Eukaryotic Gene Structure and Its Role in Genetic Disease

Part 1: The Elements of Gene Structure

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

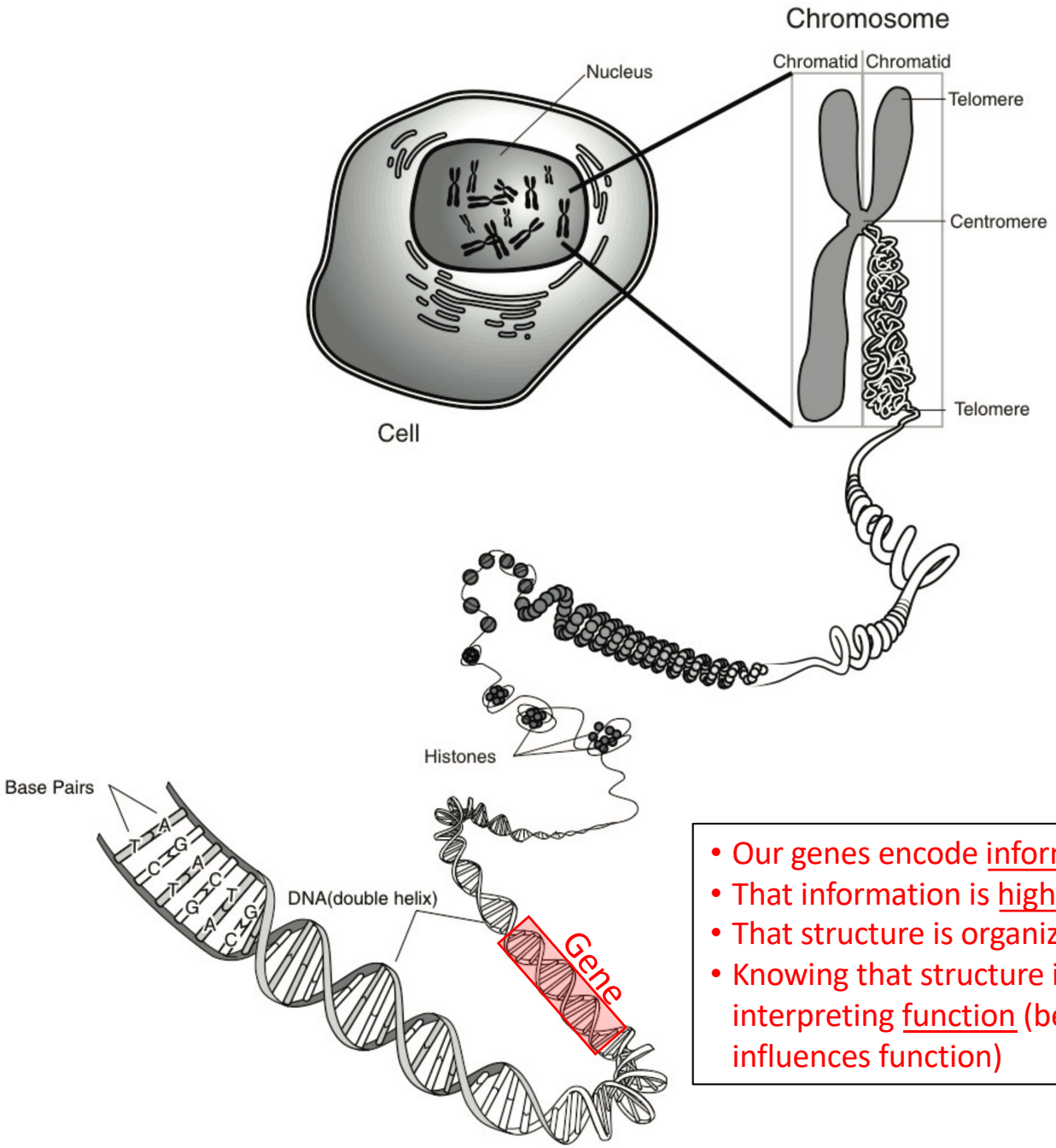
Outline

1. What is a Gene?
2. Defining Gene Structure
3. How Splicing Works
4. Decoding Gene Structure
5. For More Information

Outline

1. **What is a Gene?**
2. Defining Gene Structure
3. How Splicing Works
4. Decoding Gene Structure
5. For More Information

The Eukaryotic Cell

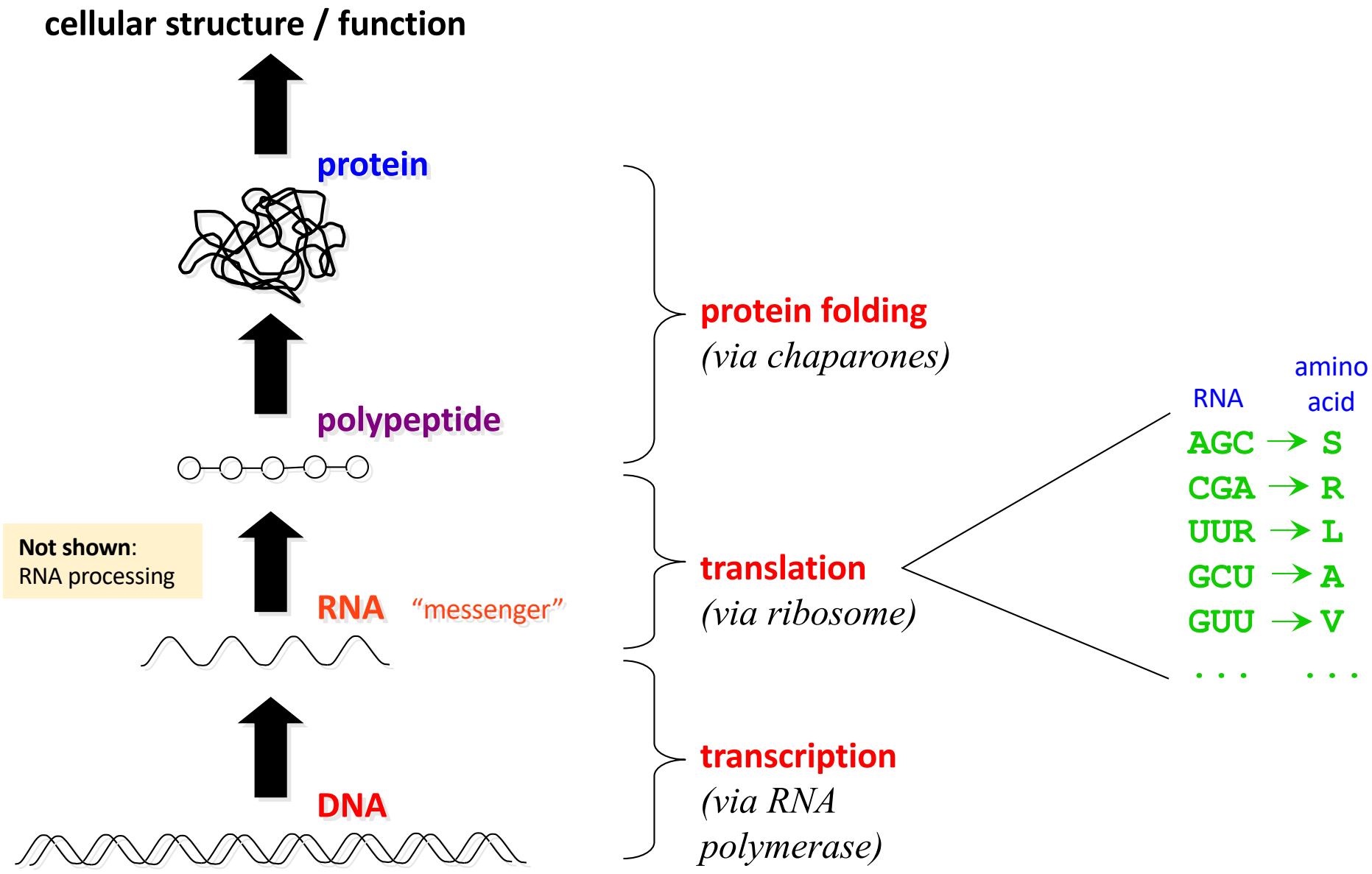


Sister chromatids after replication

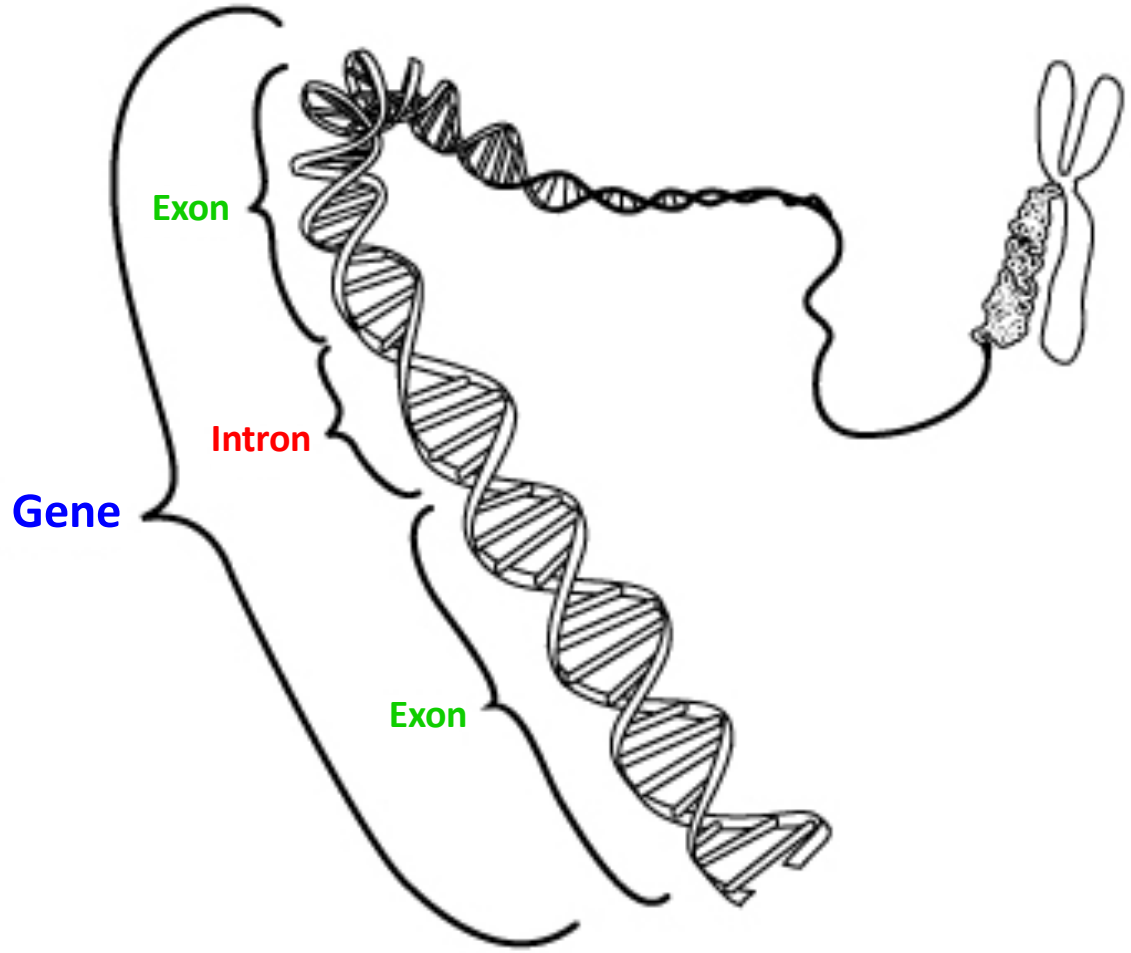
- Our genes encode information.
- That information is highly structured.
- That structure is organized hierarchically.
- Knowing that structure is important for interpreting function (because structure influences function)

NHGRI

The Central Dogma

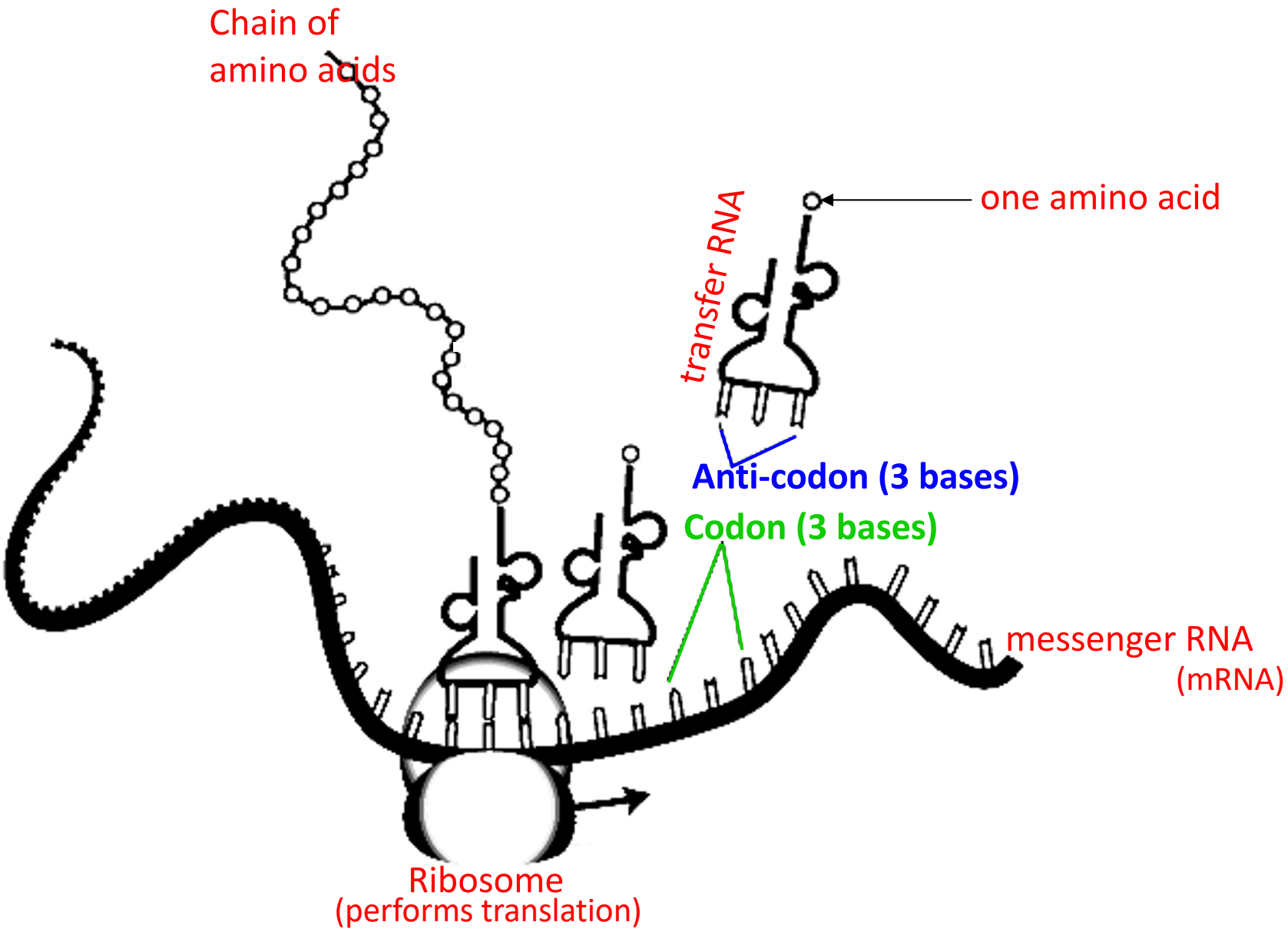


Genes: Exons and Introns



NHGRI

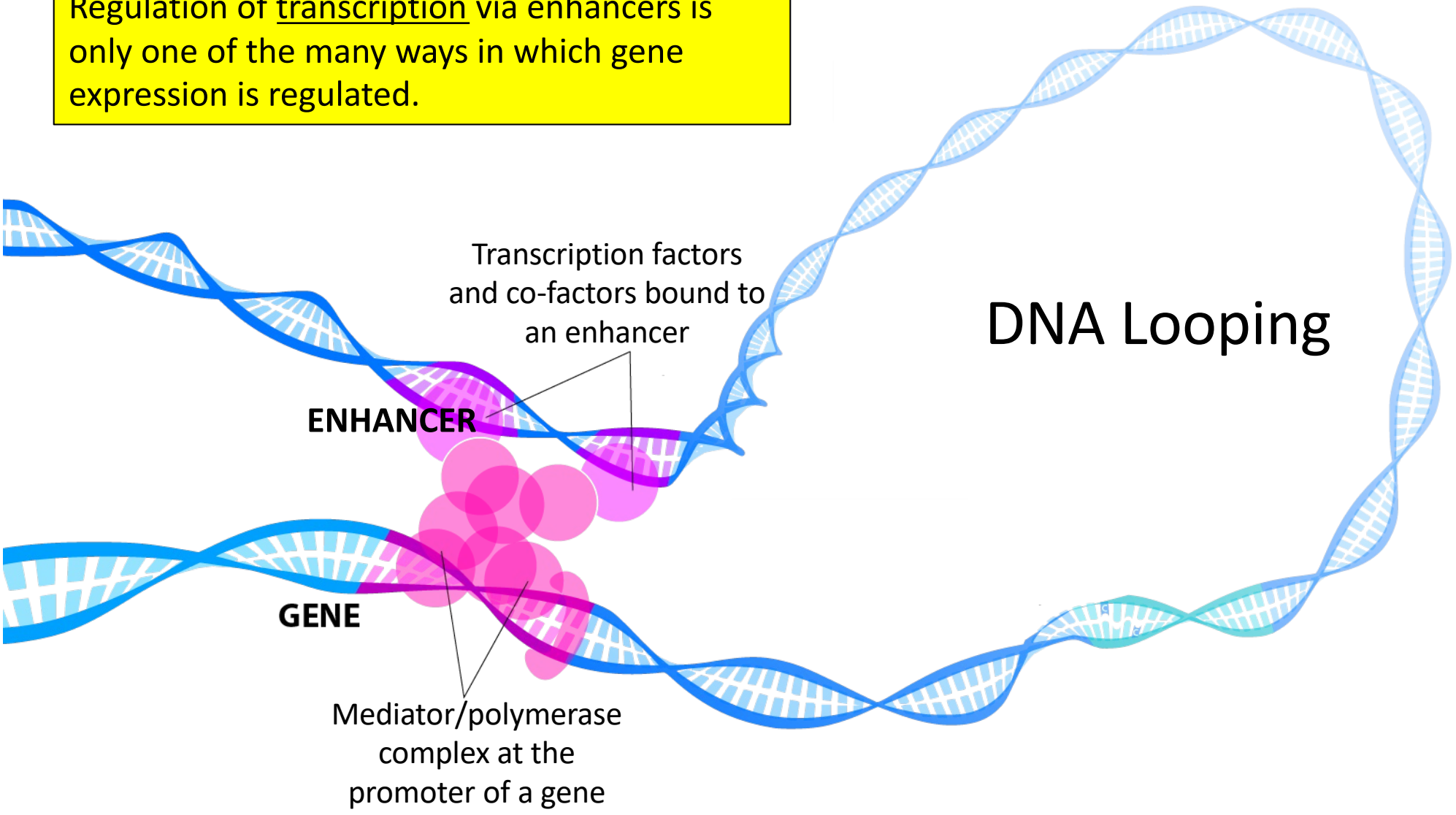
Translation: RNA to Protein



NHGRI

Gene Expression is Regulated

Regulation of transcription via enhancers is only one of the many ways in which gene expression is regulated.

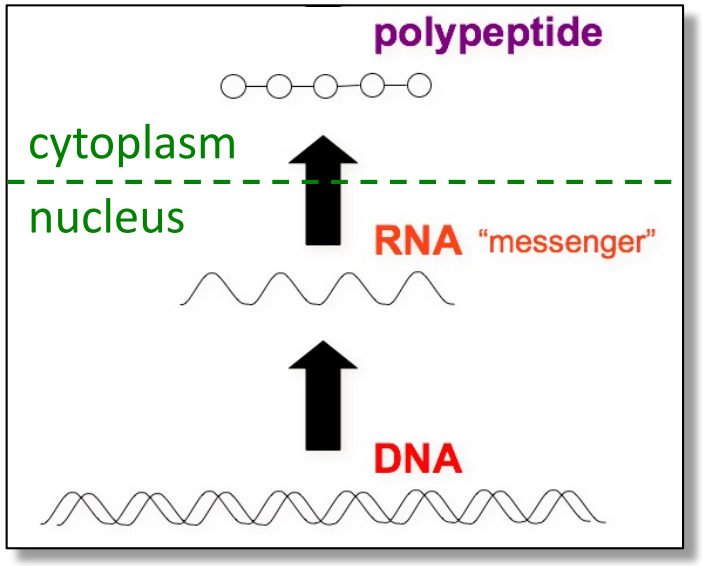


Adapted from figure by Kevin Song (Wikimedia Commons)

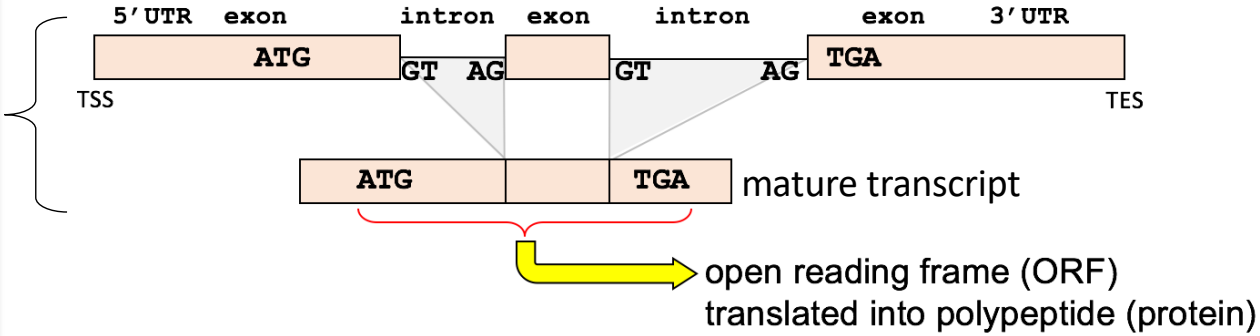
Outline

1. What is a Gene?
2. **Defining Gene Structure**
3. How Splicing Works
4. Decoding Gene Structure
5. For More Information

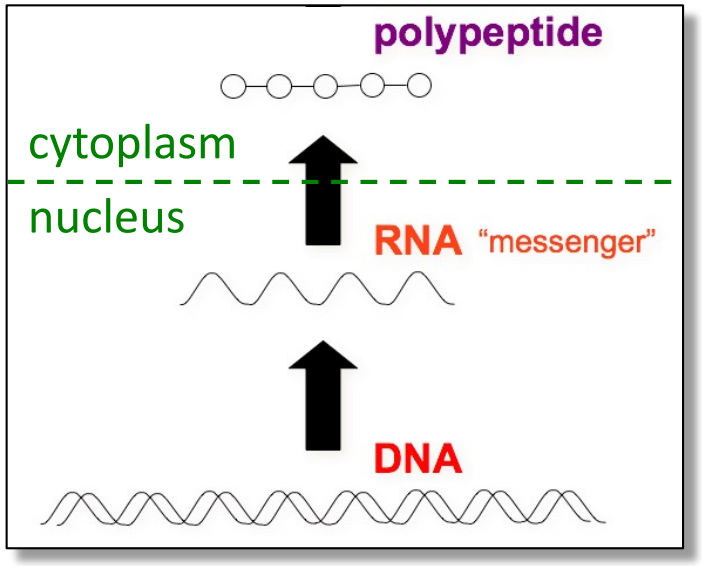
Gene Structure



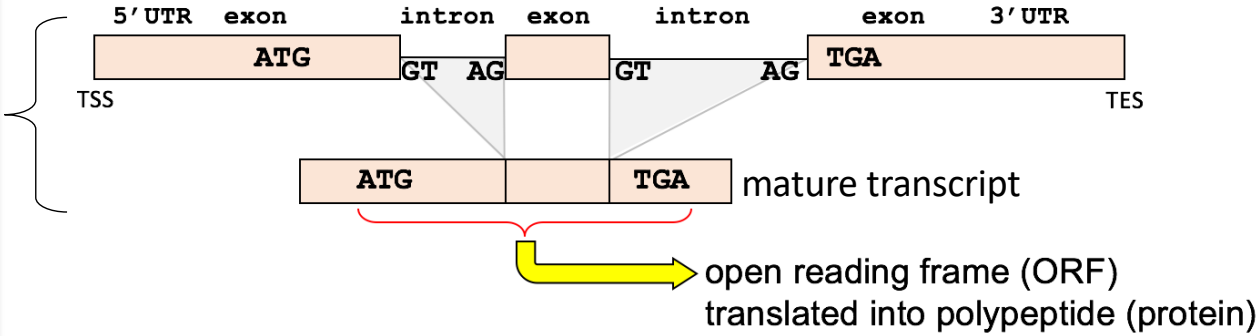
95% of human genes contain introns.



Gene Structure

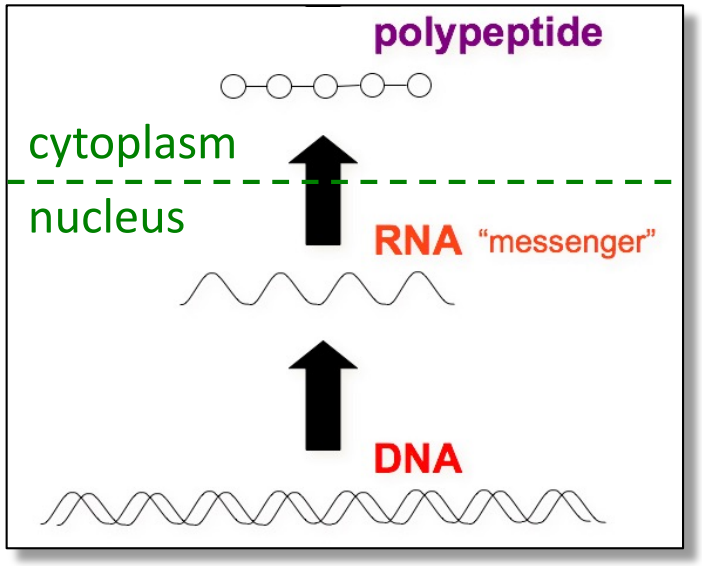


95% of human genes contain introns.

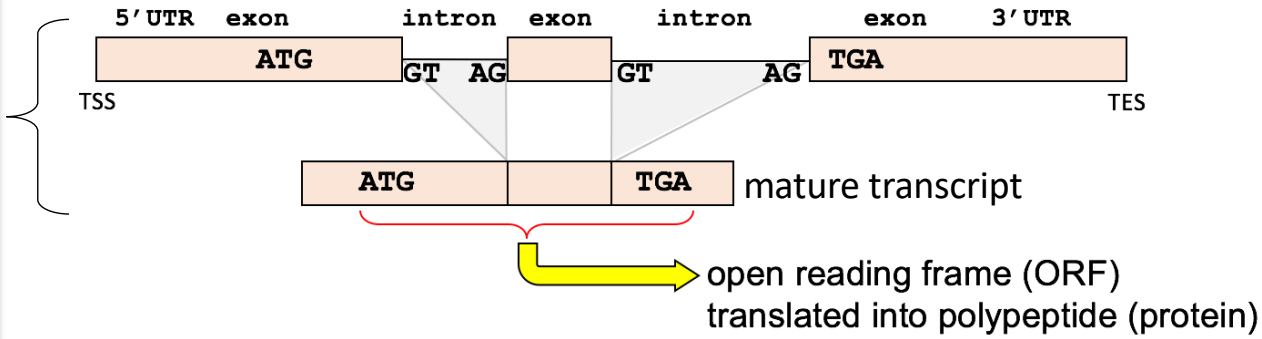


The way that a gene is spliced can impact how it is later translated.

Gene Structure



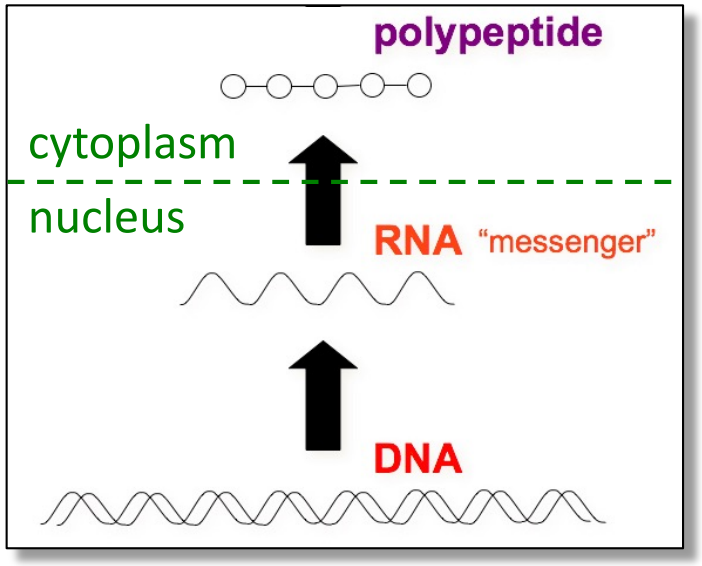
95% of human genes contain introns.



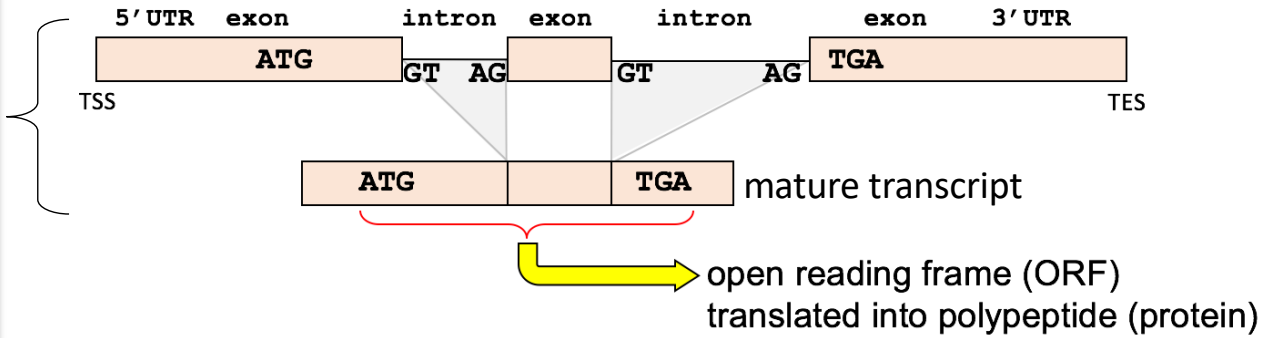
The way that a gene is spliced can impact how it is later translated.

- Failing to remove an intron within the ORF would result in additional nucleotides being translated into amino acids

Gene Structure



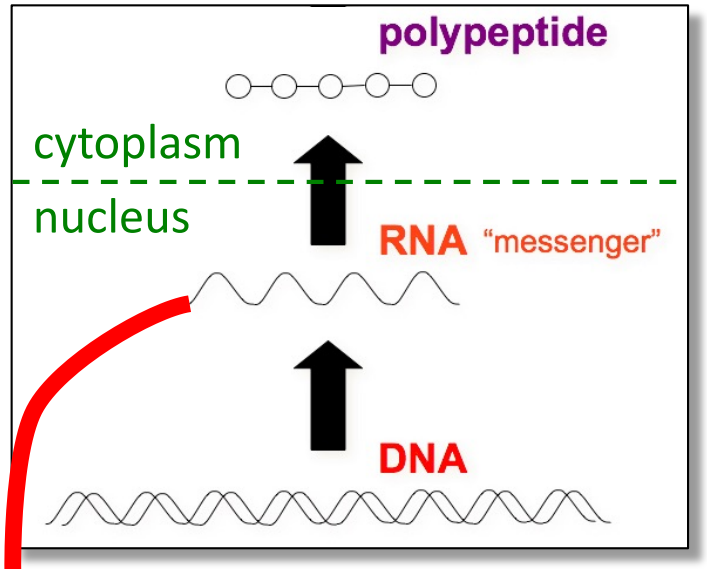
95% of human genes contain introns.

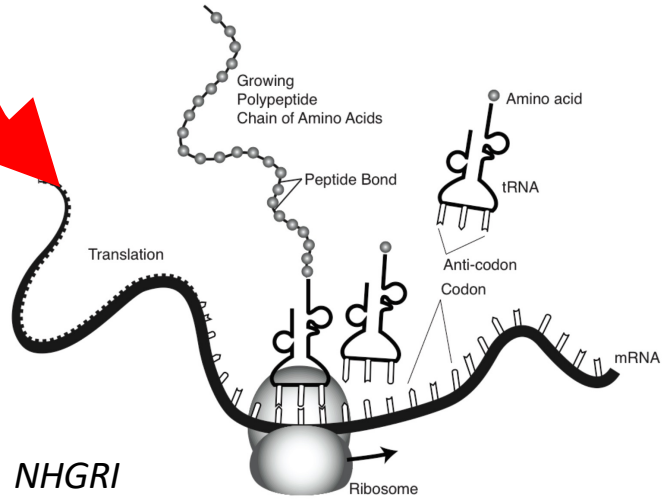


The way that a gene is spliced can impact how it is later translated.

- 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

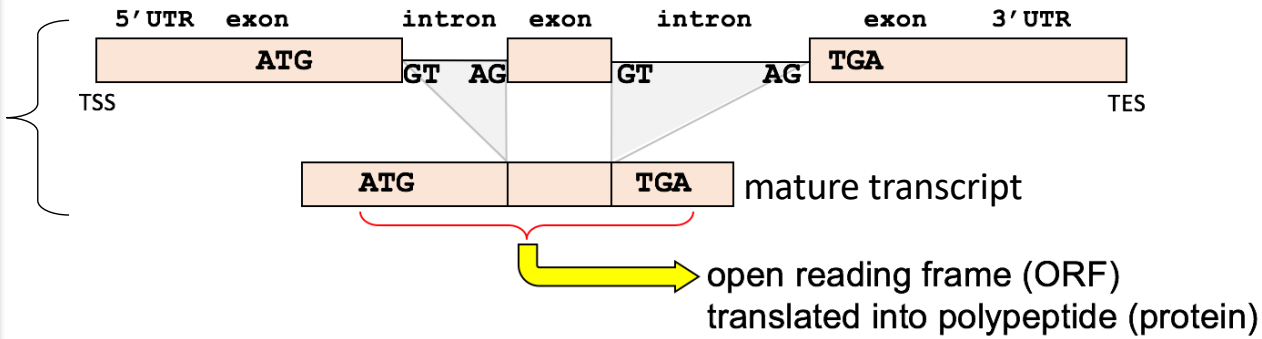
Gene Structure





NHGRI

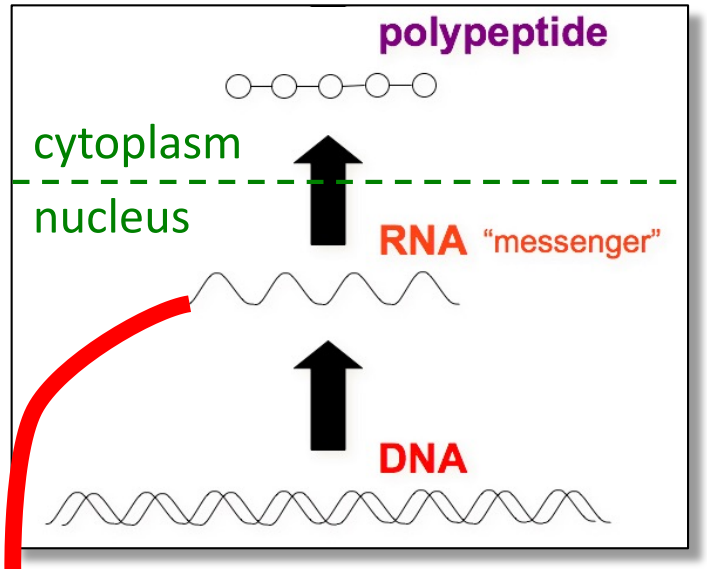
95% of human genes contain introns.

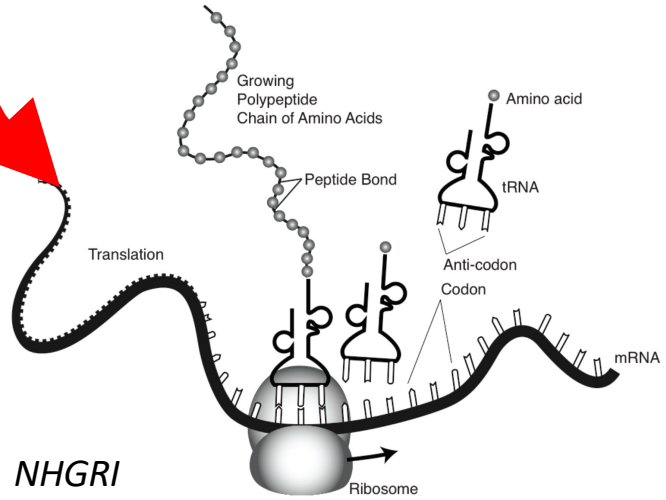


The way that a gene is spliced can impact how it is later translated.

- 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

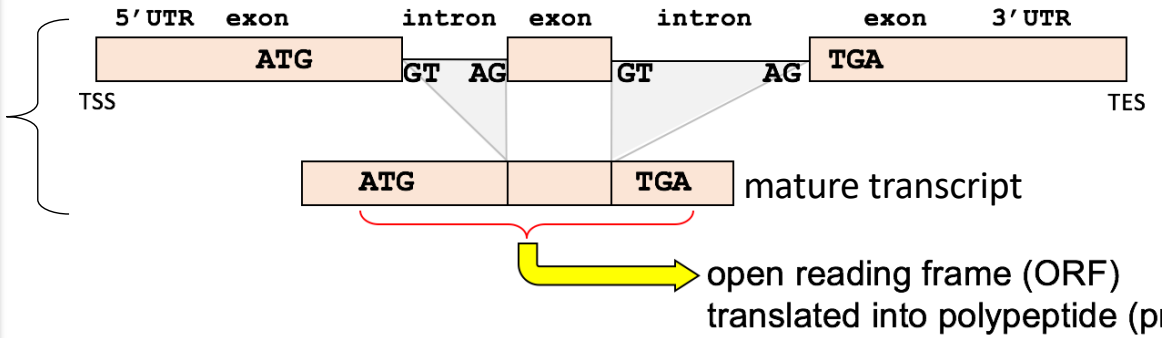
Gene Structure





NHGRI

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 ***
```

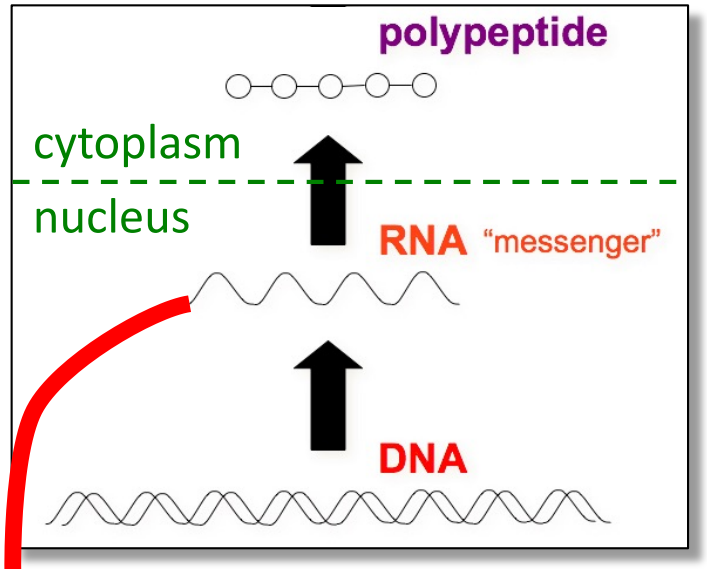
phase 2:

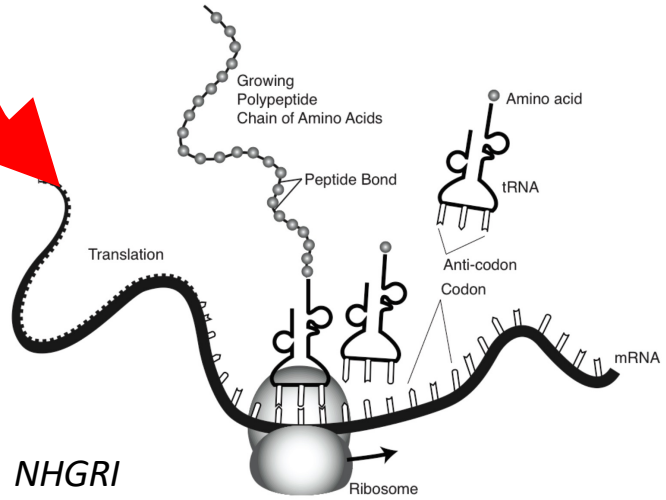
```
A TGG ACC ACC CAA TTG TGG TTG AGC AGC CAG ATG CCT GGA CAG AGG ACA ATG GCT TCT GA
  trp thr thr gln leu trp leu ser ser gln met pro gly gln arg thr met ala ser => no stop
```

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 ***
```

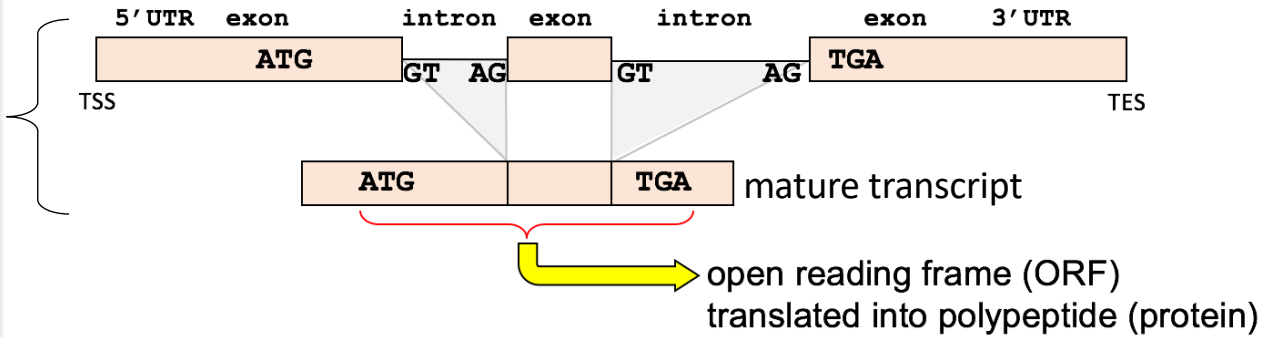
Gene Structure





NHGRI

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:

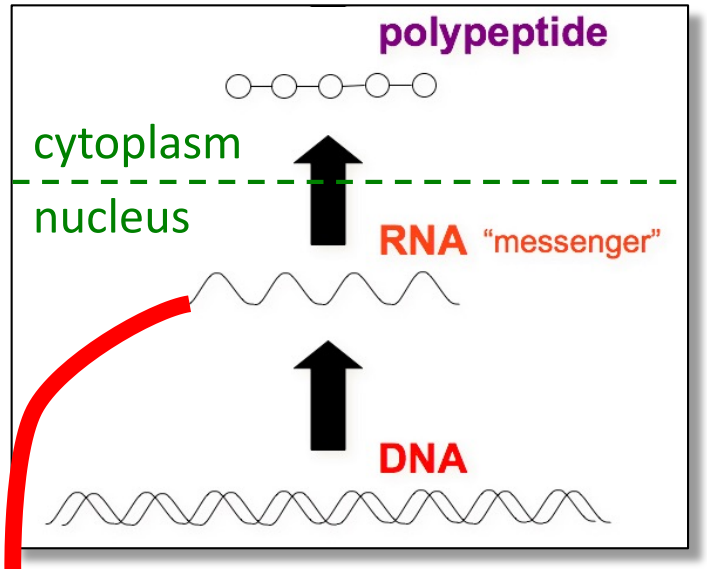
```
A TGG ACC ACC CAA TTG TGG TTG AGC AGC CAG ATG CCT GGA CAG AGG ACA ATG GCT TCT GA
  trp thr thr gln leu trp leu ser ser gln met pro gly gln arg thr met ala ser => no stop
```

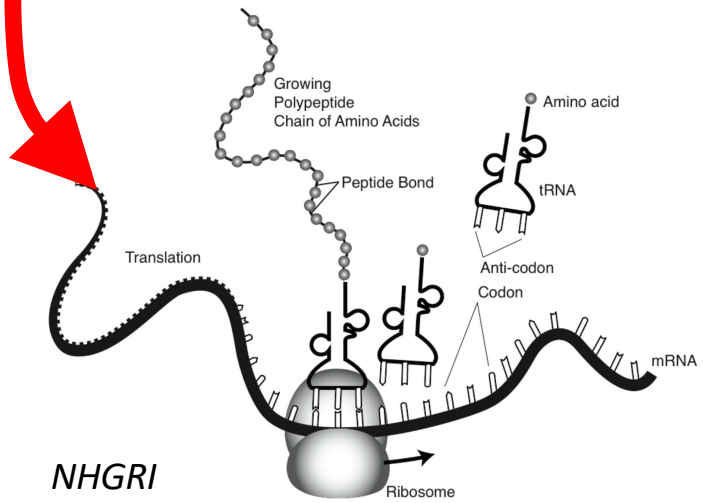
phase 1:

```
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 ***
```

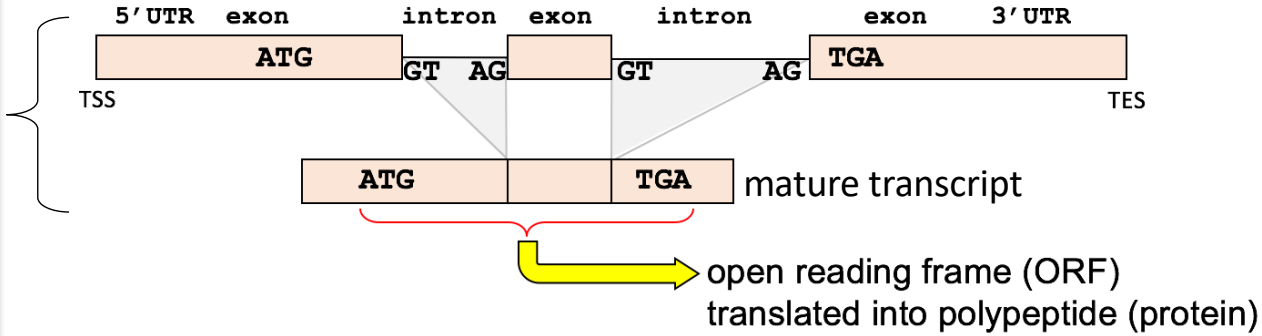
Changes in splicing can change the reading frame, and that can be highly disruptive.

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
```

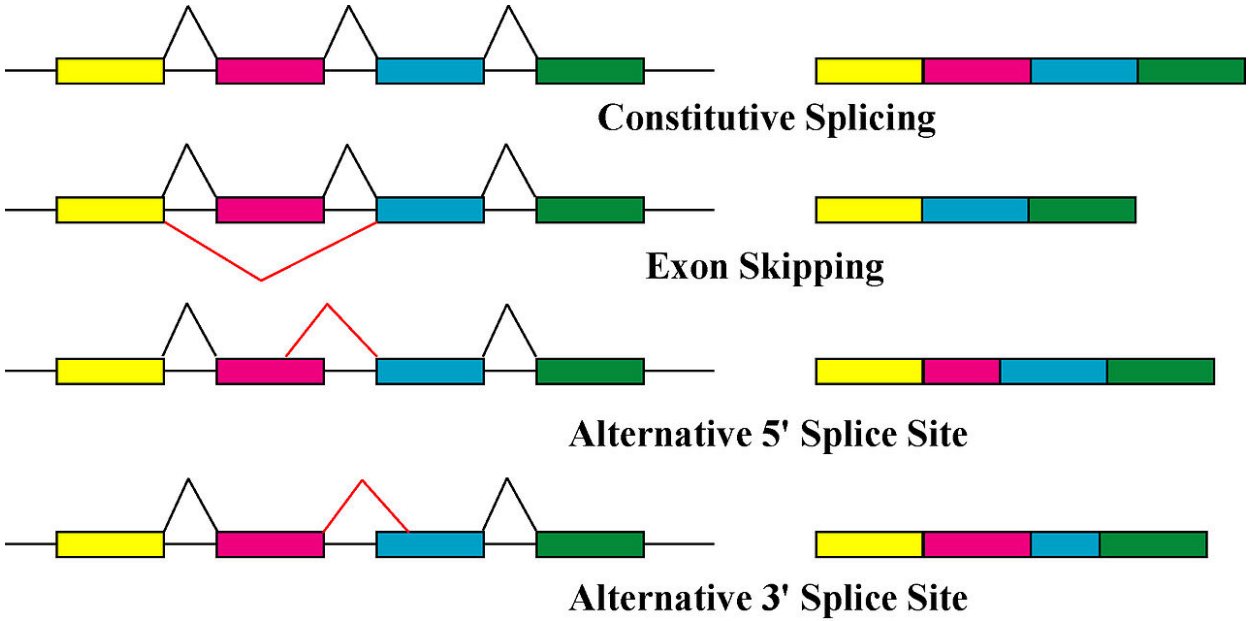
phase 1:

```
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 splicing pattern + reading frame = gene structure.

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)

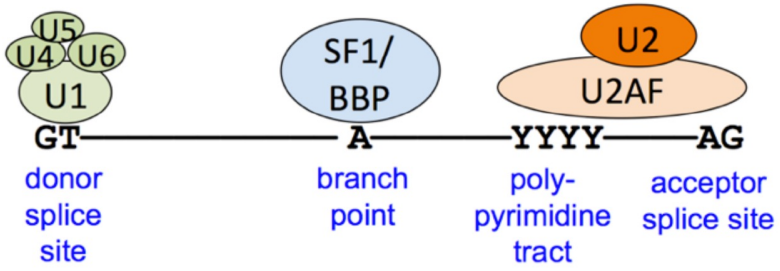
https://en.wikipedia.org/wiki/Protein_isoform

Outline

1. What is a Gene?
2. Defining Gene Structure
3. **How Splicing Works**
4. Decoding Gene Structure
5. For More Information

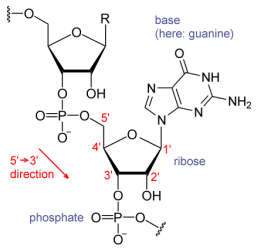
Splicing is a Complex & Highly Regulated Process

An intron:

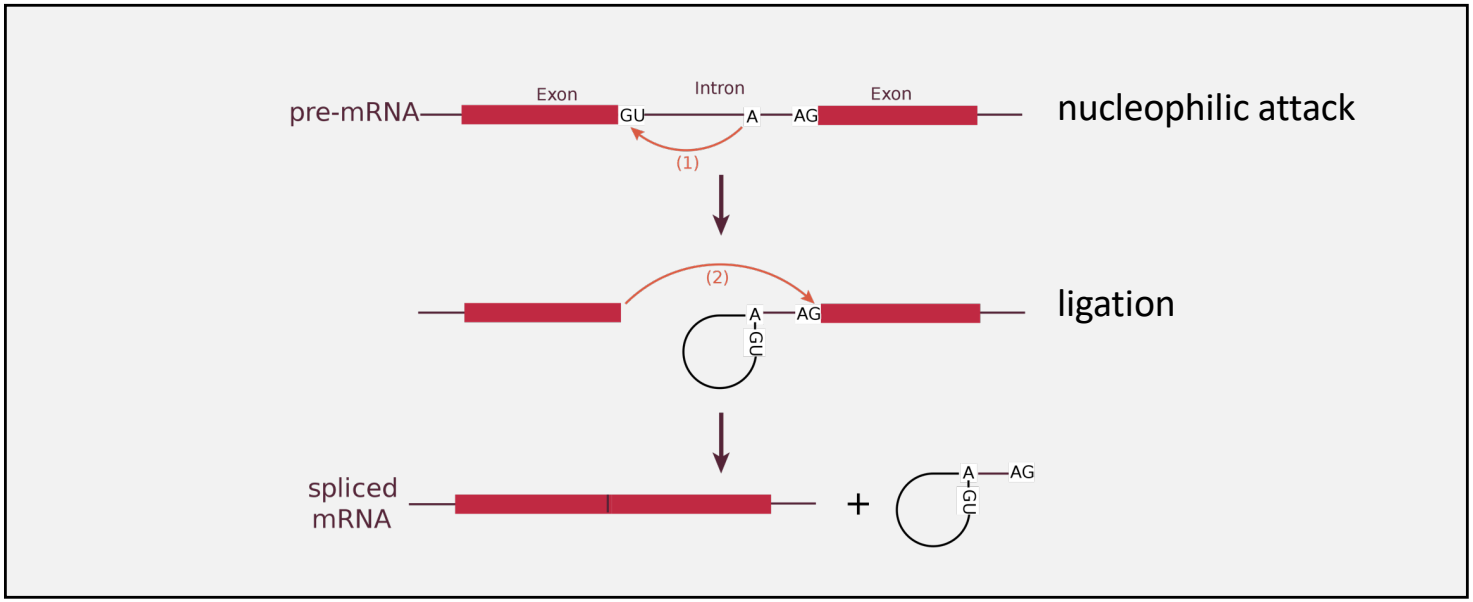


purines: A,G (R)
pyrimidines: C, T (Y)

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

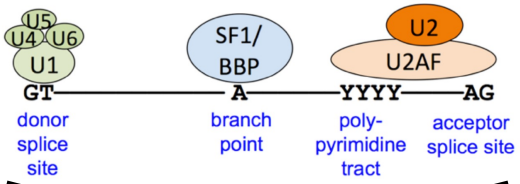


RNA nucleotides have an additional reactive 2' OH that facilitates this!



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

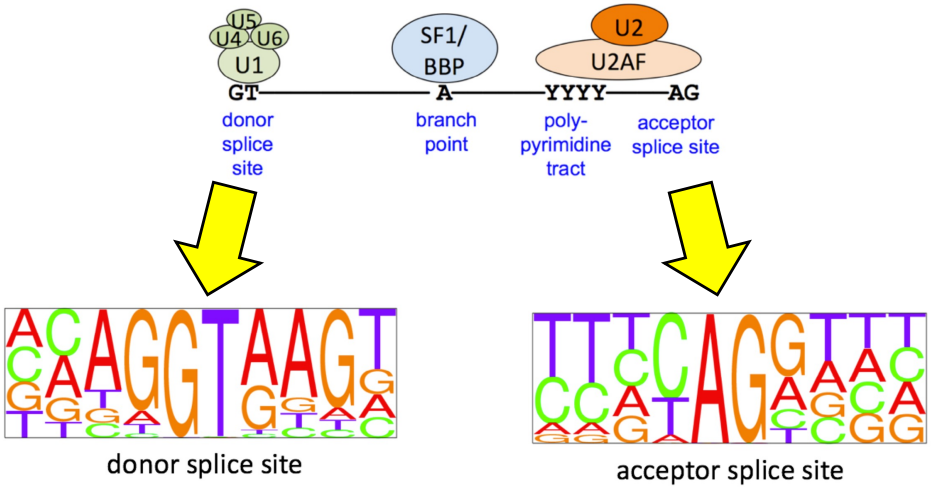
acceptor splice site

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

PWM = Probabilistic Weight Matrix

$$M = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{bmatrix} 0.25 & 0.31 & 0.56 & 0.13 & 0.00 & 0.00 & 0.69 & 0.81 & 0.38 & 0.44 \\ 0.13 & 0.38 & 0.13 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.06 \\ 0.31 & 0.25 & 0.19 & 0.81 & 1.00 & 0.00 & 0.13 & 0.13 & 0.56 & 0.19 \\ 0.31 & 0.06 & 0.13 & 0.06 & 0.00 & 1.00 & 0.18 & 0.06 & 0.06 & 0.31 \end{bmatrix}$$

Splicing is a Complex & Highly Regulated Process

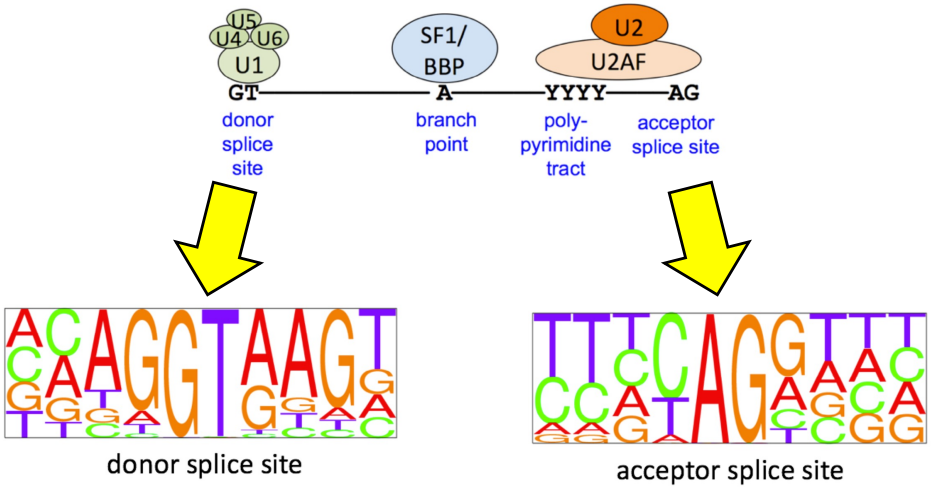


The splicing code and the protein code reside in the same space, and jointly influence the selective landscape for gene sequences.

See: Kornblihtt *et al.*, 2013 for full description and figures

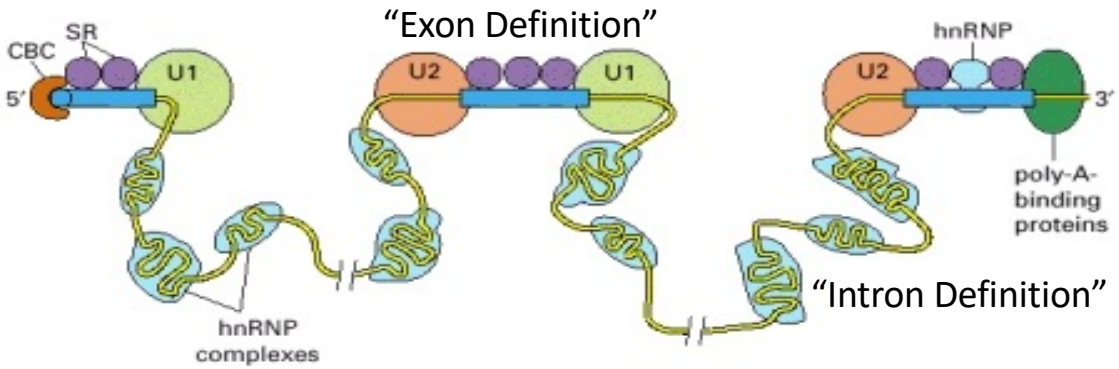
hnRNP = heterogeneous nuclear ribonucleoprotein
SR protein = serine-argenine-rich protein
ISS = intronic splicing silencer
ISE = intronic splicing enhancer
ESS = exonic splicing silencer
ESE = exonic splicing enhancer

Splicing is a Complex & Highly Regulated Process



See: Kornblihtt *et al.*, 2013 for full description and figures

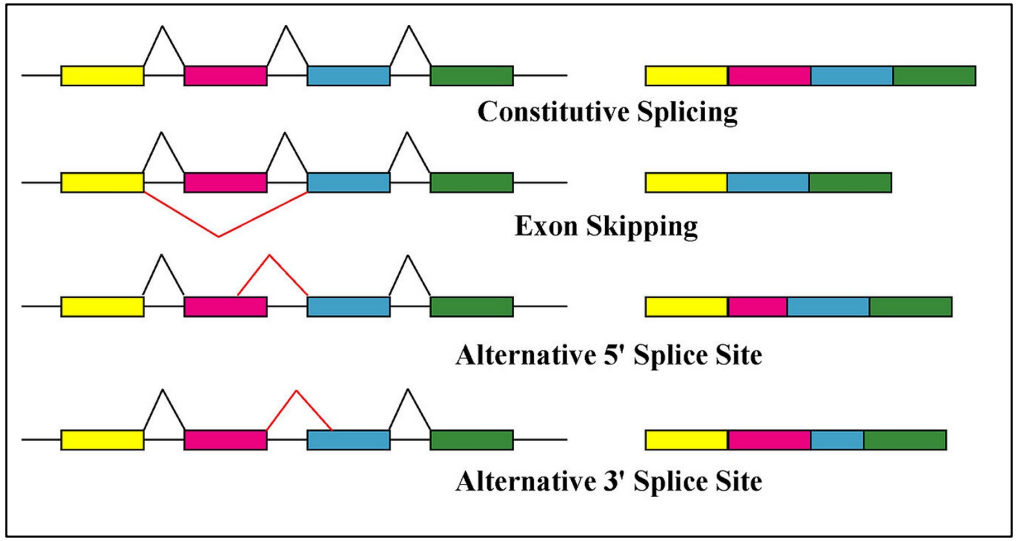
Splice sites alone are not sufficient to determine splicing patterns!



Alberts *et al.*, 2002

Splicing Regulation Can be Cell-type Specific

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



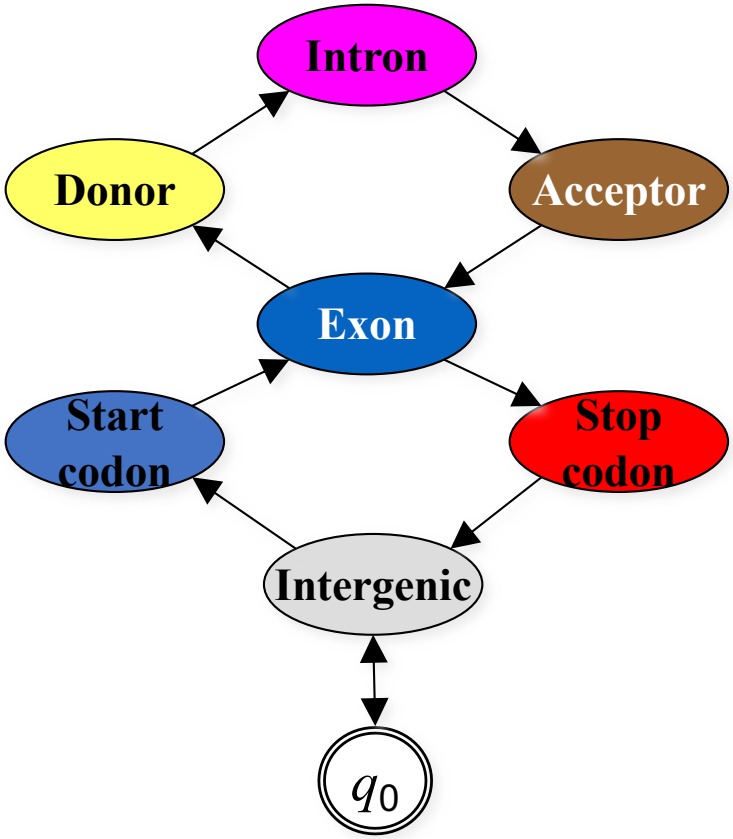
https://en.wikipedia.org/wiki/Protein_isoform

Outline

1. What is a Gene?
2. Defining Gene Structure
3. How Splicing Works
4. **Decoding Gene Structure**
5. For More Information

Statistical Models for Gene Prediction

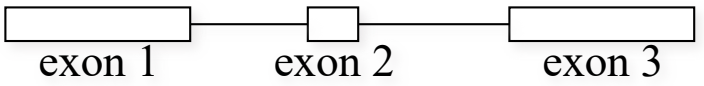
Gene structures can be predicted using statistical methods such as Hidden Markov Models (HMMs).



The input sequence: AGCTAGCAGT**ATG**TCATGGCAT**GT**TCGG**AG**GTA**GT**ACGTAG**AG**GTAGCTAGTA**TAG**GTCGATAGTA

The most probable path:

The gene prediction:



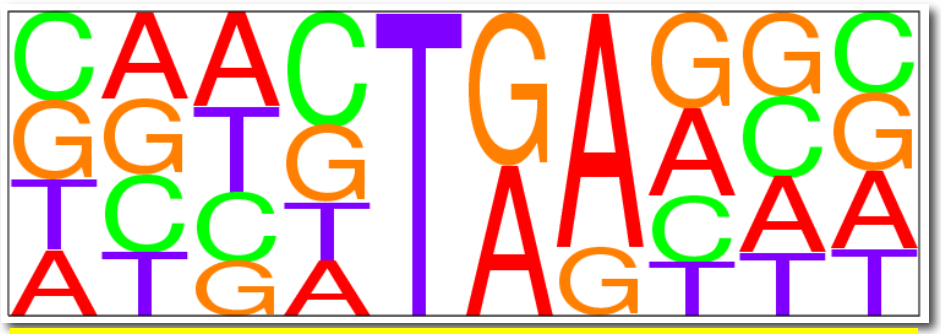
Sequence Signals

(start codons)

A T G

(stop codons)

T G A
T A A
T A G



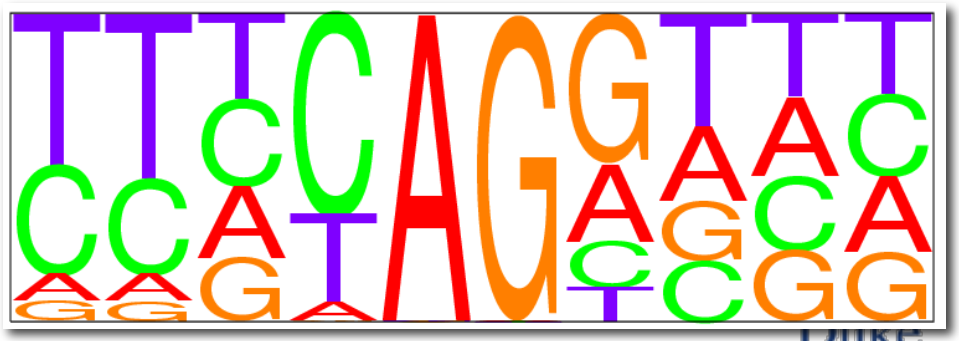
Also accepts TGG!

(donor splice sites)

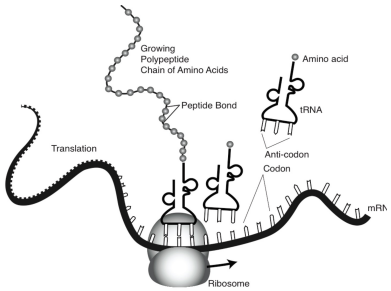
G T

(acceptor splice sites)

A G



The Genetic Code is Degenerate



Amino acid	DNA codons	Amino acid	DNA codons
Ala, A	GCU, GCC, GCA, GCG	**Ile, I**	AUU, AUC, AUA
Arg, R	CGU, CGC, CGA, CGG; AGA, AGG	**Leu, L**	CUU, CUC, CUA, CUG; UUA, UUG
Asn, N	AAU, AAC	**Lys, K**	AAA, AAG
Asp, B			
Cys, C	UGU, UGC	**Pro, P**	CCU, CCC, CCA, CCG
Gln, Q	CAA, CAG	**Ser, S**	UCU, UCC, UCA, UCG; AGU, AGC
Glu, E	GAA, GAG	**Thr, T**	ACU, ACC, ACA, ACG
Gln or Glu, Z	CAA, CAG; GAA, GAG	**Trp, W**	UGG
Gly, G	GGU, GGC, GGA, GGG	**Tyr, Y**	UAU, UAC
His, H	CAU, CAC	**Val, V**	GUU, GUC, GUA, GUG
START	AUG	**STOP**	UAA, UGA, UAG

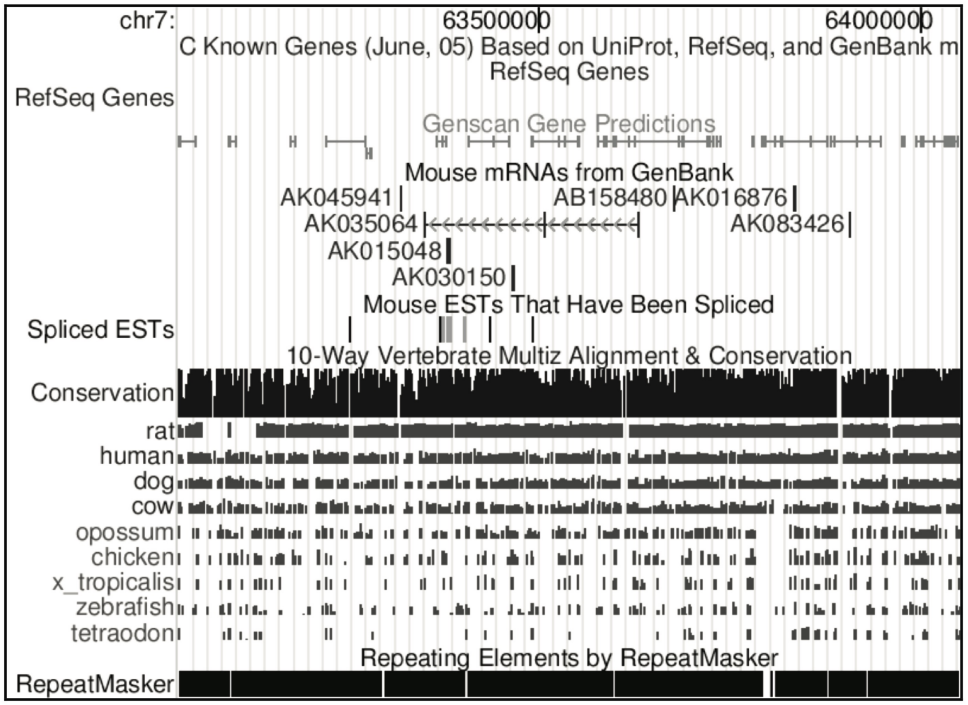
Not all codons occur with the same frequency! This **codon bias** is a major statistical signal we can use in identifying genes and their reading frames.

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

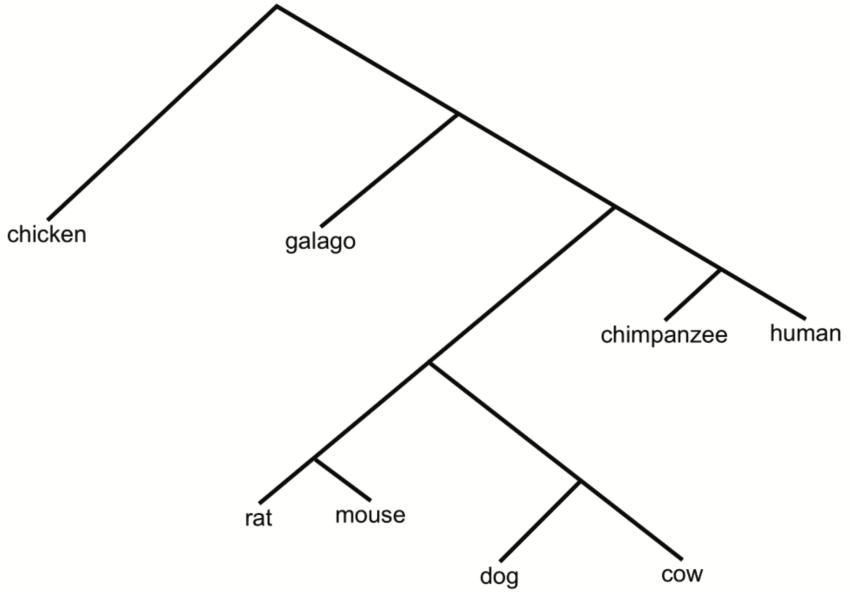
Evolutionary Conservation

Many genes are conserved over evolutionary time.

We can use this fact by observing patterns of conservation across related species.

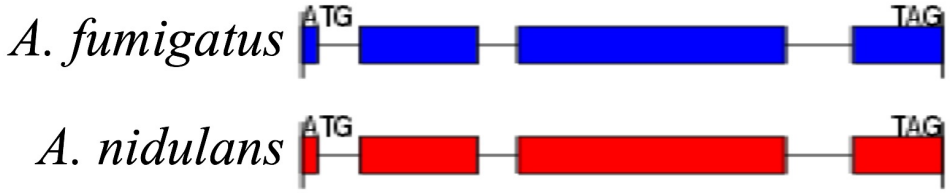


UCSC Genome Browser showing a region of the mouse genome.



Levels of Conservation

As another example, consider this gene that is conserved between two species of fungi:



In this gene's exons, amino acid conservation is higher than the nucleotide conservation (and vice-versa for introns):

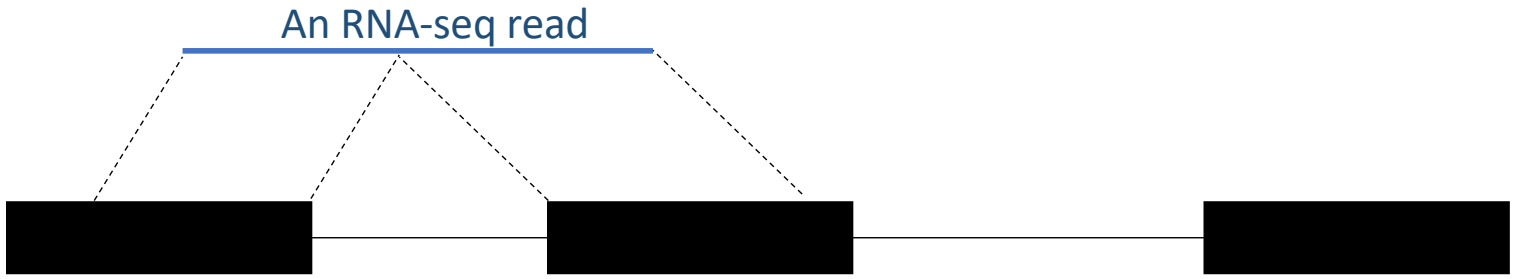
feature	amino acid conservation		nucleotide conservation
exon 1	100%	>	71%
intron 1	14%	<	51%
exon 2	98%	>	85%
intron 2	29%	<	49%
exon 3	97%	>	82%
intron 3	9%	<	49%
exon 4	96%	>	83%

Note that introns do contain functional elements related to the regulation of splicing, so they show some conservation.

This is because in protein-coding genes, the primary action of natural selection is to conserve protein function.

Deducing Gene Structure Experimentally

We can align RNA-seq reads to the genome (using spliced-alignment) to identify gene structures:



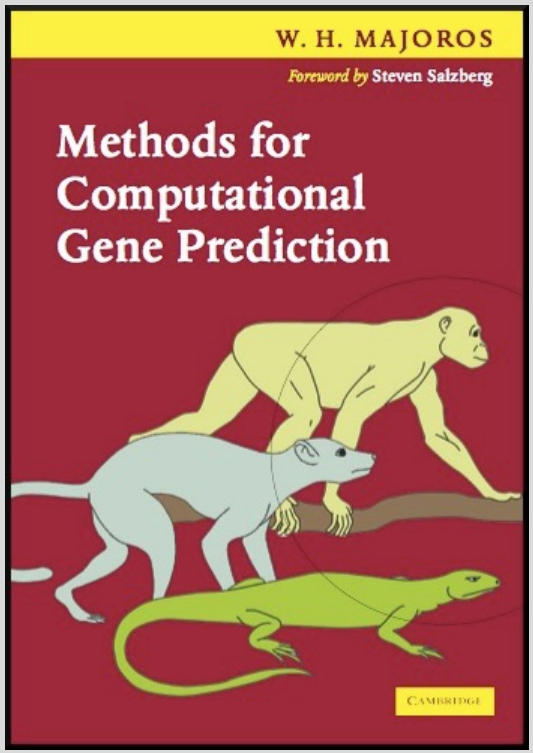
Each spliced read gives us evidence of a splice junction (intron).

Some spliced aligners: STAR, TopHat

Outline

1. What is a Gene?
2. Defining Gene Structure
3. How Splicing Works
4. Decoding Gene Structure
5. **For More Information**

For More Info



Methods for Computational Gene Prediction
by W.H. Majoros
with a foreword by Steven L. Salzberg

CAMBRIDGE UNIVERSITY PRESS

1. Introduction
2. Mathematical Preliminaries
3. Overview of Gene Prediction
4. Gene Finder Evaluation
5. A Toy Exon Finder
6. Hidden Markov Models
7. Signal and Content Sensors
8. Generalized Hidden Markov Models
9. Comparative Gene Finding
10. Machine Learning Methods
11. Tips and Tricks
12. Advanced Topics

References

- Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol. 2013 Mar;14(3):153-65. doi: 10.1038/nrm3525. Epub 2013 Feb 6. Erratum in: Nat Rev Mol Cell Biol. 2013 Mar;14(3). doi:10.1038/nrm3560. PMID: 23385723.

Questions?