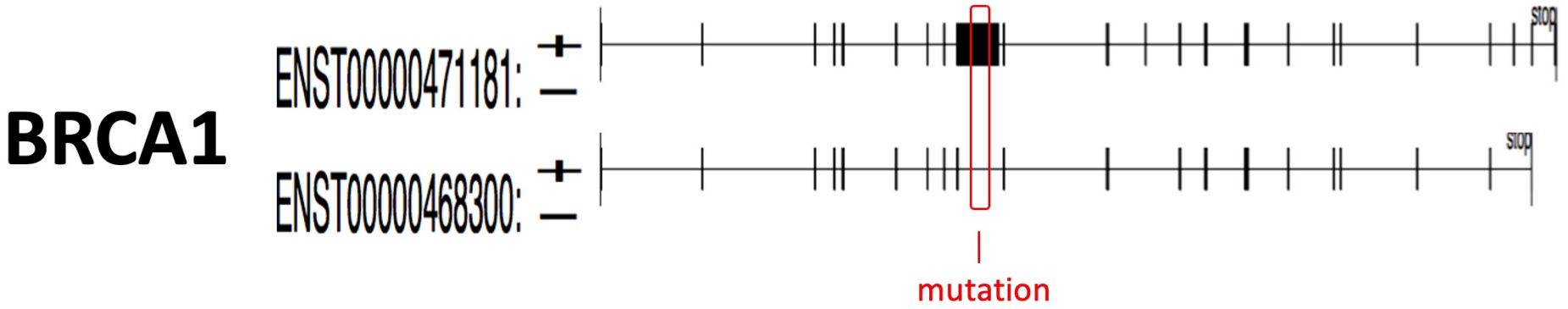


Variants ⇔ Gene Structure

1. Gene structure influences how we interpret the function of genetic variants
2. Variants can alter gene structure

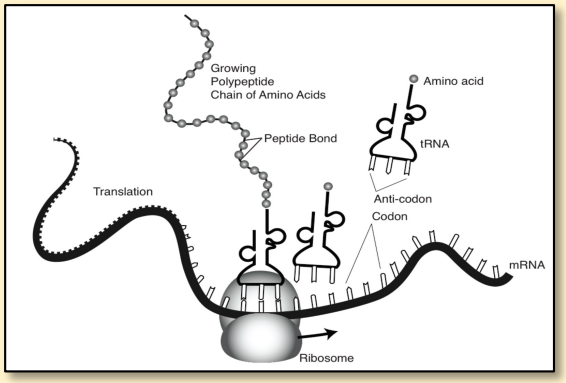
Interpreting Variants in Genes

Gene structure influences variant interpretation:



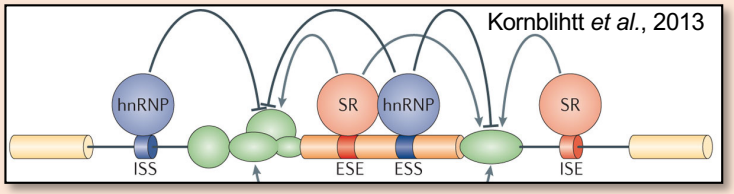
Variants in exons can:

- Alter amino acids
- Modify protein domains or signal peptides
- Modify the reading frame
- Alter splicing
- etc.



Variants in introns can:

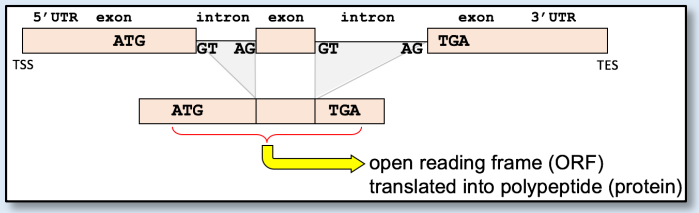
- Alter splicing via splicing regulation
- Impact gene regulation (intronic enhancers)
- Impact a miRNA gene within the intron



Variants Can Alter Gene Structure

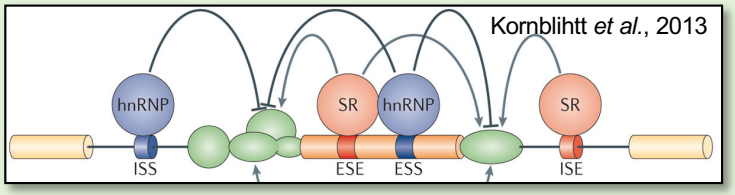
Variants can alter the protein-coding reading frame ("ORF"):

- Interrupt a start codon, or create a new start codon
- Interrupt a stop codon, or create a new stop codon
- Cause a frameshift (indels)



Variants can alter splicing:

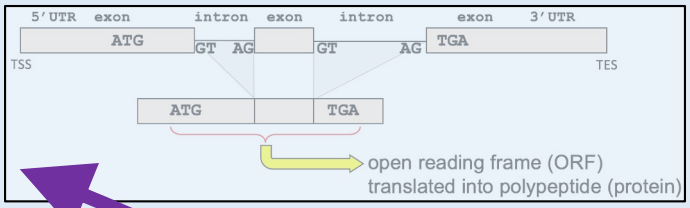
- Interrupt a splice site, or create a new splice site
- Alter splicing regulatory elements



Variants Can Alter Gene Structure

Variants can alter the protein-coding reading frame ("ORF"):

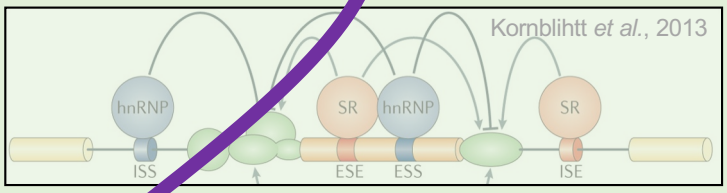
- Interrupt a start codon, or create a new start codon
- Interrupt a stop codon, or create a new stop codon
- Cause a frameshift (indels)



cascading effects

Variants can alter splicing:

- Interrupt a splice site, or create a new splice site
- Alter splicing regulatory elements

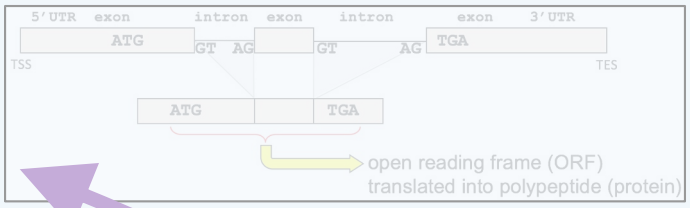


Note that changes to splicing can change the reading frame!

Variants Can Alter Gene Structure

Variants can alter the protein-coding reading frame ("ORF"):

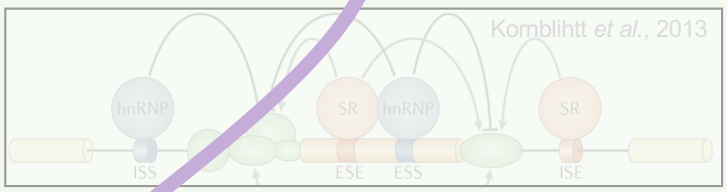
- Interrupt a start codon, or create a new start codon
- Interrupt a stop codon, or create a new stop codon
- Cause a frameshift (indels)



cascading effects

Variants can alter splicing:

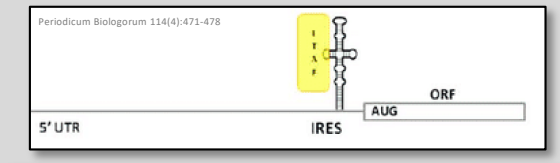
- Interrupt a splice site, or create a new splice site
- Alter splicing regulatory elements

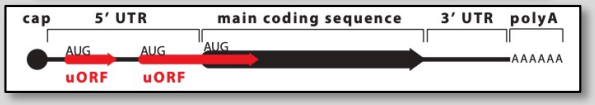


Note that changes to splicing can change the reading frame!

Variants can do other things:

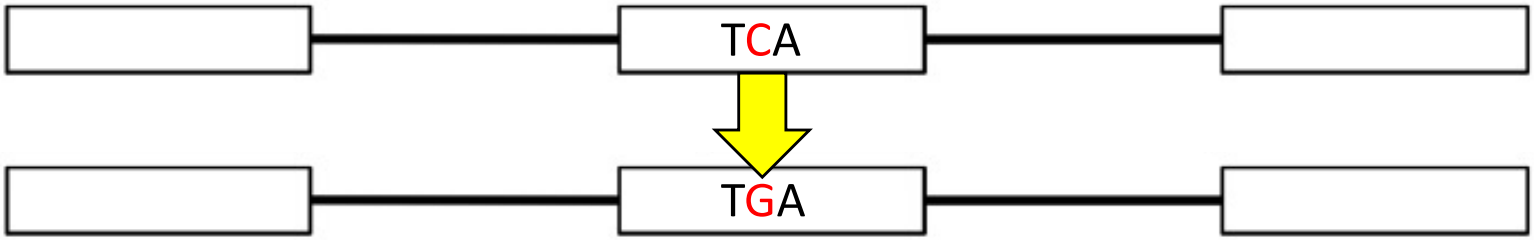
- Create secondary structures such as hairpin loops
- Create or interrupt upstream open reading frames (uORF)
- Create or interrupt internal ribosome entry sites (IRES)





Variants Can Modify the Reading Frame

A common way to modify the reading frame is to introduce a pre-termination codon (PTC):

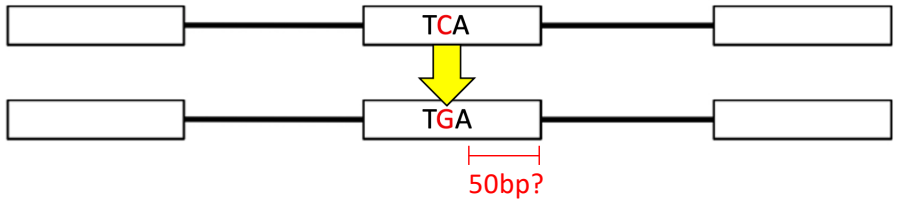


- A PTC is only functional if it occurs in the translation reading frame.
- If the PTC occurs late in the gene, it can result in protein truncation, which may or may not be deleterious—via loss of function or gain of function (e.g., *dominant-negative*).
- If it occurs early in the gene, it can result in nonsense-mediate decay (NMD) . . .

Nonsense Mediated Decay (NMD)

The 50 bp "rule":

- PTC at least 50-55 bp upstream of last exon junction
- But PTCs far upstream can escape NMD



Some PTCs escape NMD:

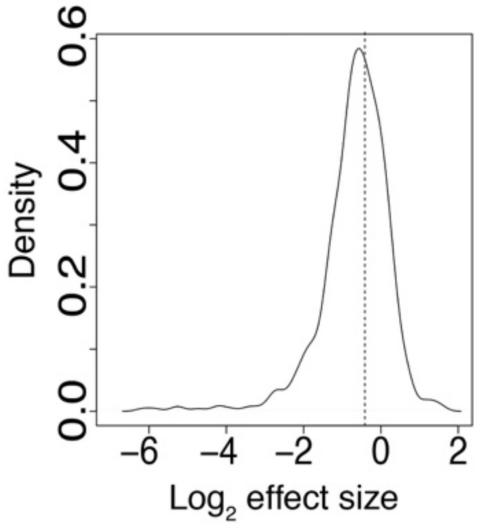
A Novel *FLCN* c.1489_1490delTG Mutation that Escapes the Nonsense-Mediated Decay System

Yong-Jin Park[1], Seog-Ki Lee[2], Seong-Ho Kang[3], Sook-Jin Jang[3], Dae-Soo Moon[3], and Geon Park[3]

[1]Department of Emergency Medicine, Chosun University College of Medicine, Gwangju, [2]Department of Thoracic and Cardiovascular Surgery, Chosun University College of Medicine, Gwangju, and [3]Department of Laboratory Medicine, Chosun University College of Medicine, Gwangju, South Korea

Abstract. A novel *FLCN* c.1489_1490delTG (p.Val497Glyfs*22) mutation at the genomic DNA and mRNA levels was identified in a 43-year-old woman with complaining of recurrent primary spontaneous pneumothorax. The aberrant *FLCN* mRNA escaped the nonsense-mediated decay system (NMD) because of a premature termination code located in an NMD-incompetent region. To the best of our knowledge, this is the first case report of an FLCN mutation escaping the NMD.

NMD effect size varies:



(Majoros, 1017)
(also: Rosenberg, 2015)

➔ On average, NMD results in roughly a halving of expression of each affected copy

➔ The remaining undegraded transcripts can have deleterious gain-of-function effects (e.g., dominant-negative)

Variants Can Cause Frameshifts

Translation reading frames:

phase 0:

```
ATG GAC CAC CCA ATT GTG GTT GAG CAG CCA GAT GCC TGG ACA GAG GAC AAT GGC TTC TGA
met asp his pro ile val val glu gln pro asp ala trp thr glu asp asn gly phe ***
```

phase 2:

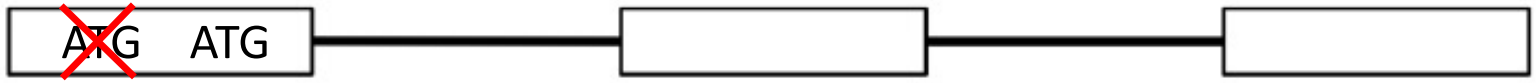
```
A TGG ACC ACC CAA TTG TGG TTG AGC AGC CAG ATG CCT GGA CAG AGG ACA ATG GCT TCT GA
  trp thr thr gln leu trp leu ser ser gln met pro gly gln arg thr met ala ser => no stop
```

phase 1:

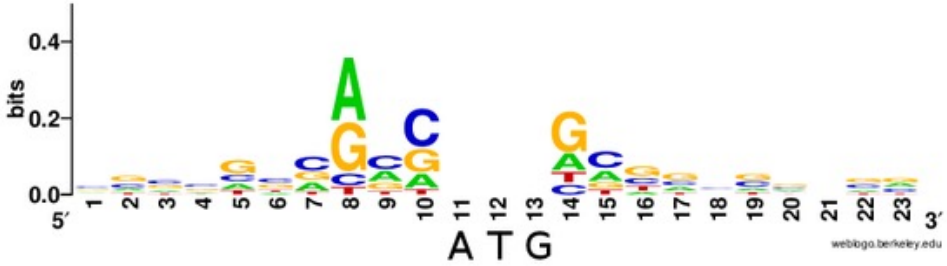
```
AT GGA CCA CCC AAT TGT GGT TGA GCA GCC AGA TGC CTG GAC AGA GGA CAA TGG CTT CCA TG
   gly pro pro asn cys gly ***
```

- Indels (insertions/deletions) can cause a frameshift if the length is not divisible by 3
- Splicing changes can also cause frameshifts
- Changes to start codons can cause frameshifts

Variants Can Change the Start Codon



Ribosome scanning model: the ribosome scans for the first start codon with a strong Kozak signal ("5' cap-dependent translation")



https://en.wikipedia.org/wiki/Kozak_consensus_sequence

Heterozygous *SSBP1* start loss mutation co-segregates with hearing loss and the m.1555A>G mtDNA variant in a large multigenerational family

Peter J Kullar, Aurora Gomez-Duran, Payam A Gammage, Caterina Garone, Michal Minczuk, Zoe Golder, Janet Wilson, Julio Montoya, Sanna Häkli, Mikko Kärppä ... Show more

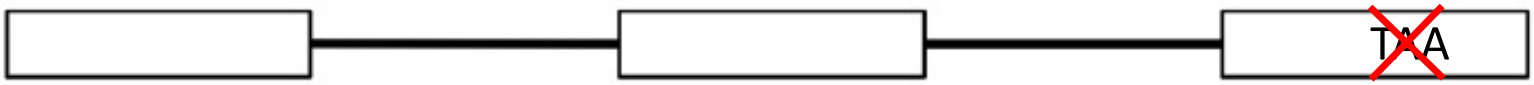
Brain, Volume 141, Issue 1, January 2018, Pages 55–62, https://doi.org/10.1093/brain/awx295
Published: 22 November 2017 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

The m.1555A>G mtDNA variant causes maternally inherited deafness, but the reasons for the highly variable clinical penetrance are not known. Exome sequencing identified a heterozygous start loss mutation in *SSBP1*, encoding the single stranded binding protein 1 (SSBP1), segregating with hearing loss in a multi-generational family transmitting m.1555A>G, associated with mtDNA depletion and multiple deletions in skeletal muscle. The *SSBP1* mutation reduced steady state SSBP1 levels leading to a perturbation of mtDNA metabolism, likely compounding the intra-mitochondrial translation defect due to m.1555A>G in a tissue-specific manner. This family demonstrates the importance of rare *trans*-acting genetic nuclear modifiers in the clinical expression of mtDNA disease.

Variants Can Disrupt the Stop Codon



No stop codon – ribosomes pile up at the poly-A tail, triggering "non-stop decay" (NSD).

Non-stop decay—a new mRNA surveillance pathway

Shobha Vasudevan, Stuart W. Peltz, and Carol J. Wilusz*

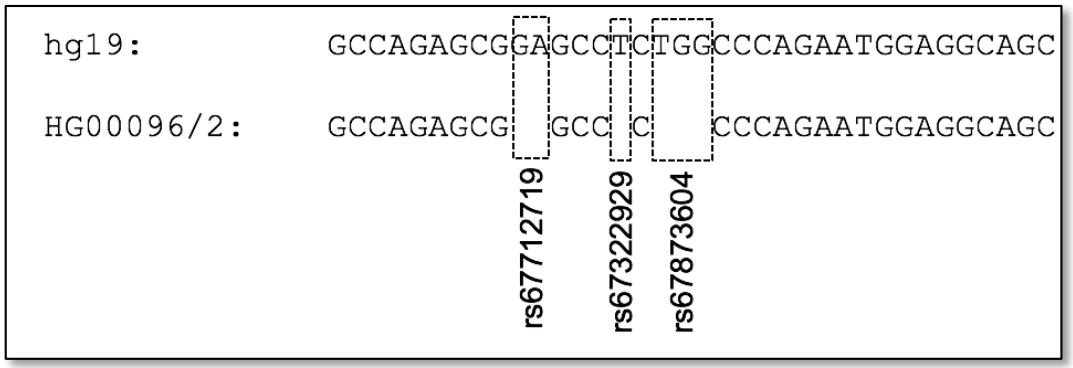
BioEssays 24:785–788, © 2002 Wiley Periodicals, Inc. **BioEssays 24.9** 785

Summary

Gene expression is an inherently complex process and errors often occur during the transcription and processing of mRNAs. Several surveillance mechanisms have evolved to check the fidelity at each step of mRNA manufacture. Two recent reports describe the identification of a novel pathway in eukaryotes that recognizes and degrades mRNAs that lack a stop codon.[1,2] The non-stop decay mechanism releases ribosomes stalled at the 3′ end of a mRNA and stimulates the exosome to rapidly degrade the transcript. *BioEssays* 24:785–788, 2002.

Variants are Not Independent!

The first two deletions shown below would shift the reading frame, but together they affect only two amino acids:



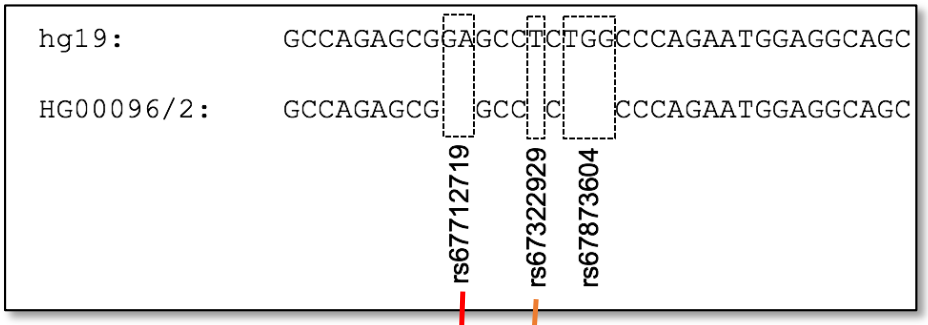
4869 / 5008 = 97% of Thousand Genomes haplotypes contain the first two variants

In Thousand Genomes, the first two variants overwhelmingly occur together. Because that haplotype is so common, it is unlikely to be deleterious.

Ensembl's VEP tool predicts that the first two variants would be deleterious.

This is not anecdotal! Every individual in the 1000 Genomes Project sample has one or more compensatory frameshifts (median = 7 per individual) affecting ≤30 amino acids.

Example Continued




```
reference:     AGA GCG GAG CCT CTG GCC CAG AAT GGA GGC AGC ...(1662bp)... TGA
1st deletion:  AGA GCG GCC TCT GGC CCA GAA TGG AGG CAG CAG ...(651bp).... TAG => truncation
```

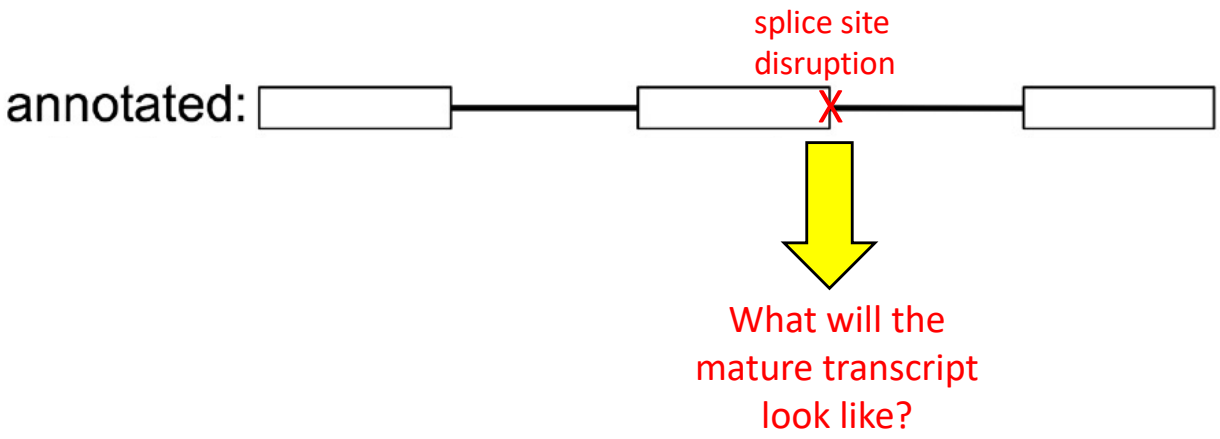
```
reference:     AGA GCG GAG CCT CTG GCC CAG AAT GGA GGC AGC ...(1662bp)... TGA
2nd deletion:  AGA GCG GAG CCC TGG CCC AGA ATG GAG GCA GCA ...(1032bp)... TGA => truncation
```

All three variants:

```
reference:     AGA GCG GAG CCT CTG GCC CAG AAT GGA GGC AGC ...(1662bp)... TGA
HG00096/2:     AGA GCG GCC         CCC CAG AAT GGA GGC AGC ...(1662bp)... TGA  => local protein change (4aa)
```

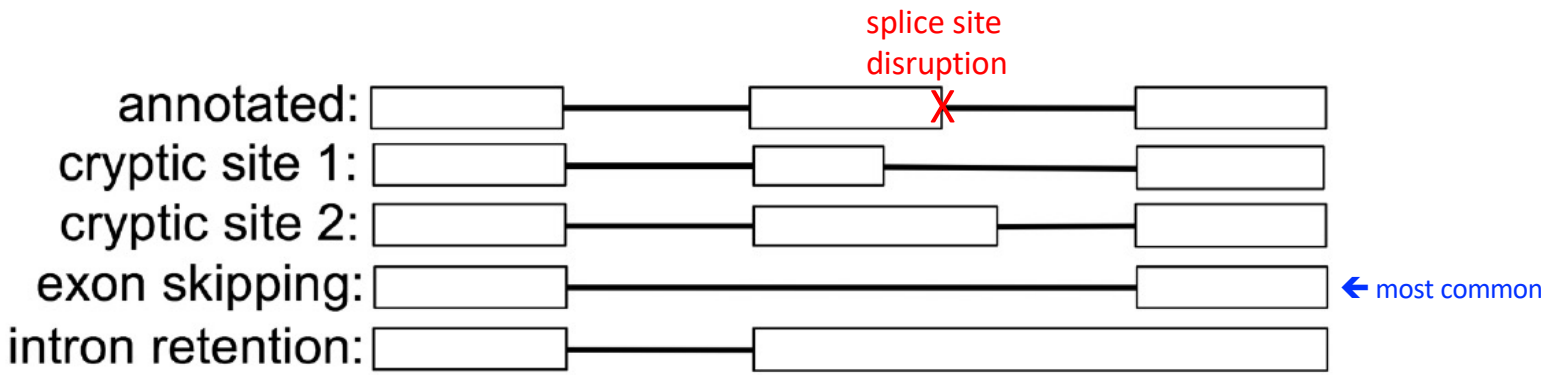
Changes to Splicing

Variants that disrupt an existing splice site can have <u>multiple possible outcomes</u>:



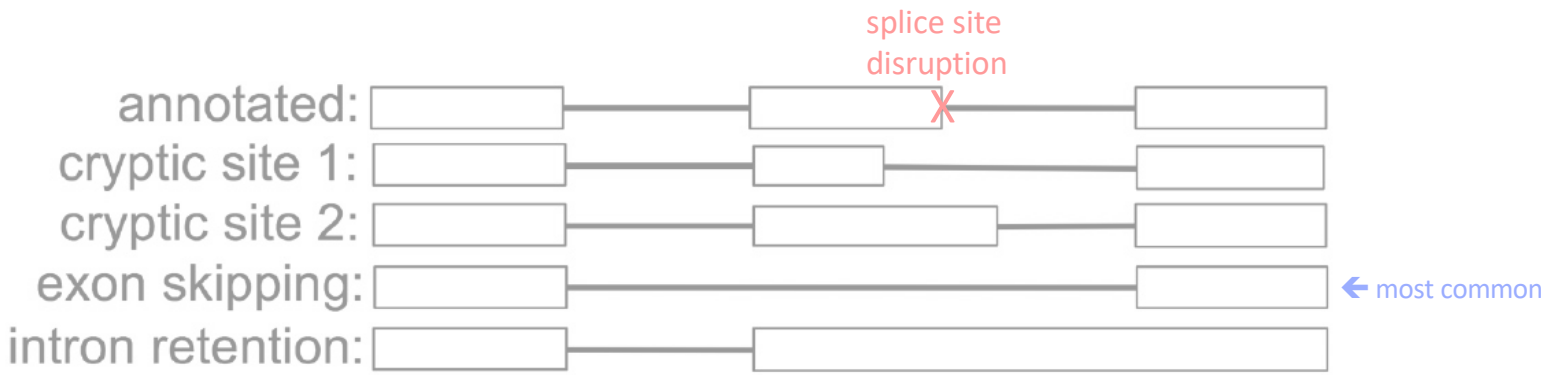
Changes to Splicing

Variants that disrupt an existing splice site can have <u>multiple possible outcomes</u>:



Changes to Splicing

Variants that disrupt an existing splice site can have multiple possible outcomes:

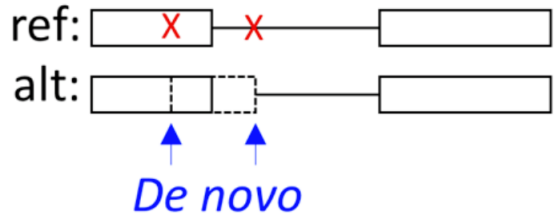


Splicing is stochastic! One variant can result in a mixture of transcripts with different structures in the same cell.

De Novo Splice Sites

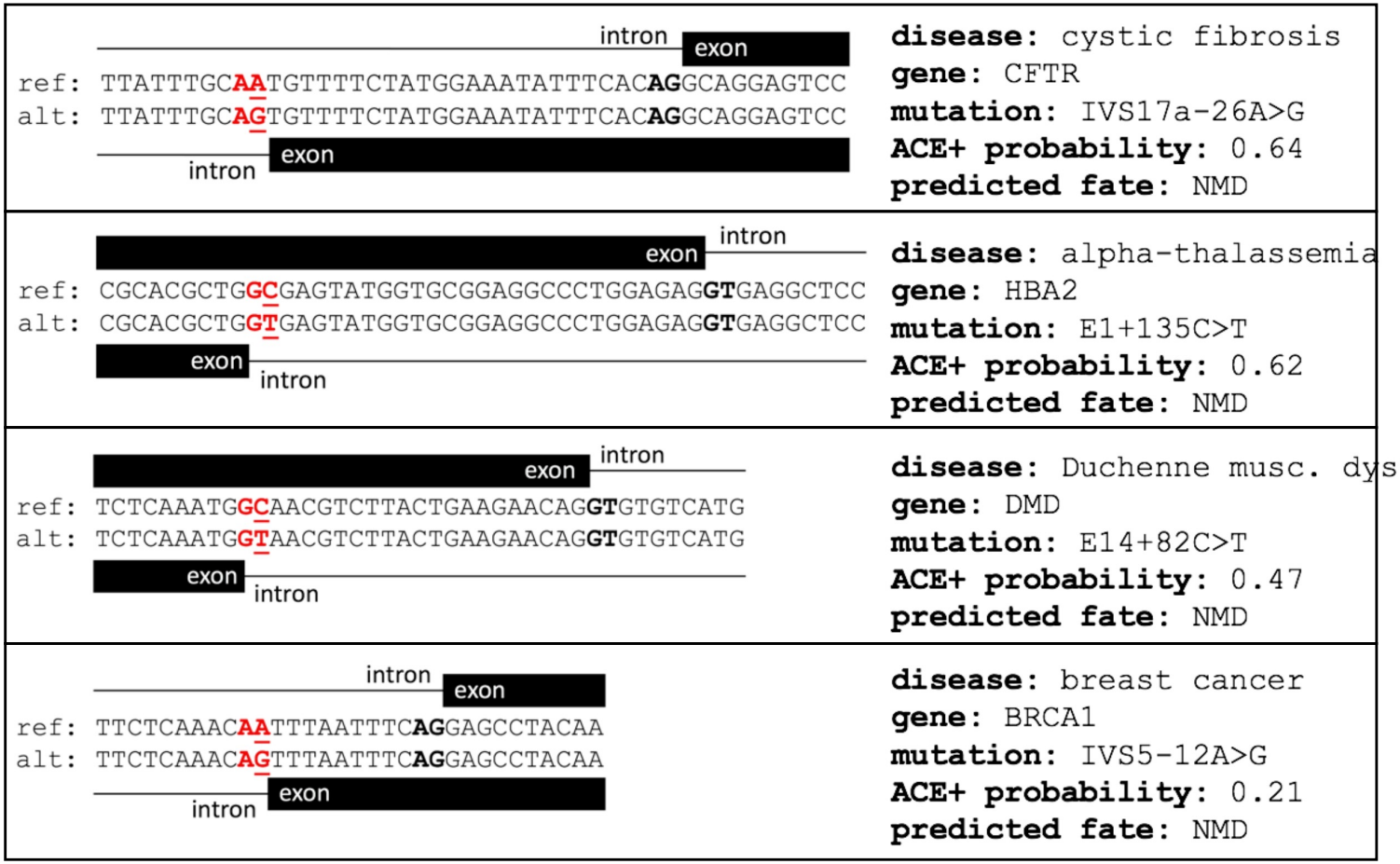
A variant can create an entirely new splice site – this is not as unlikely as you might think!

De novo sites:



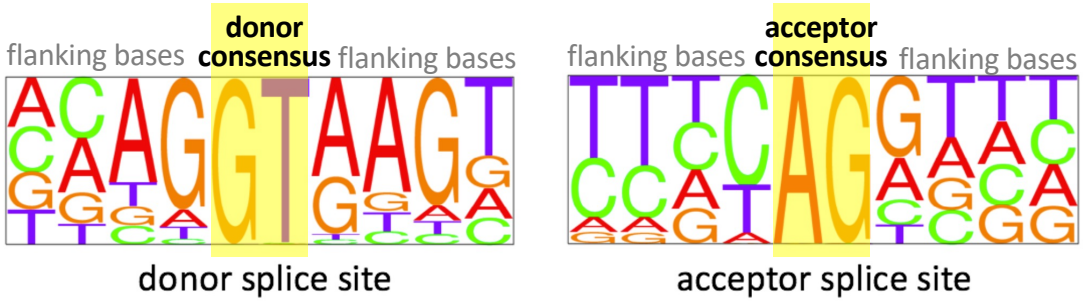
De Novo Splice Sites Can Cause Disease!

DBASS
Data Base of Aberrant Splice Sites

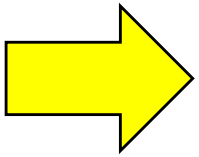


Variants Can Alter Splice Site Strength

A mutation that does not change the donor (GT) or acceptor (AG) consensus, but changes one or more flanking bases can strengthen or weaken an existing splice site:



- Weaken an existing splice site
- Strengthen a cryptic splice site



Change isoform ratios

> Nat Biotechnol. 2004 May;22(5):535-46. doi: 10.1038/nbt964.

Alternative splicing in disease and therapy

Mariano A Garcia-Blanco [1], Andrew P Baraniak, Erika L Lasda

Affiliations + expand

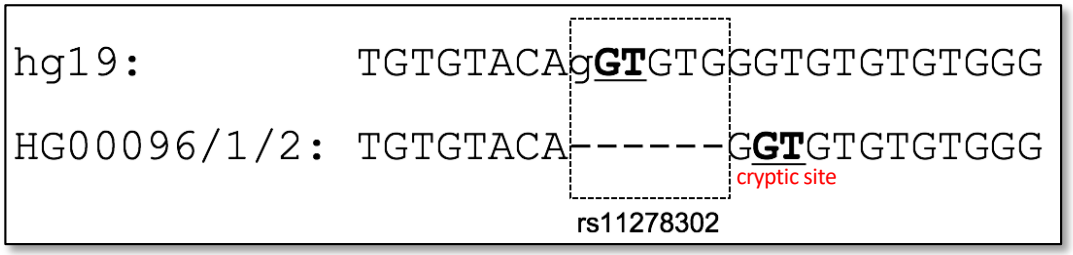
PMID: 15122293 DOI: 10.1038/nbt964

➔ mutations that lead to even subtle changes in the ratio of MAPT isoforms 3R and 4R cause an inherited form of dementia

Context Is Important!

The local context of a splice-altering variant (SAV) is important.

In the example below, an entire splice site is deleted, but there is a “cryptic splice” site nearby, and after the deletion the cryptic site is actually predicted to be stronger than the original site.



Context Is Important!

The local context of a splice-altering variant (SAV) is important.

In the example below, an entire splice site is deleted, but there is a "cryptic splice" site nearby, and after the deletion the cryptic site is actually predicted to be stronger than the original site.

New splice site is stronger than the original

Allele frequency = 0.23

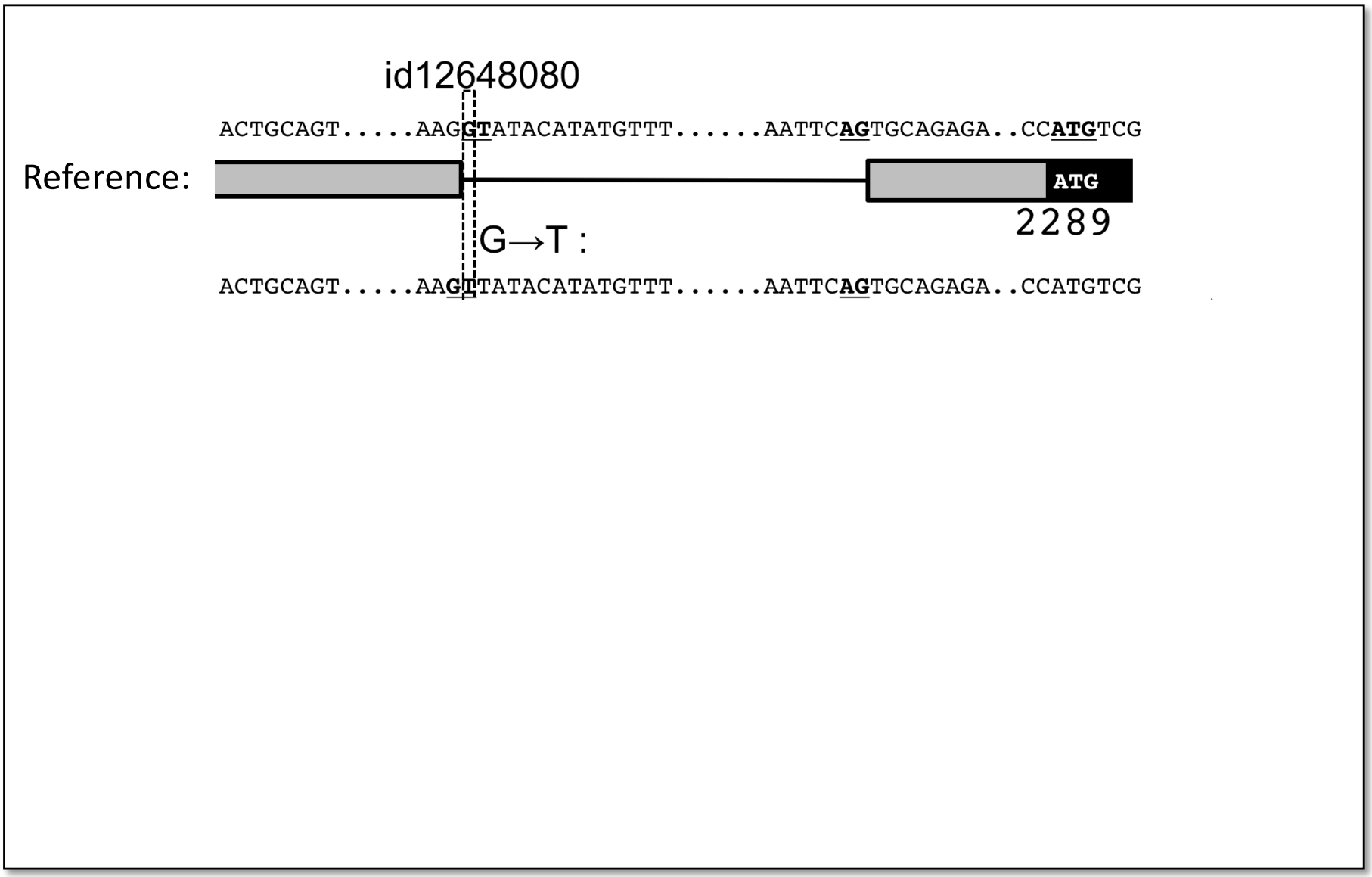
The result is a loss of two amino acids and no change to the reading frame.

Because the variant is common, it is unlikely to be deleterious.

Ensembl's VEP tool predicts that this variant would be deleterious, because it deletes an entire splice site.

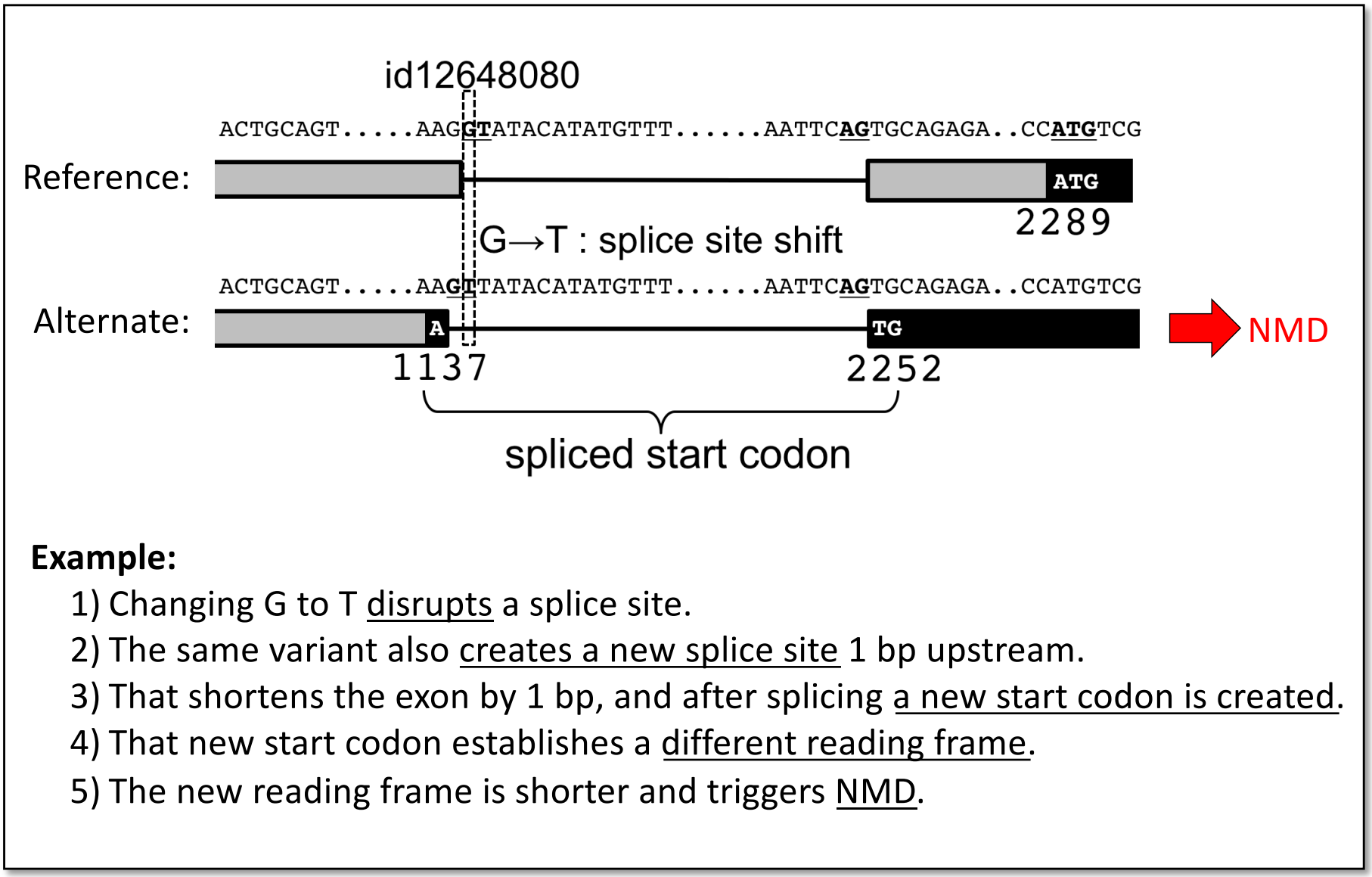
Cascading Effects

Changes to splicing often have cascading downstream effects:



Cascading Effects

Changes to splicing often have cascading downstream effects:

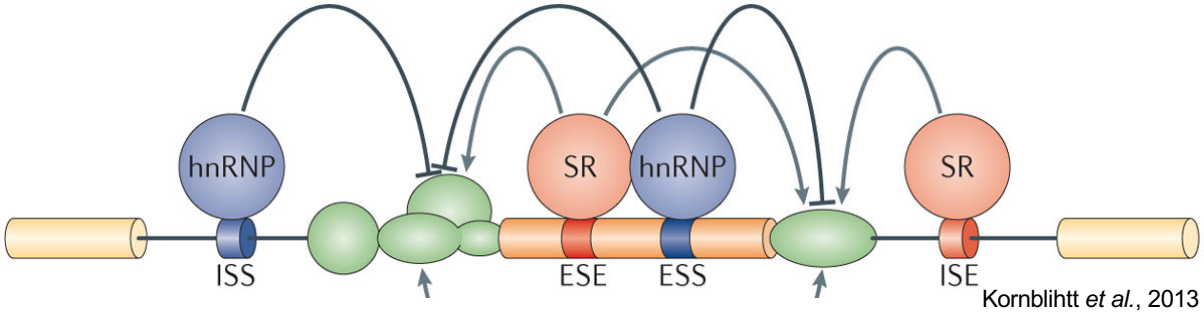


Example:

1) Changing G to T disrupts a splice site.
2) The same variant also creates a new splice site 1 bp upstream.
3) That shortens the exon by 1 bp, and after splicing a new start codon is created.
4) That new start codon establishes a different reading frame.
5) The new reading frame is shorter and triggers NMD.

Variants Can Alter Splicing Regulatory Elements

Variants that alter splicing enhancers/silencers can result in repression or activation of splice sites:

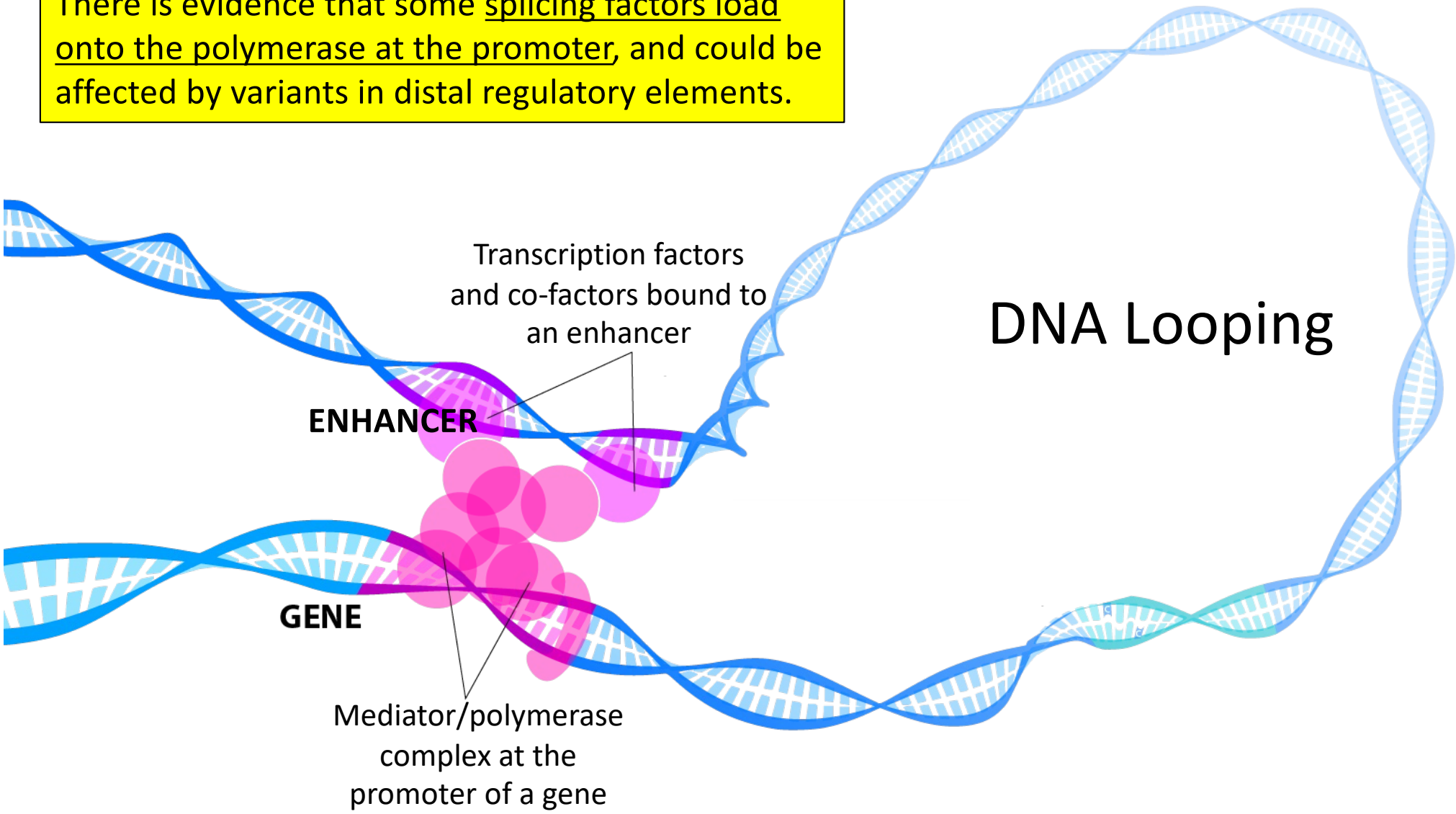


Kornblihtt *et al.*, 2013

These are the most difficult cases to predict. Current methods are focusing on deep-learning neural networks, but their predictive accuracy isn't perfect, and they aren't always interpretable.

Distal Variants Might Impact Gene Structure

There is evidence that some splicing factors load onto the polymerase at the promoter, and could be affected by variants in distal regulatory elements.



Adapted from figure by Kevin Song (Wikimedia Commons)

Questions?