A Zero-Knowledge Based Introduction to Biology

Sandeep Chinchali, Jim Notwell

26 September 2014

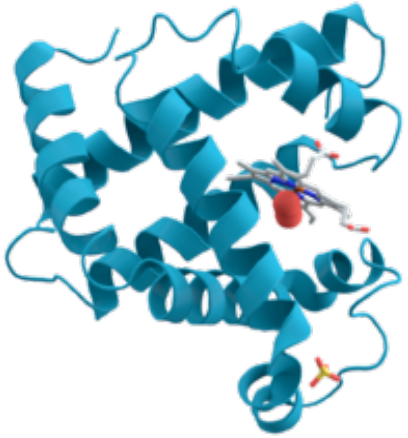
Q: What is your genome?

Q: What is your genome?

A: The sum of your hereditary information.

Human Genome

- 3 billion base pairs: A,T,G,C
- Full DNA sequence in virtually all cells
- DNA is the ***blueprint for life***:
  - Cookbook with many “recipes” for proteins - genes
  - Proteins do most of the work in biology
  - *Yet, only ~2% of the genome is protein-coding genes!*



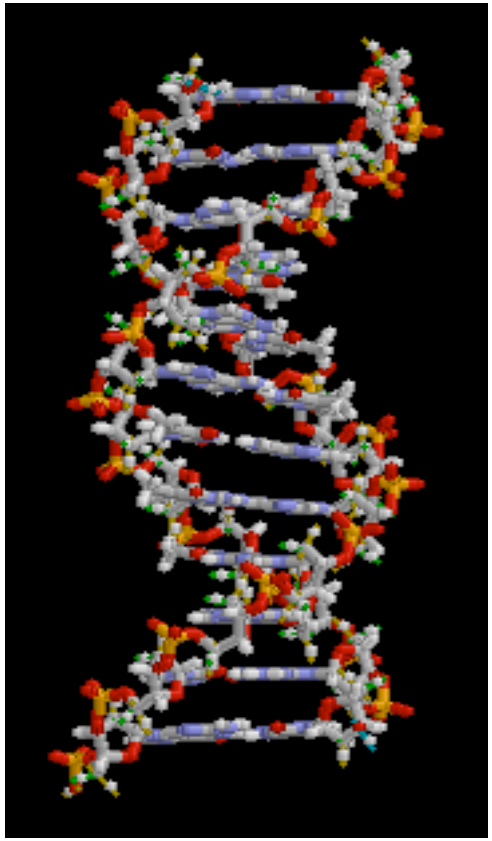
What does the rest of the genome do?

- 3 billion base pairs – 2% coding, 5-10% regulatory
- Organism's complexity NOT correlated with number of genes!
- Human (20-25k genes) vs. Rice (51k genes)
- 1 million Regulatory elements (switches) enable:
  - Precise control for turning genes on/off
  - Diverse cell types (lung, heart, skin)

- *Analogy: Making specific recipes (genes) from a large cookbook (genome) at a given time*

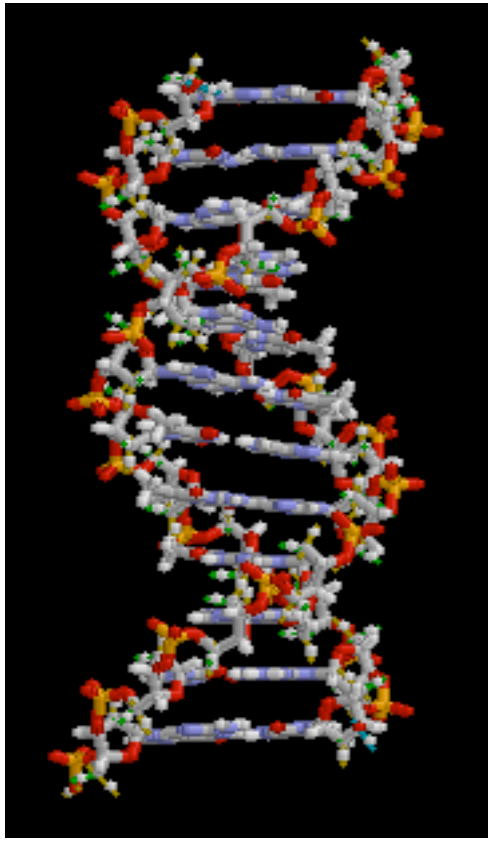
Quick Recap

DNA: "Blueprints" for a cell



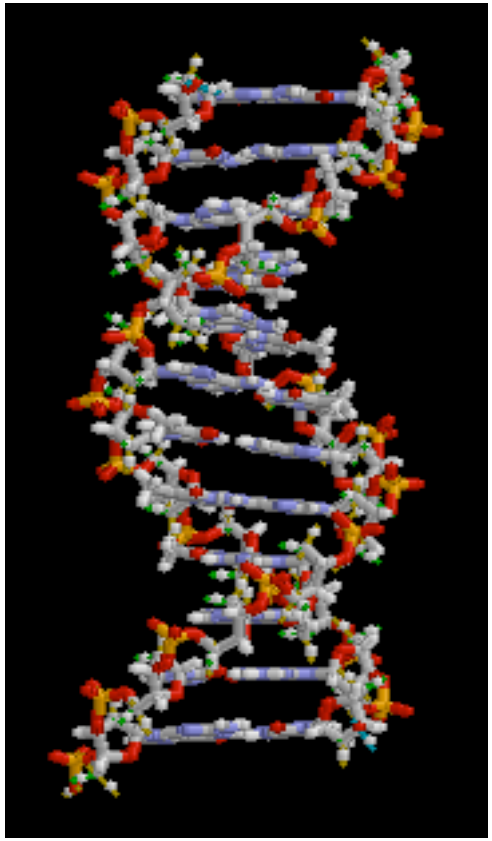
- Genetic information encoded in long strings
- Deoxyribonucleic acid (DNA) comes in four bases: adenine (A), thymine (T), guanine (G) , and cytosine (C)

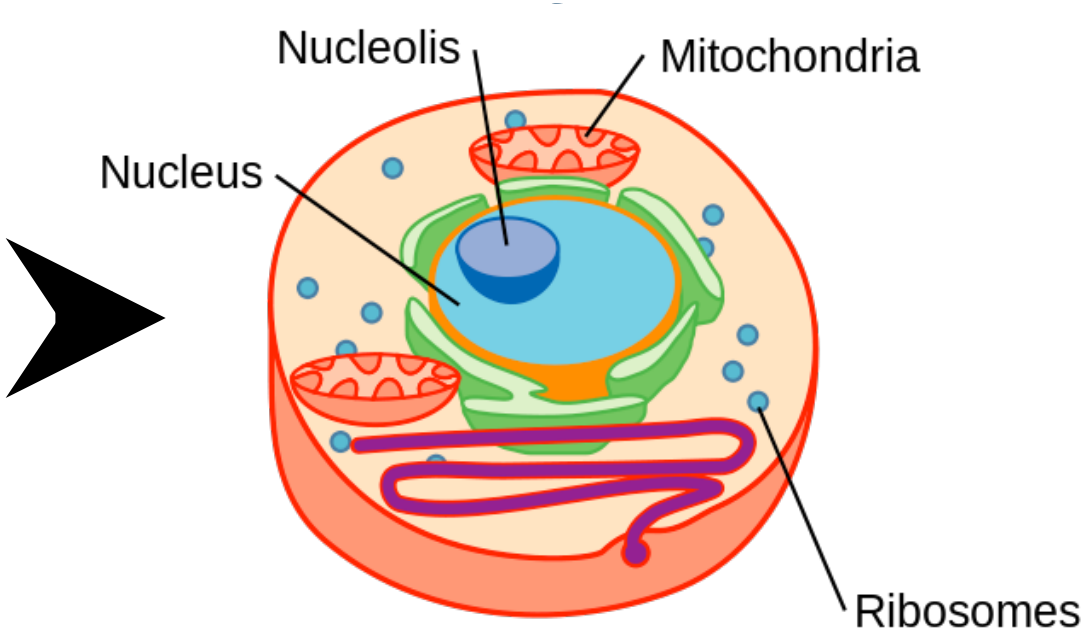
From DNA to Organism



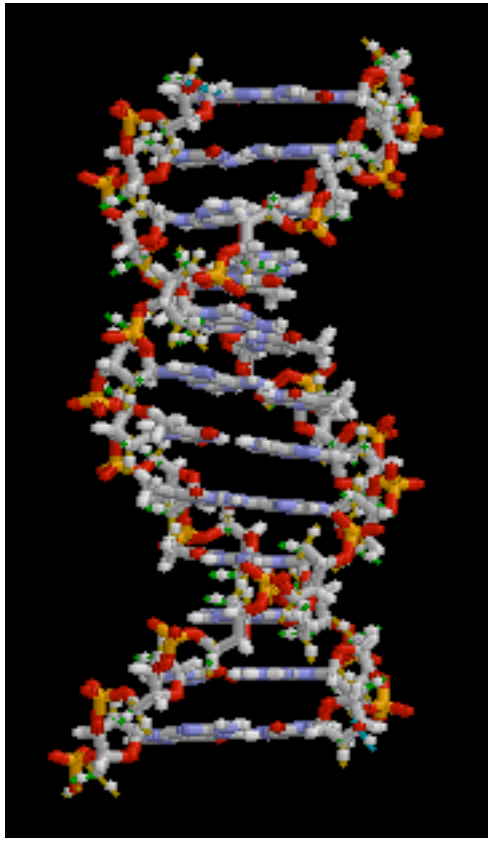
You are composed of ~ 10 trillion cells

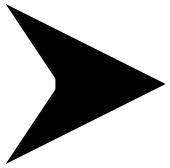
From DNA to ~~Organism~~ Cell

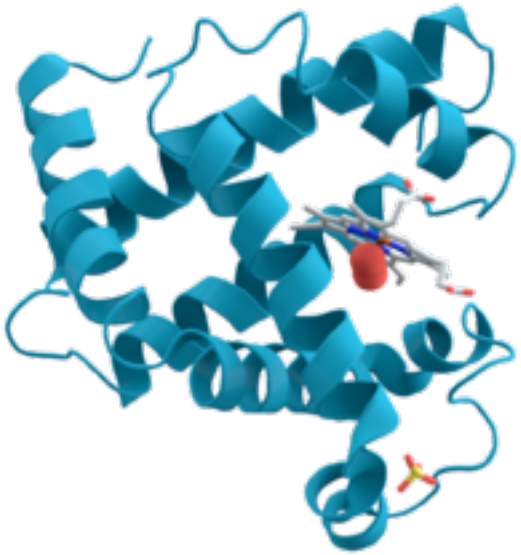




From DNA to ~~Organism~~ ~~Cell~~ Protein





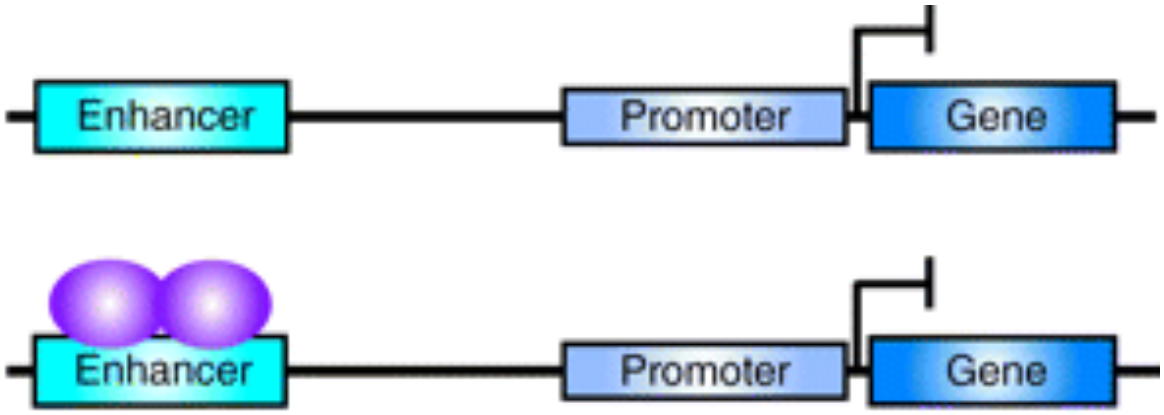


Proteins do most of the work in biology

Q: How does ***one genome*** encode a variety of cell types in a complex organism?

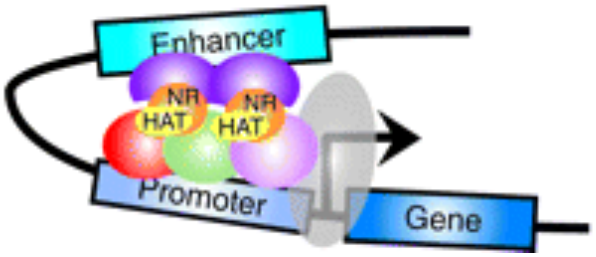
Regulatory Elements

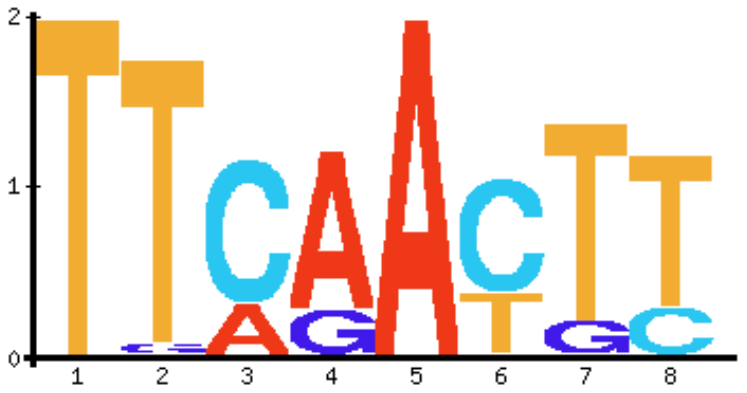
- ~ 20-25k genes
  - Expression Modulated by Regulatory elements
    - Enhancer, Promoters, Silencers
- CS analogy:
  - Genes are like variable assignments (a = 7)
  - Regulatory elements are control flow, complex logic



Controlling Gene Expression

- Transcription factors (TFs):
  - Proteins that recognize sequence motifs in enhancers, promoters
  - Combinatorial switches that turn genes on/off





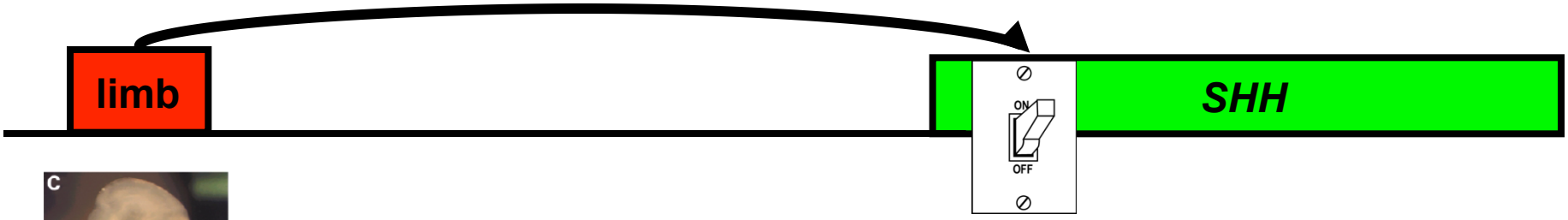
How does the genome influence human disease?

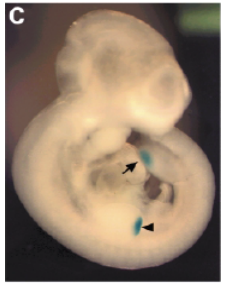
Disease Implications

MUTATIONS

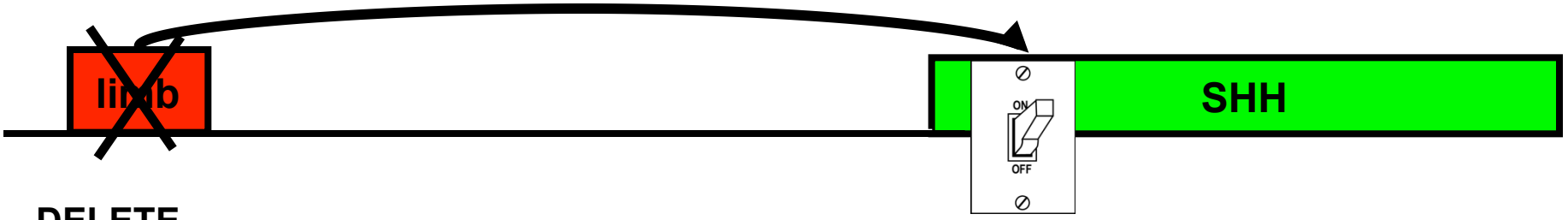
- **Brain**
- **Limb**
- **Other**

Limb Enhancer 1Mb away from Gene



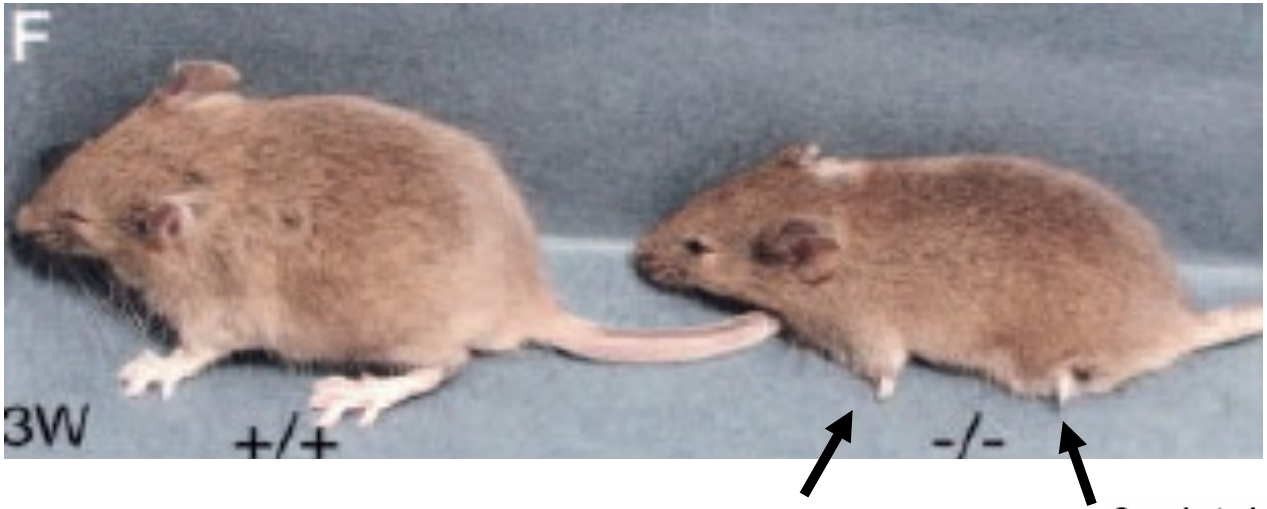


Enhancer Deletion



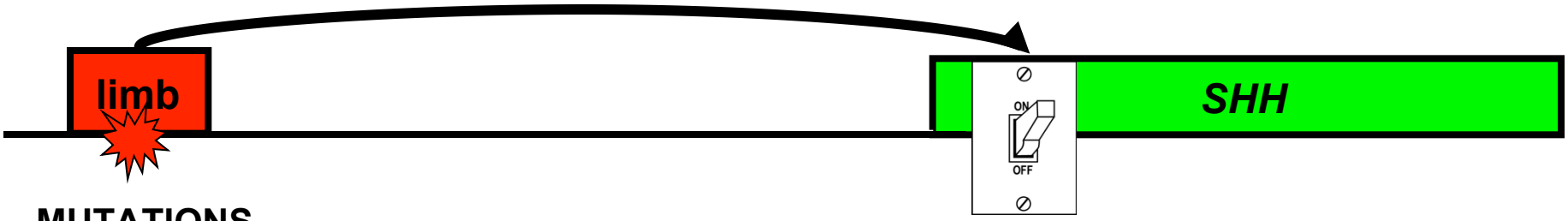
DELETE

- **Limb**



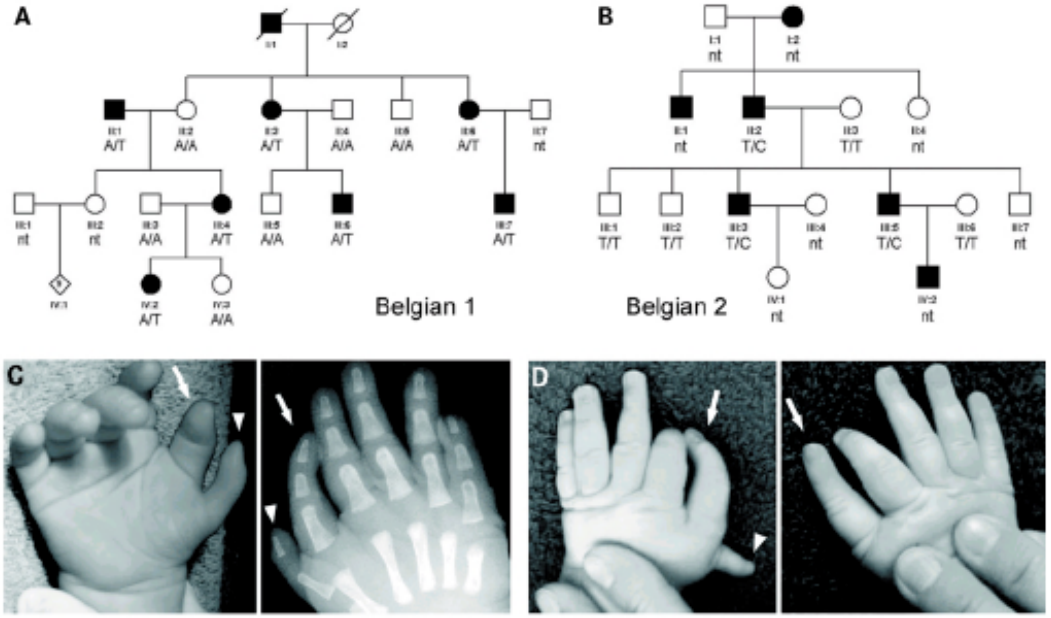
Sagai et al. *Development* 2005 132: 797-803

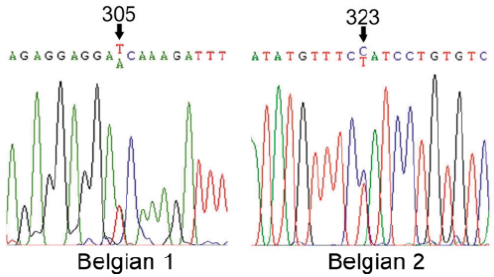
Enhancer 1bp Substitution



MUTATIONS

- **Limb**



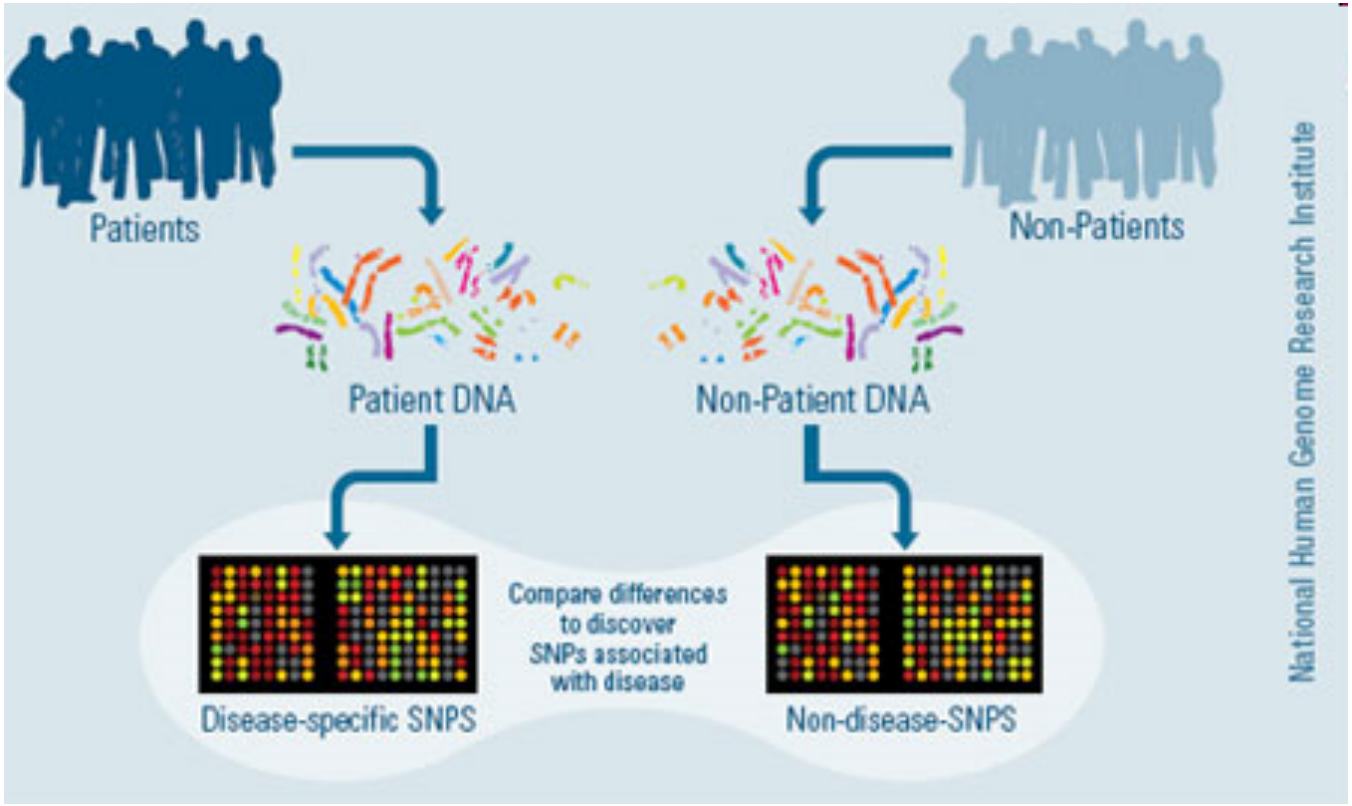


Lettice et al. *HMG* 2003 12: 1725-35

Genome Wide Assocation Study (GWAS):

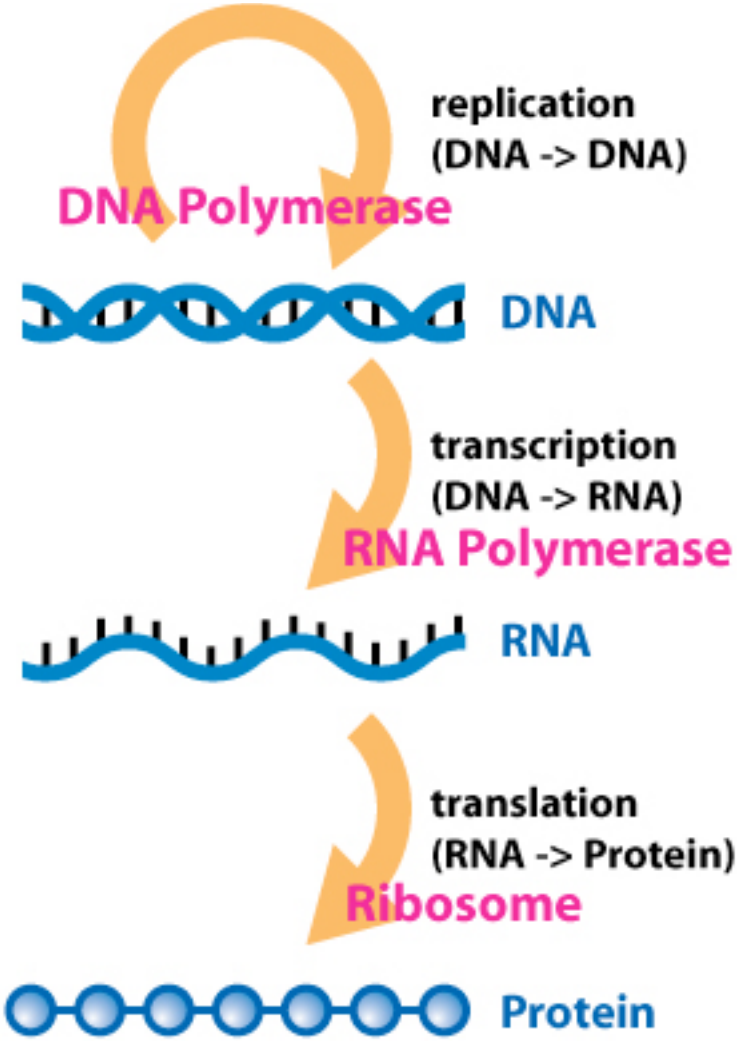
80% of GWAS SNPs are noncoding (hard to interpret)

Active area of research

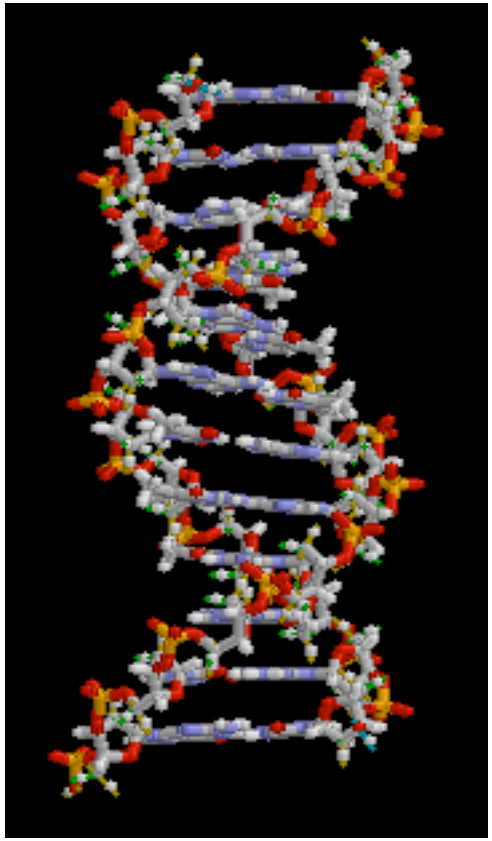


How exactly do genes code for proteins?

Central Dogma of Biology



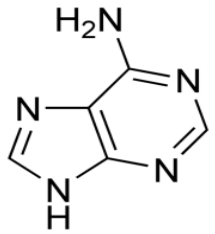
DNA: "Blueprints" for a cell



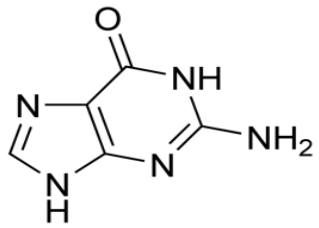
- Genetic information encoded in long strings
- Deoxyribonucleic acid comes in four bases: adenine, thymine, guanine, and cytosine

Nucleobase Complementary Pairing

purines



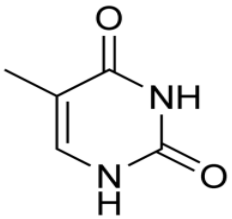
Adenine (A)

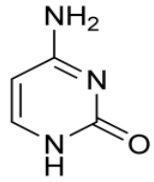


Guanine (G)

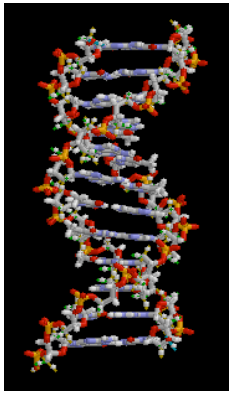
Thymine (T)

Cytosine (C)

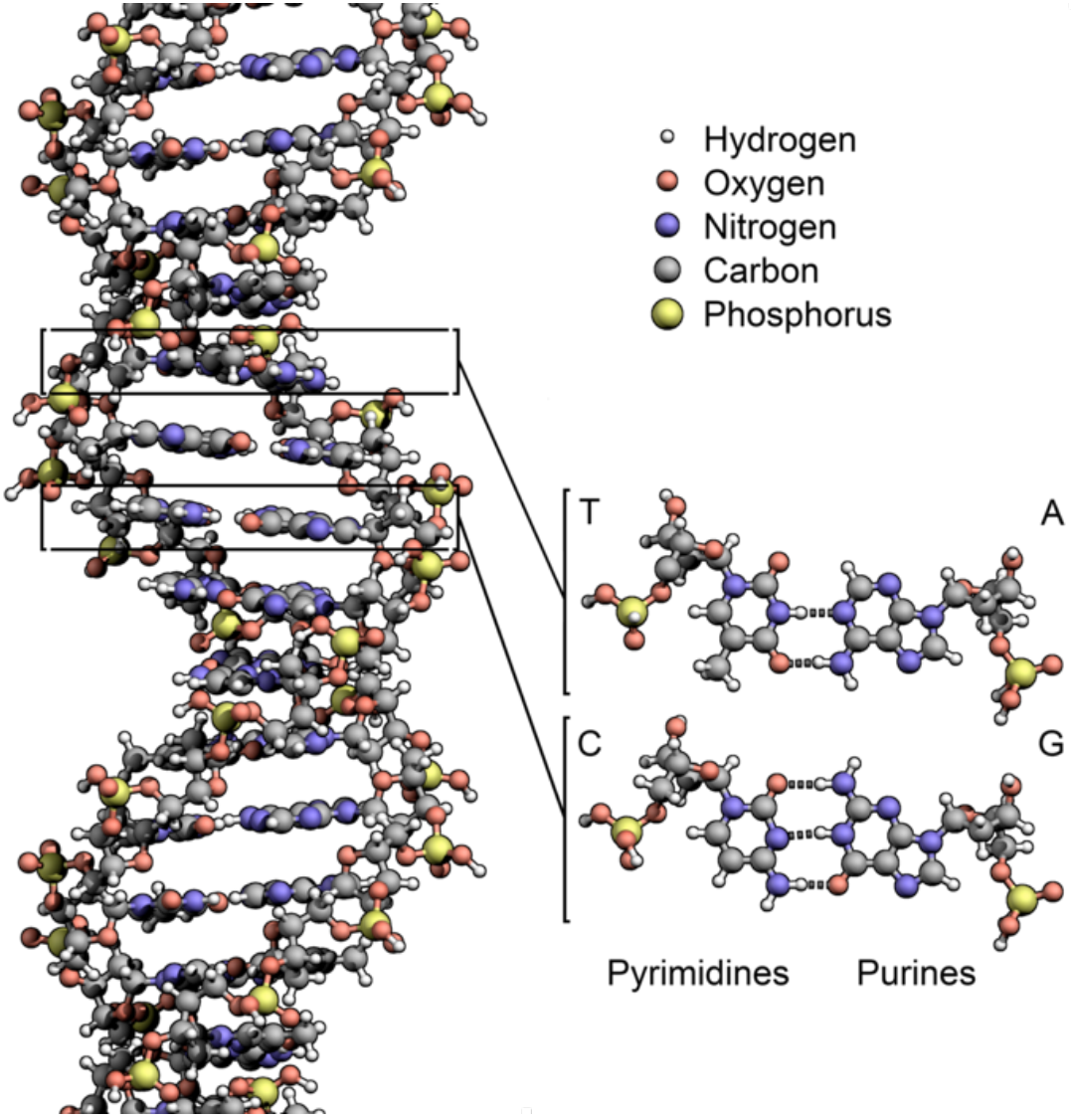




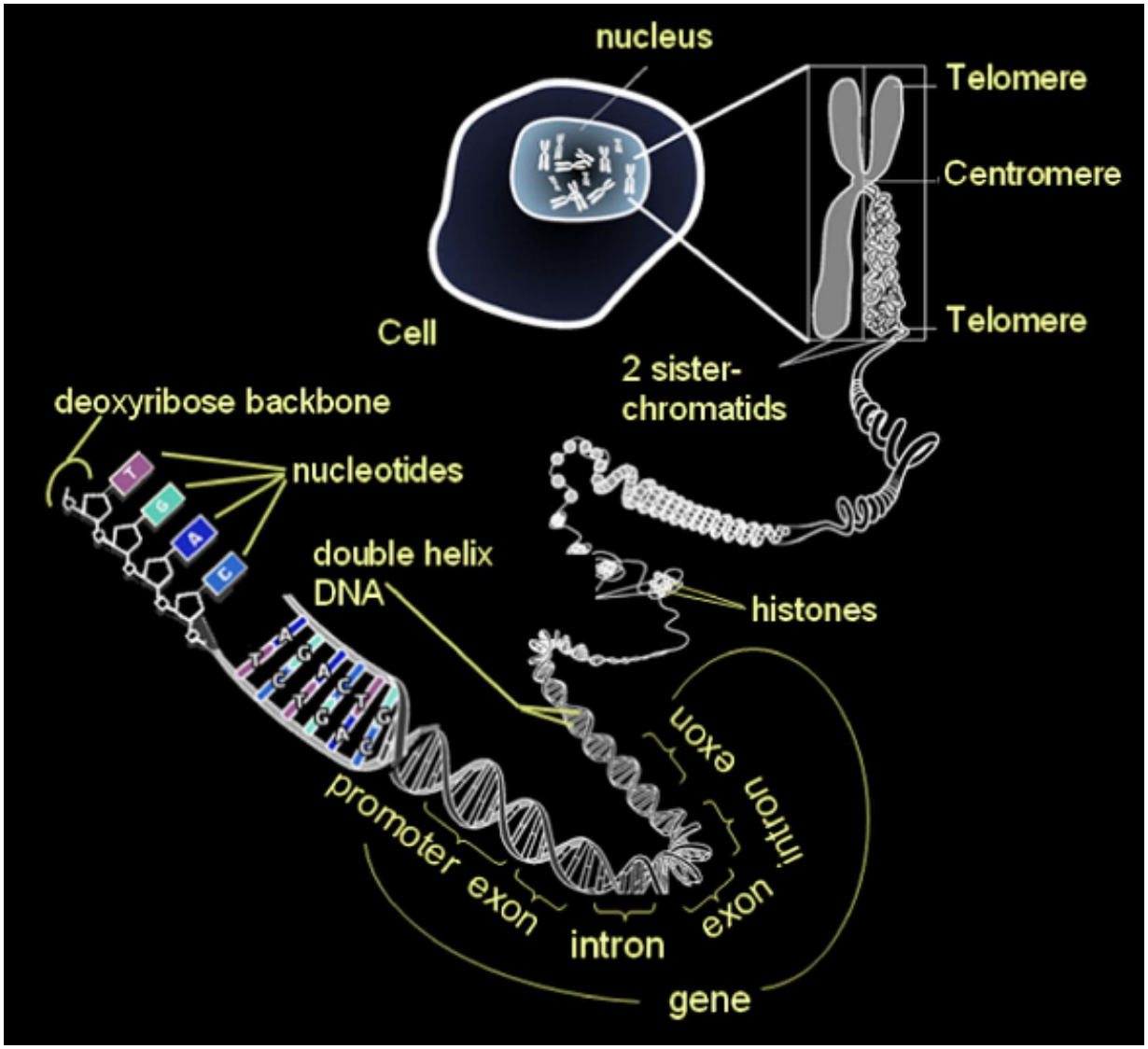
pyrimidines



DNA Double Helix



DNA Packaging



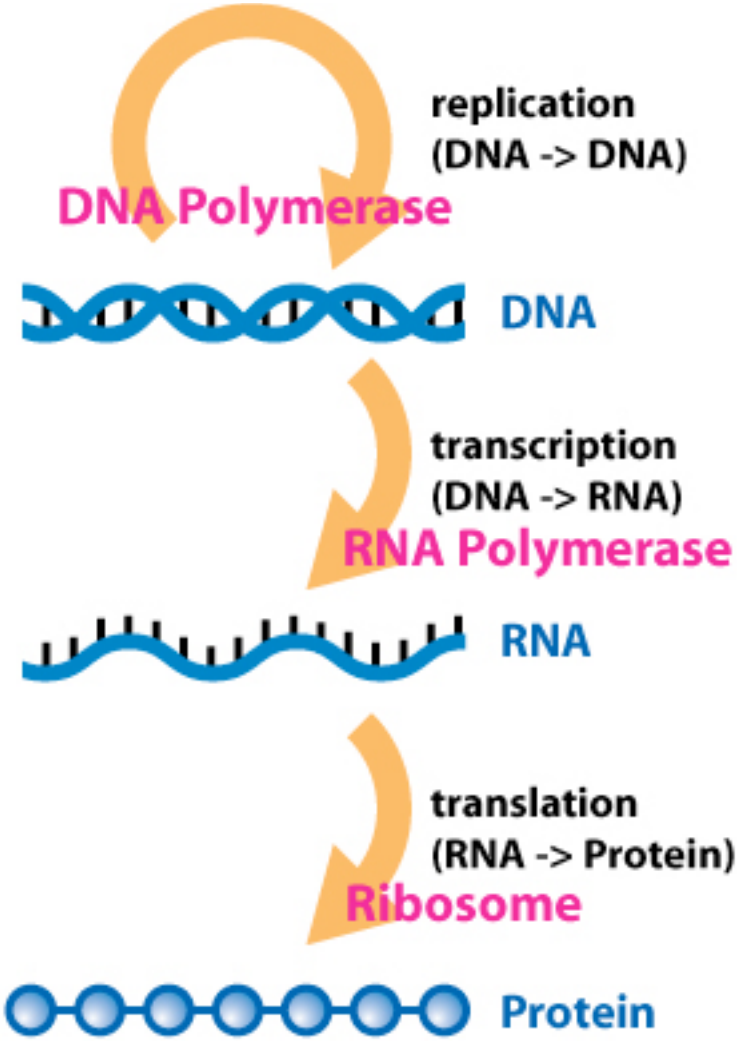
Q: What is your genome?

A:The sum of your hereditary information.

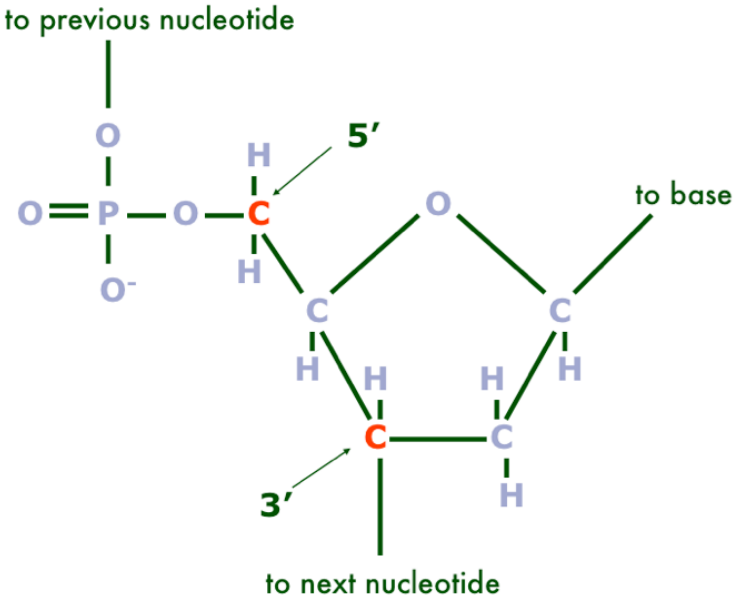
Q: What is your genome?

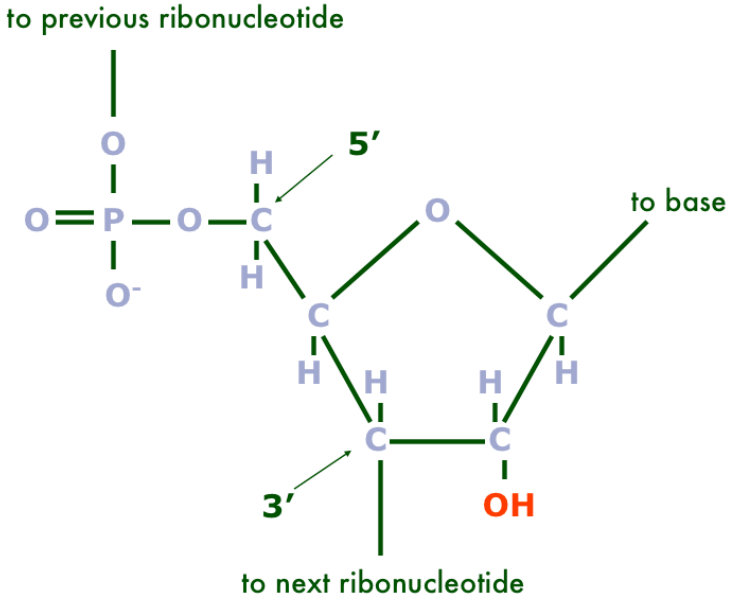
A:The sum of your hereditary information. Humans bundle two copies of the genome into 46 chromosomes in **every cell**

Central Dogma of Biology



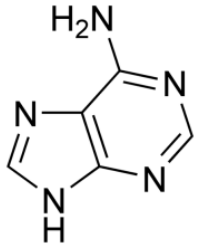
DNA vs RNA


to previous nucleotide
5′
to base
3′
to next nucleotide

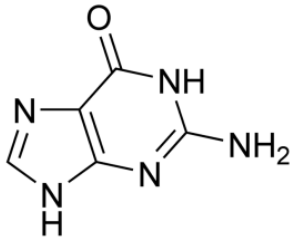

to previous ribonucleotide
5′
to base
3′
OH
to next ribonucleotide

RNA Nucleobases

purines



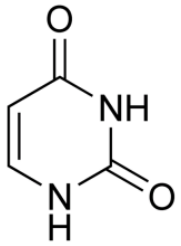
Adenine (A)

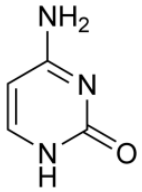


Guanine (G)

Uracil (U)

Cytosine (C)



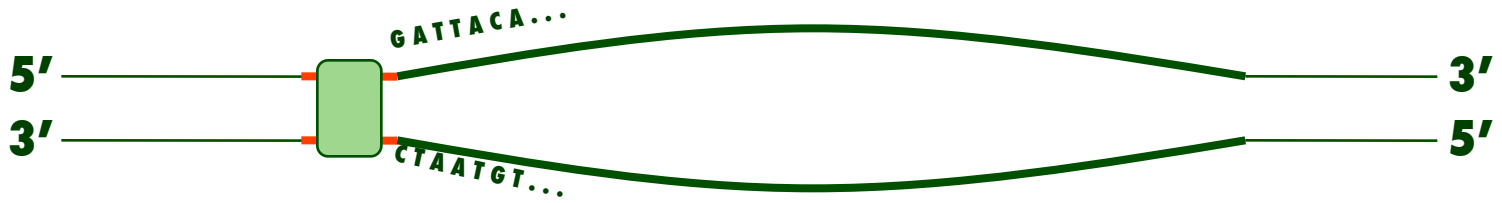


pyrimidines

Gene Transcription

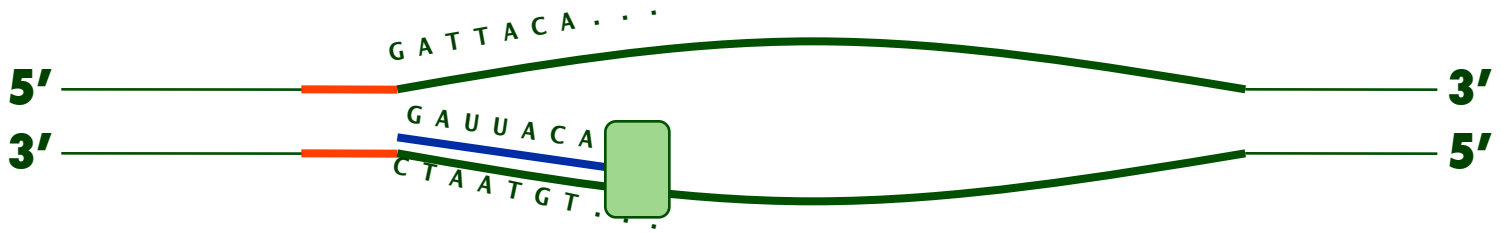
Gene Transcription

Gene Transcription



Strands are separated (DNA helicase)

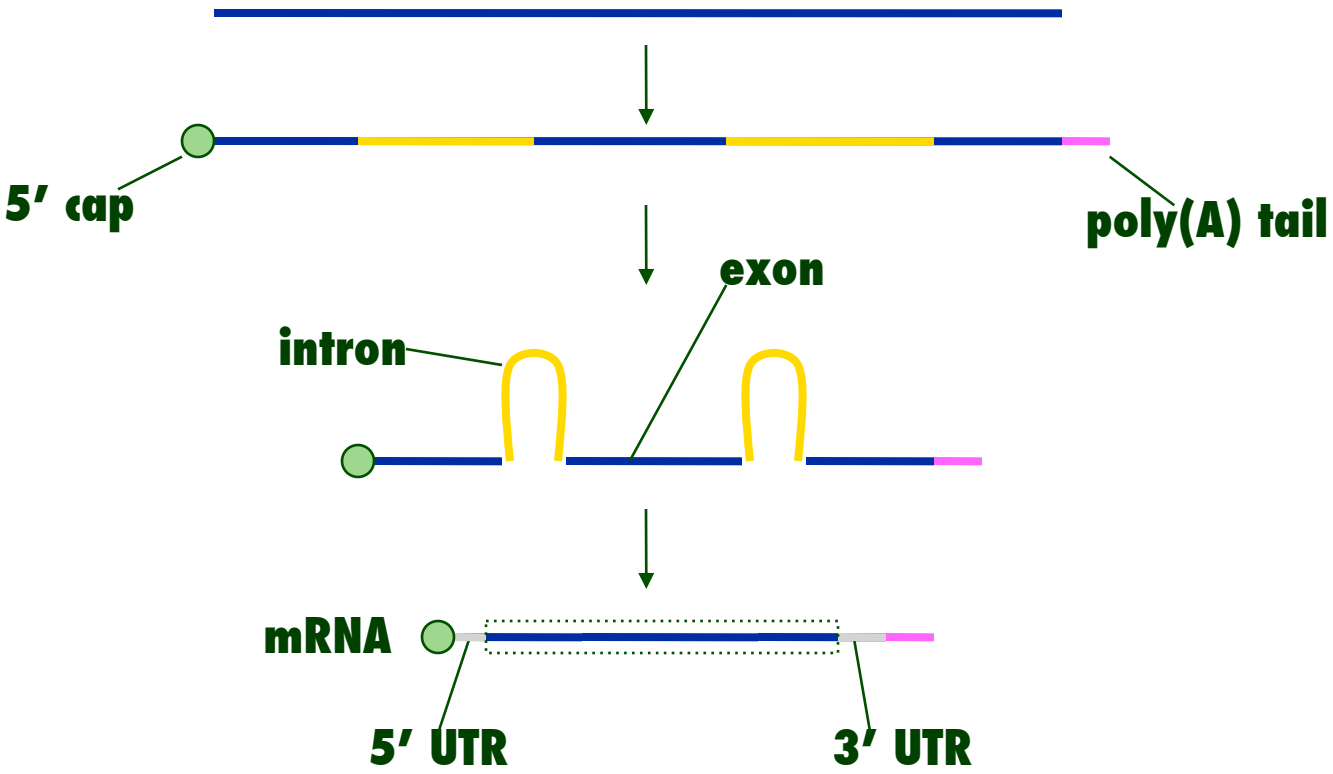
Gene Transcription



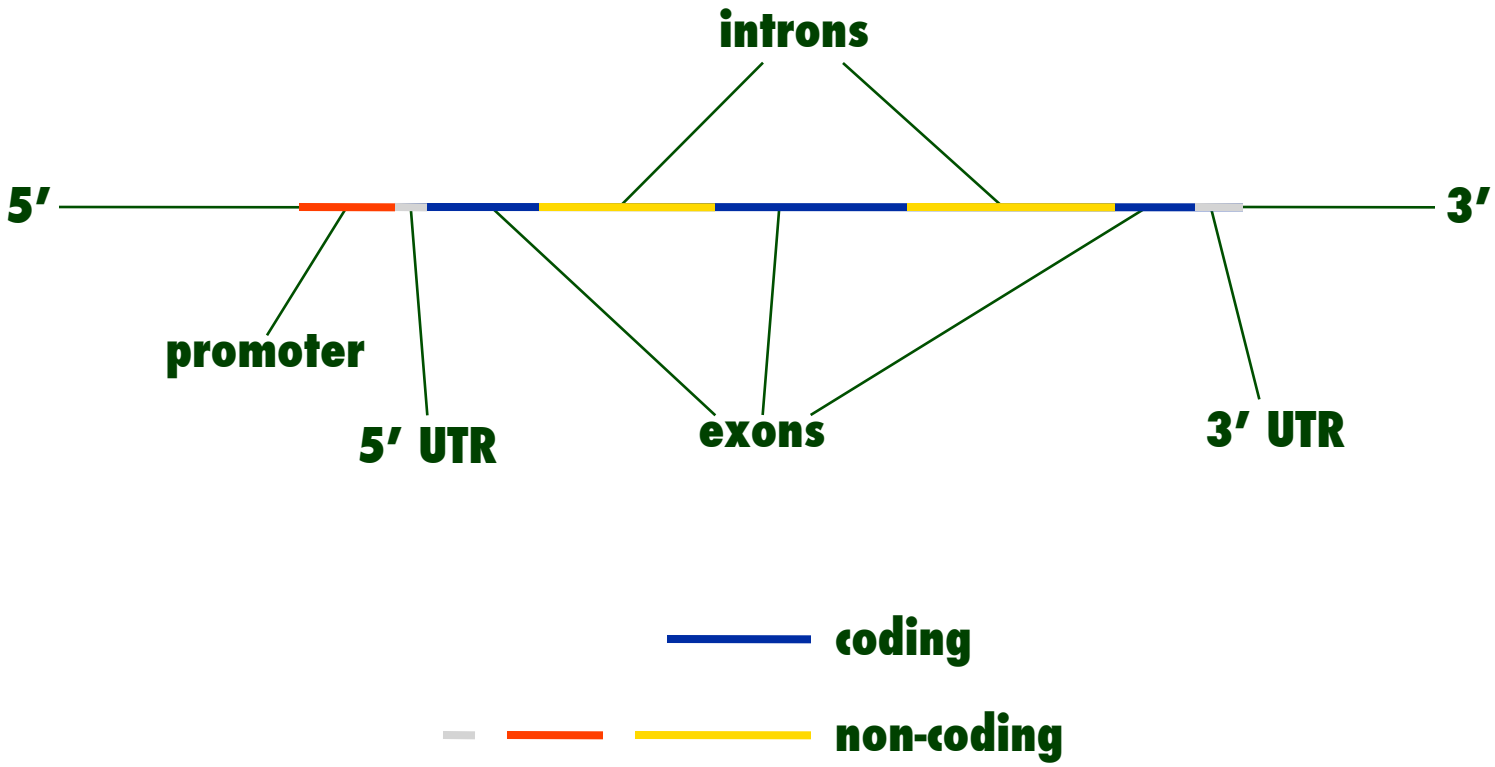
An RNA copy of the 5'→3' sequence is created from the 3'→5' template

Gene Transcription

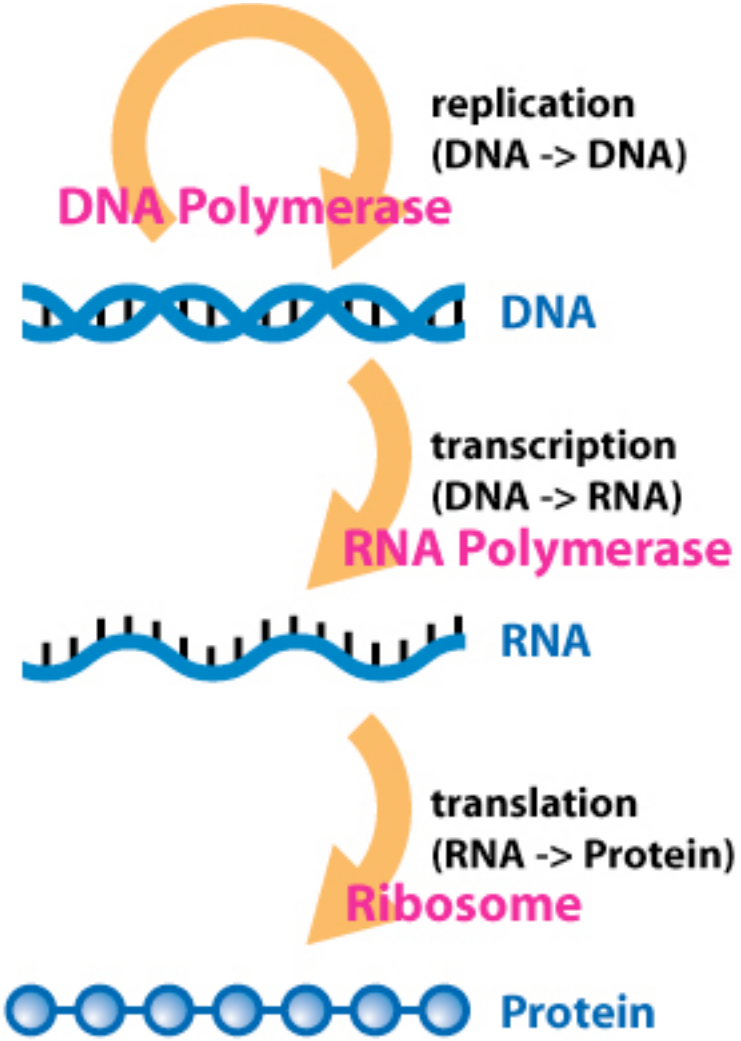
RNA Processing



Gene Structure



Central Dogma of Biology



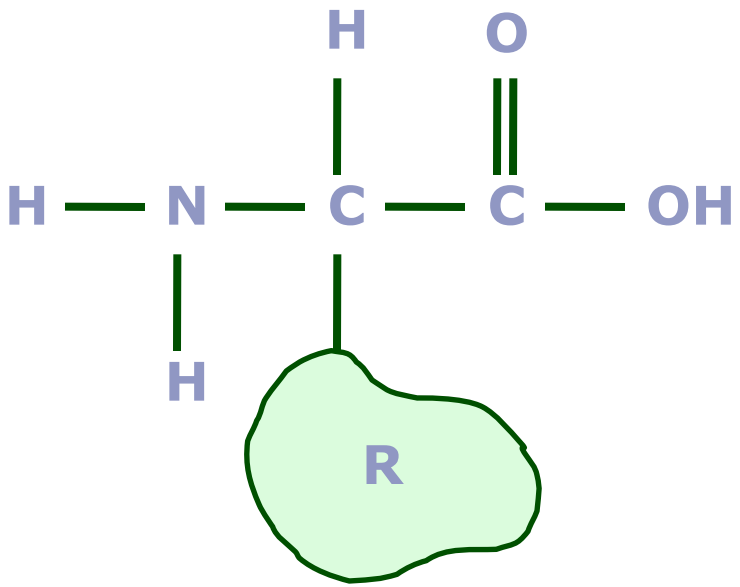
From RNA to Protein

- Proteins are long strings of amino acids joined by peptide bonds

- Translation from RNA sequence to amino acid sequence performed by ribosomes

- 20 amino acids → 3 RNA letters required to specify a single amino acid

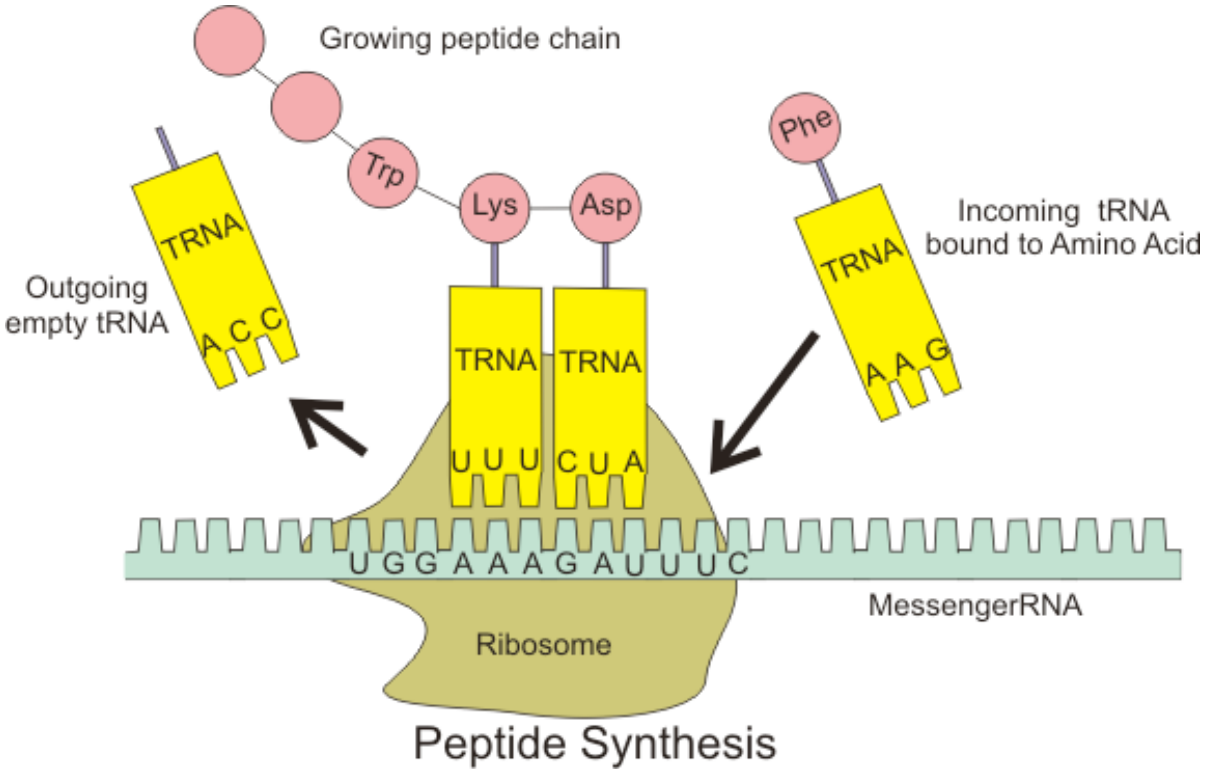
Amino Acid



There are 20 standard amino acids

Alanine
Arginine
Asparagine
Aspartate
Cysteine
Glutamate
Glutamine
Glycine
Histidine
Isoleucine
Leucine
Lysine
Methionine
Phenylalanine
Proline
Serine
Threonine
Tryptophan
Tyrosine
Valine

Translation

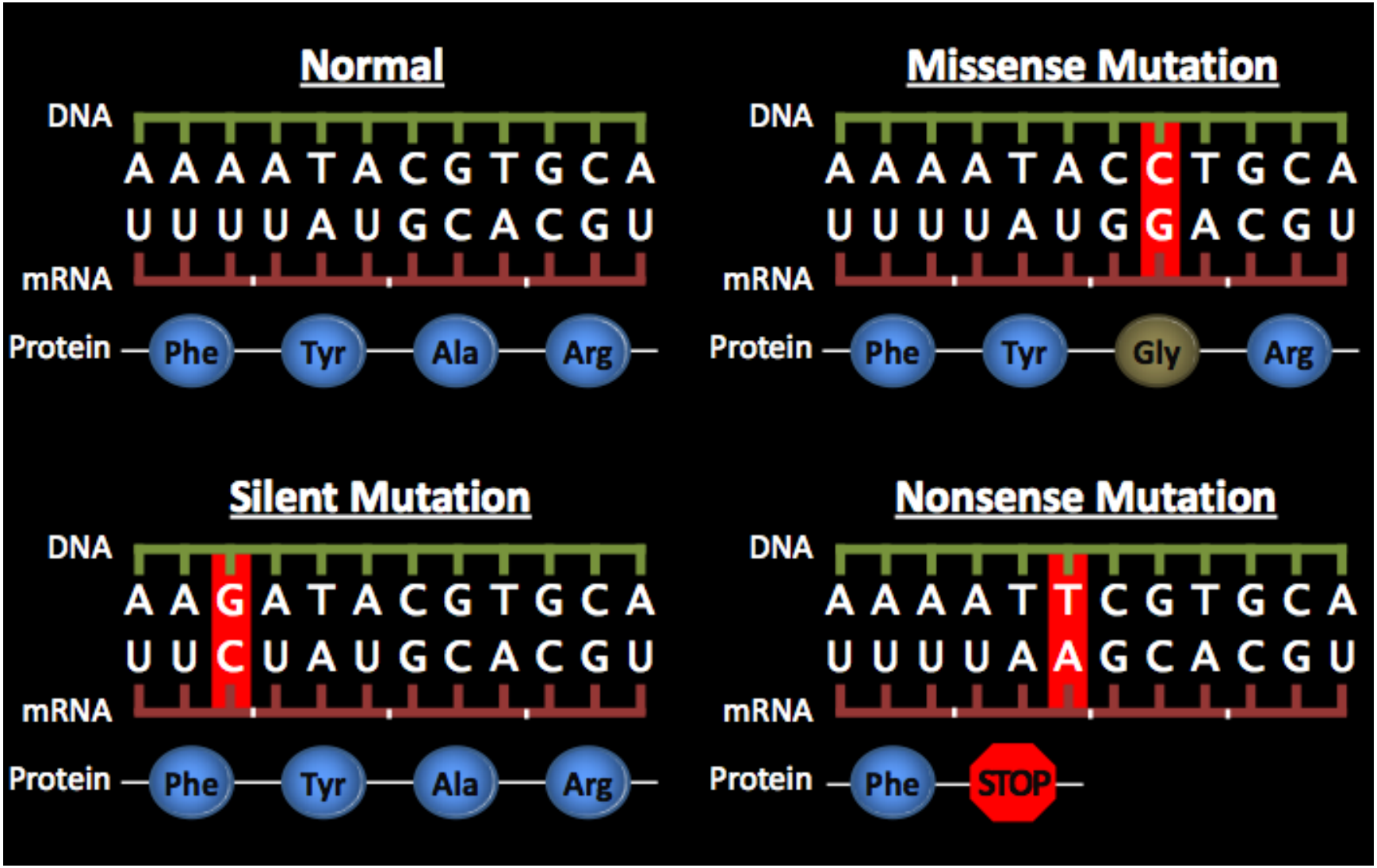


The ribosome (a complex of protein and RNA) synthesizes a protein by reading the mRNA in triplets (codons). Each codon is translated to an amino acid.

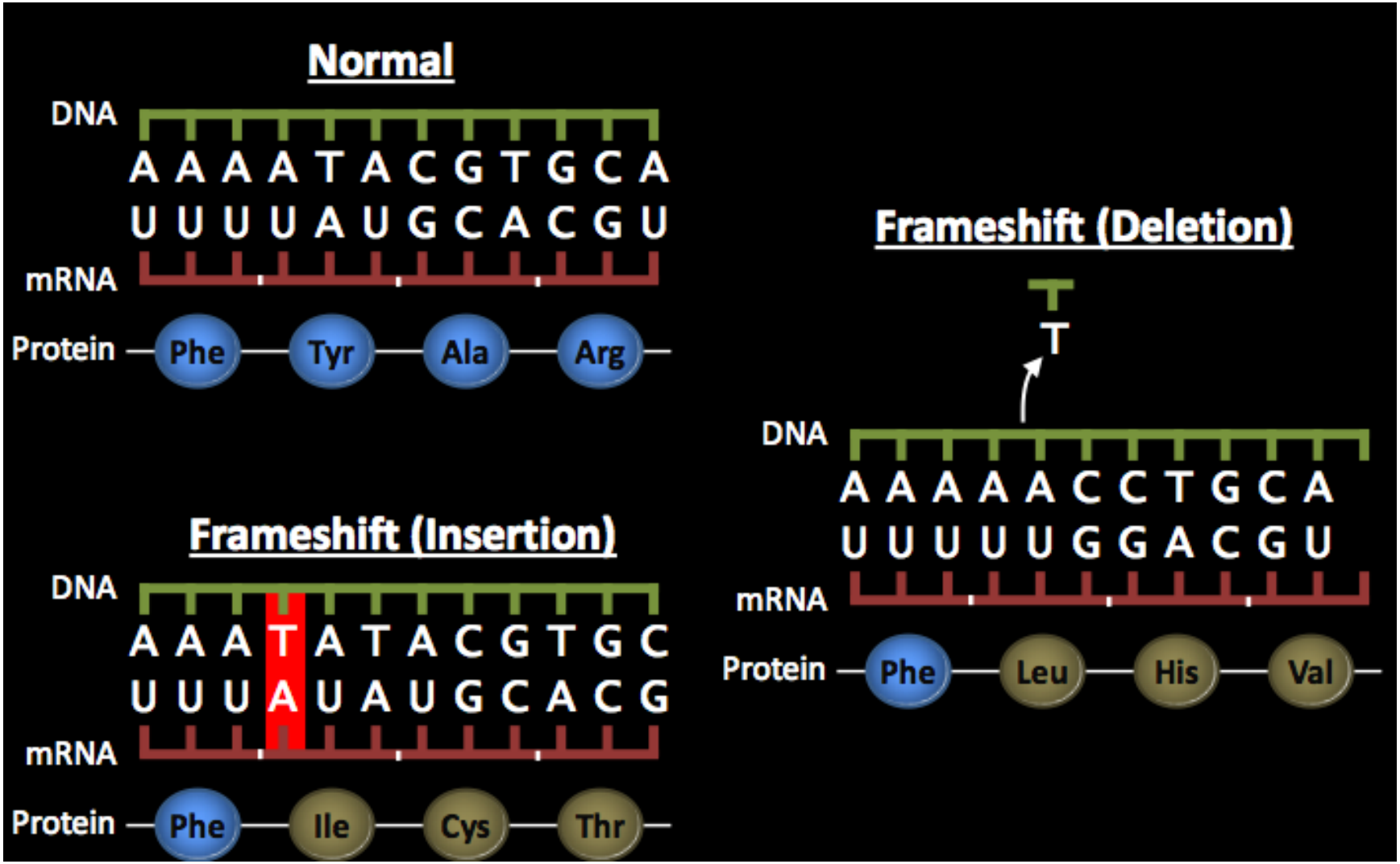
Translation

	U	C	A	G	
U	UUU **Phenylalanine** (Phe)	UCU **Serine** (Ser)	UAU **Tyrosine** (Tyr)	UGU **Cysteine** (Cys)	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA **Leucine** (Leu)	UCA Ser	UAA **STOP**	UGA **STOP**	A
	UUG Leu	UCG Ser	UAG **STOP**	UGG **Tryptophan** (Trp)	G
C	CUU **Leucine** (Leu)	CCU **Proline** (Pro)	CAU **Histidine** (His)	CGU **Arginine** (Arg)	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA **Glutamine** (Gln)	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU **Isoleucine** (Ile)	ACU **Threonine** (Thr)	AAU **Asparagine** (Asn)	AGU **Serine** (Ser)	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA **Lysine** (Lys)	AGA **Arginine** (Arg)	A
	AUG **Methionine** (Met) or **START**	ACG Thr	AAG Lys	AGG Arg	G
G	GUU **Valine** (Val)	GCU **Alanine** (Ala)	GAU **Aspartic acid** (Asp)	GGU **Glycine** (Gly)	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA **Glutamic acid** (Glu)	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

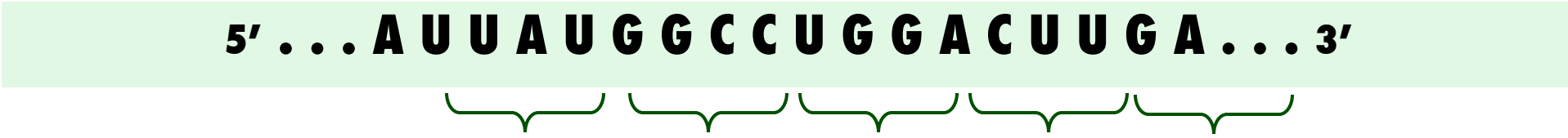
Single Nucleotide Changes



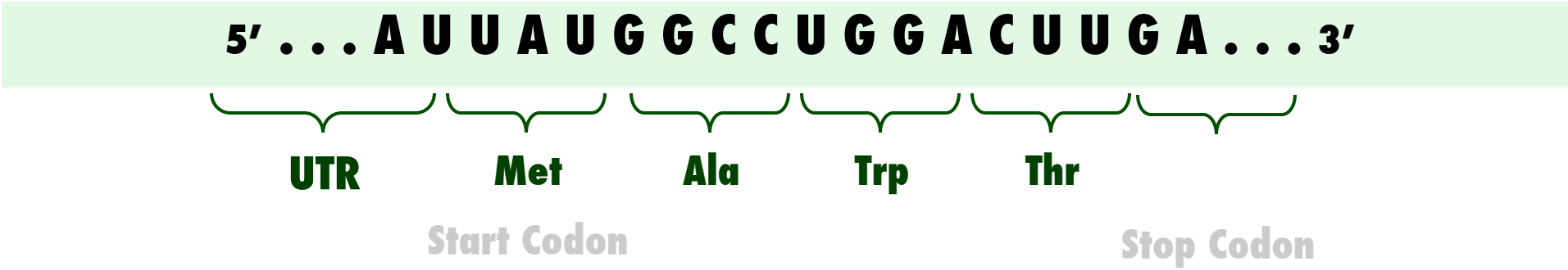
Single Nucleotide Changes



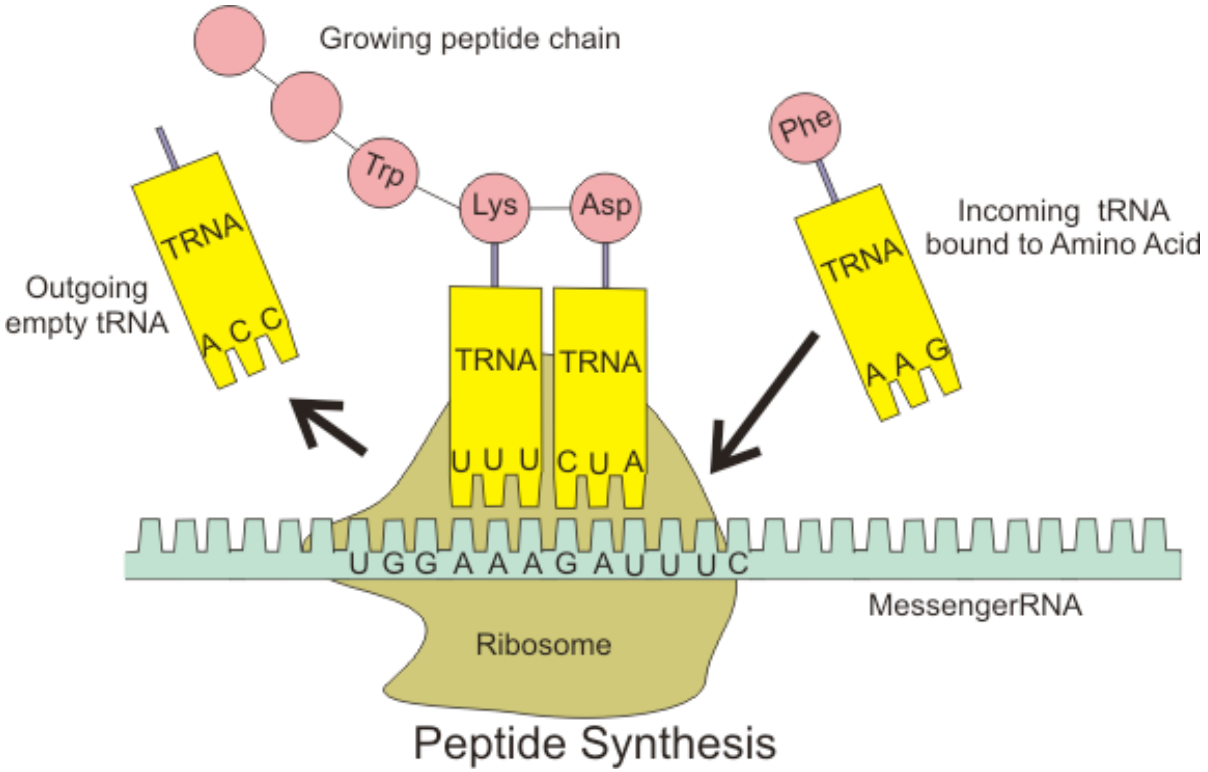
Translation



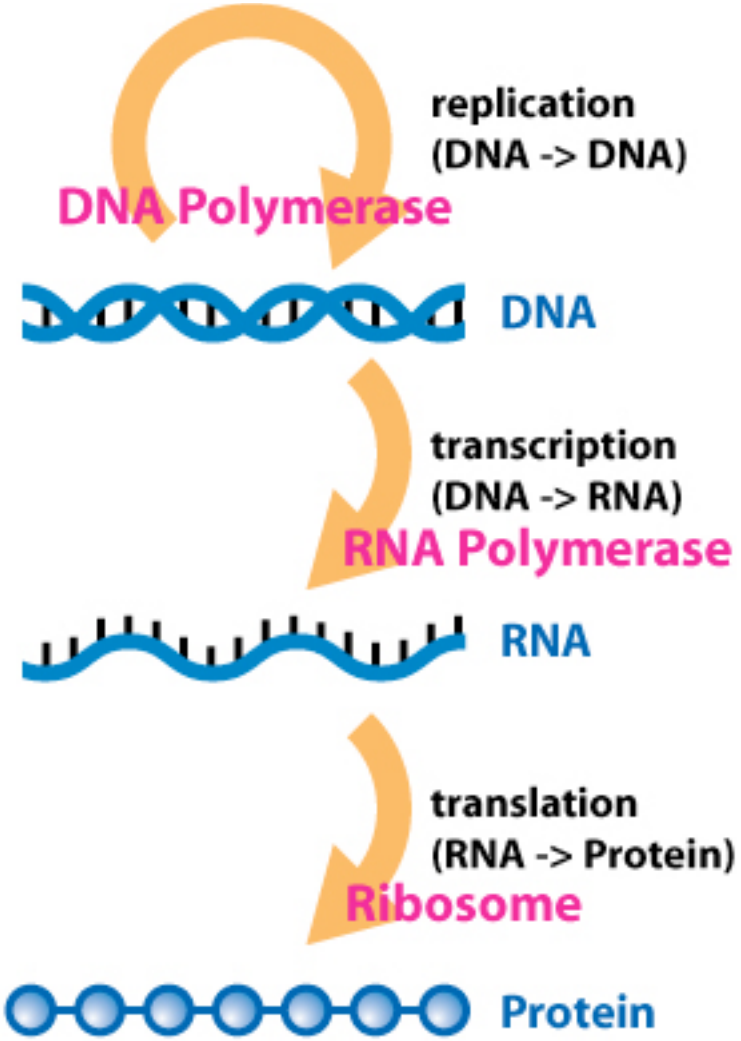
Translation

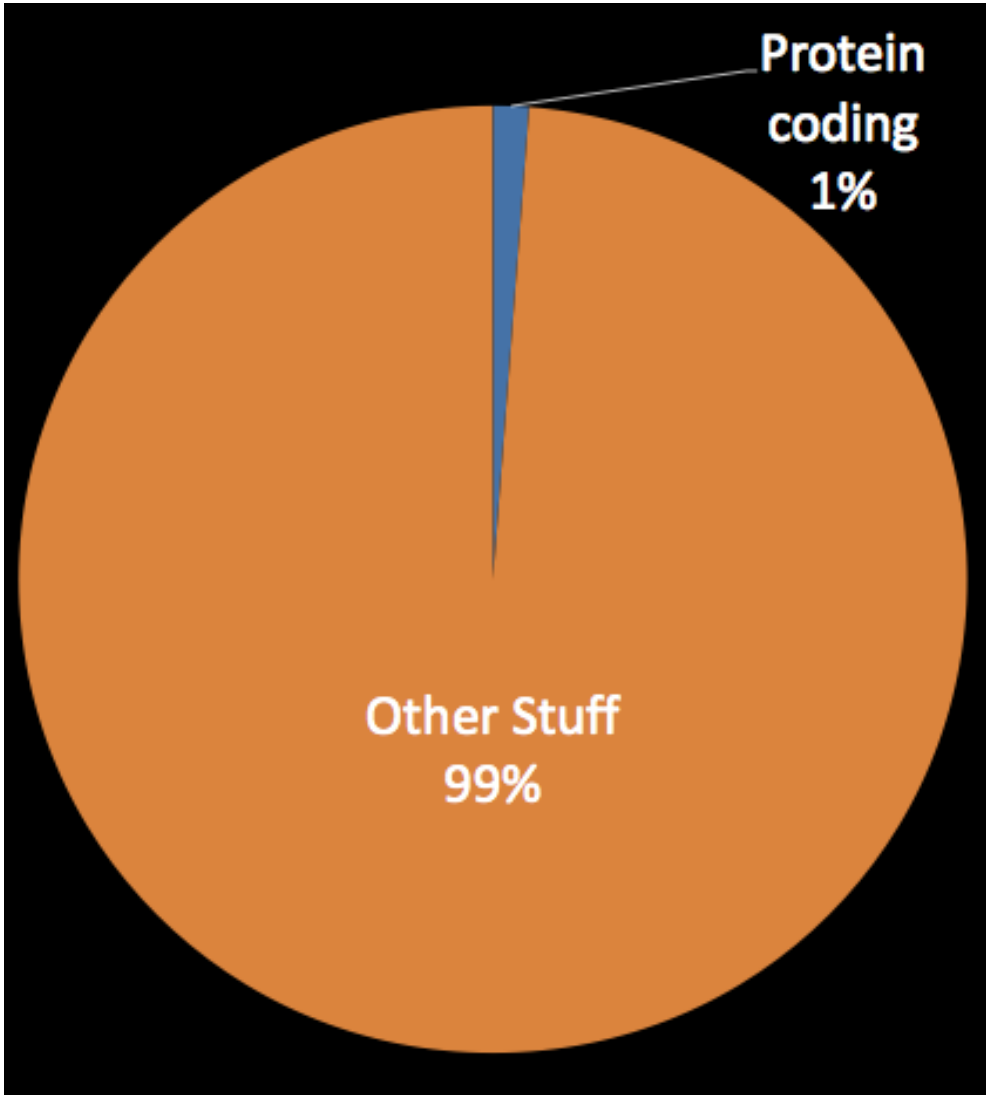


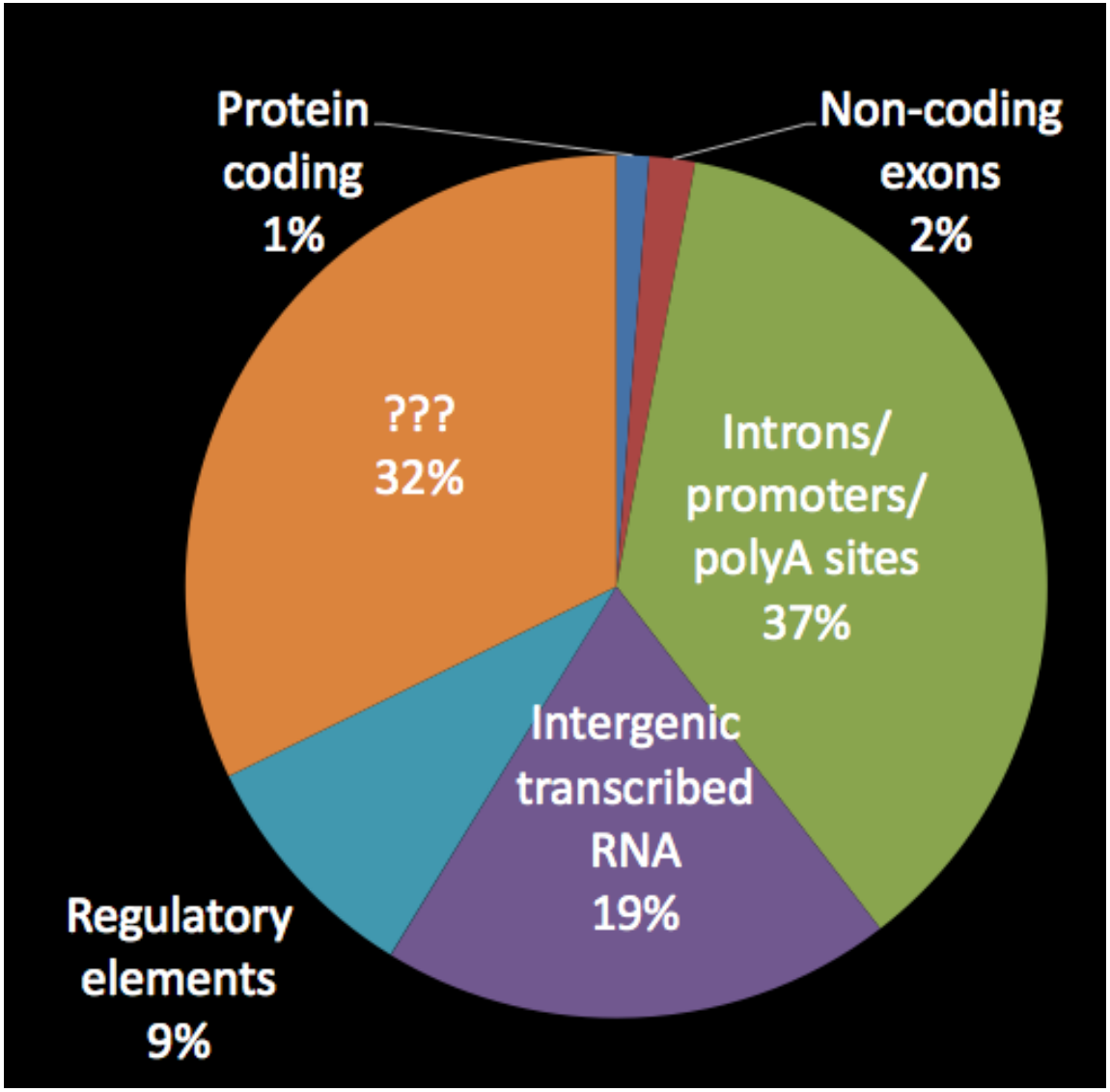
Translation

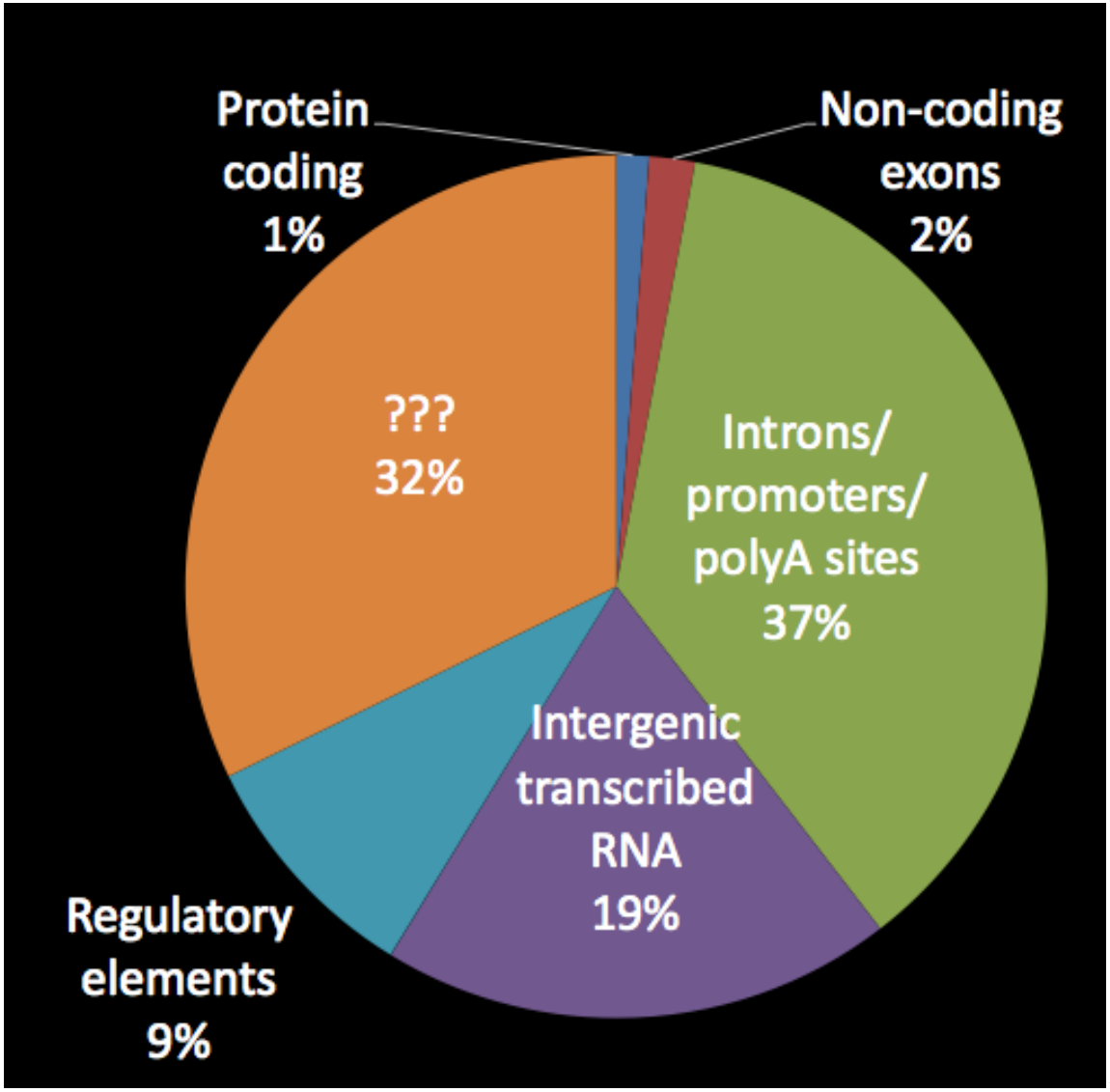


Central Dogma of Biology

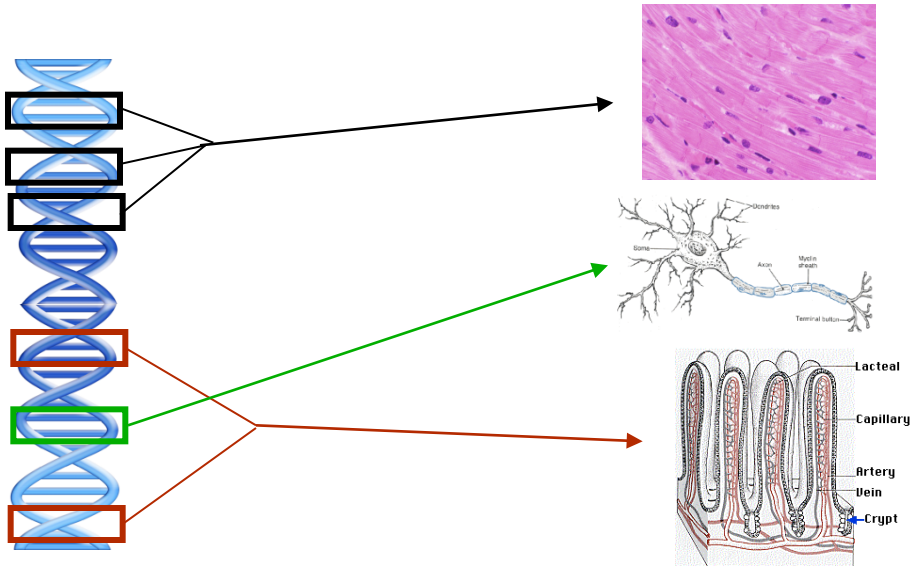



Protein coding 1%
Other Stuff 99%


Protein coding 1%
Non-coding exons 2%
??? 32%
Introns/ promoters/ polyA sites 37%
Intergenic transcribed RNA 19%
Regulatory elements 9%


Protein coding 1%
Non-coding exons 2%
??? 32%
Introns/ promoters/ polyA sites 37%
Intergenic transcribed RNA 19%
Regulatory elements 9%

Different Cell Types

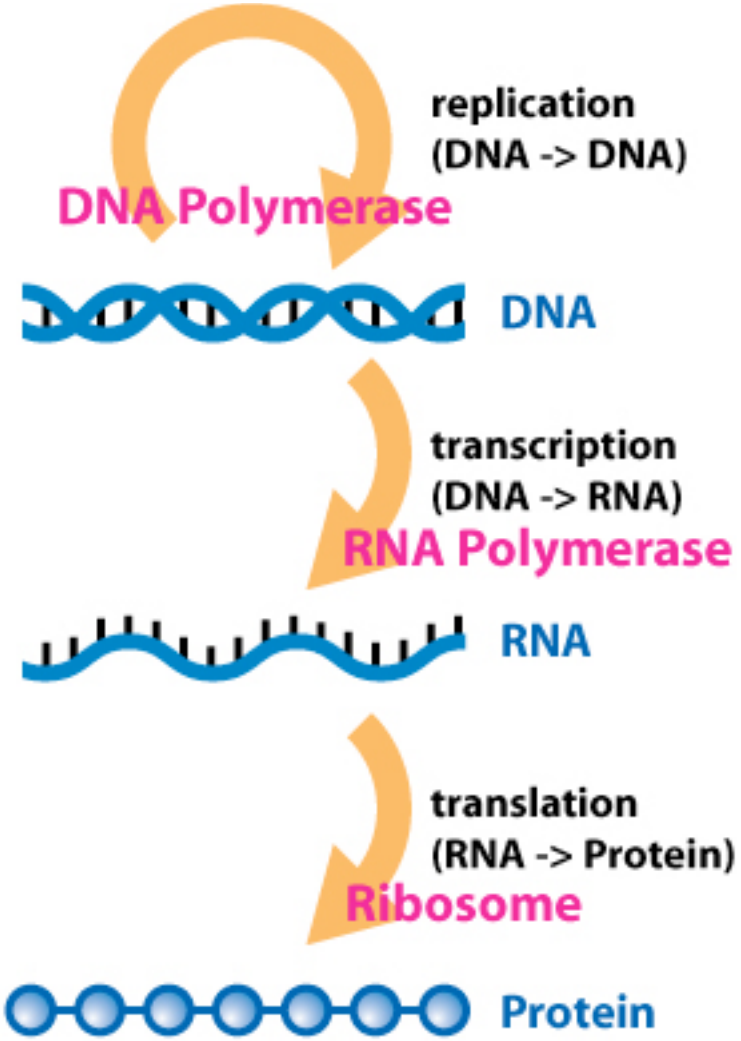


Subsets of the DNA sequence determine the identity and function of different cells

Gene Expression Regulation

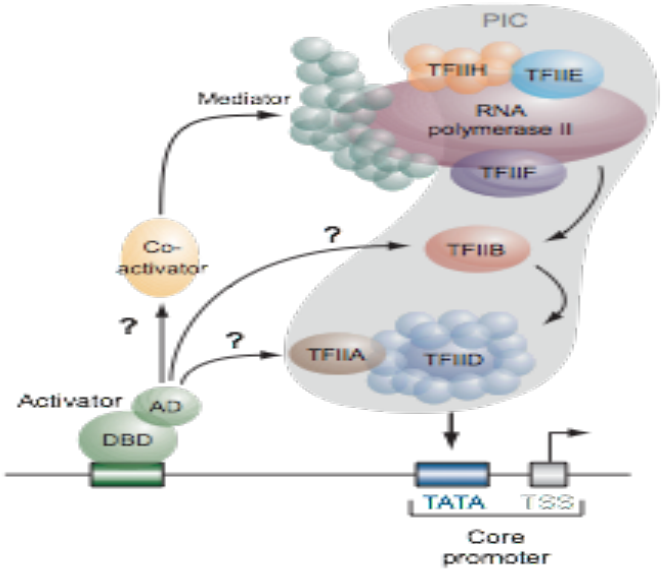
- When should each gene be expressed?
- Why? Every cell has **same DNA** but each cell expresses **different proteins**.
- Signal transduction: One signal converted to another: cascade has "master regulators" turning on many proteins, which in turn each turn on many proteins

Central Dogma of Biology



Transcription Regulation

- Transcription factors link to binding sites
- Complex of transcription factors forms
- Complex assists or inhibits formation of the RNA polymerase machinery

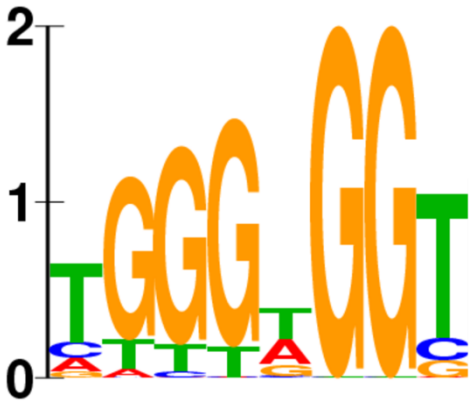


Transcription Factor Binding Sites

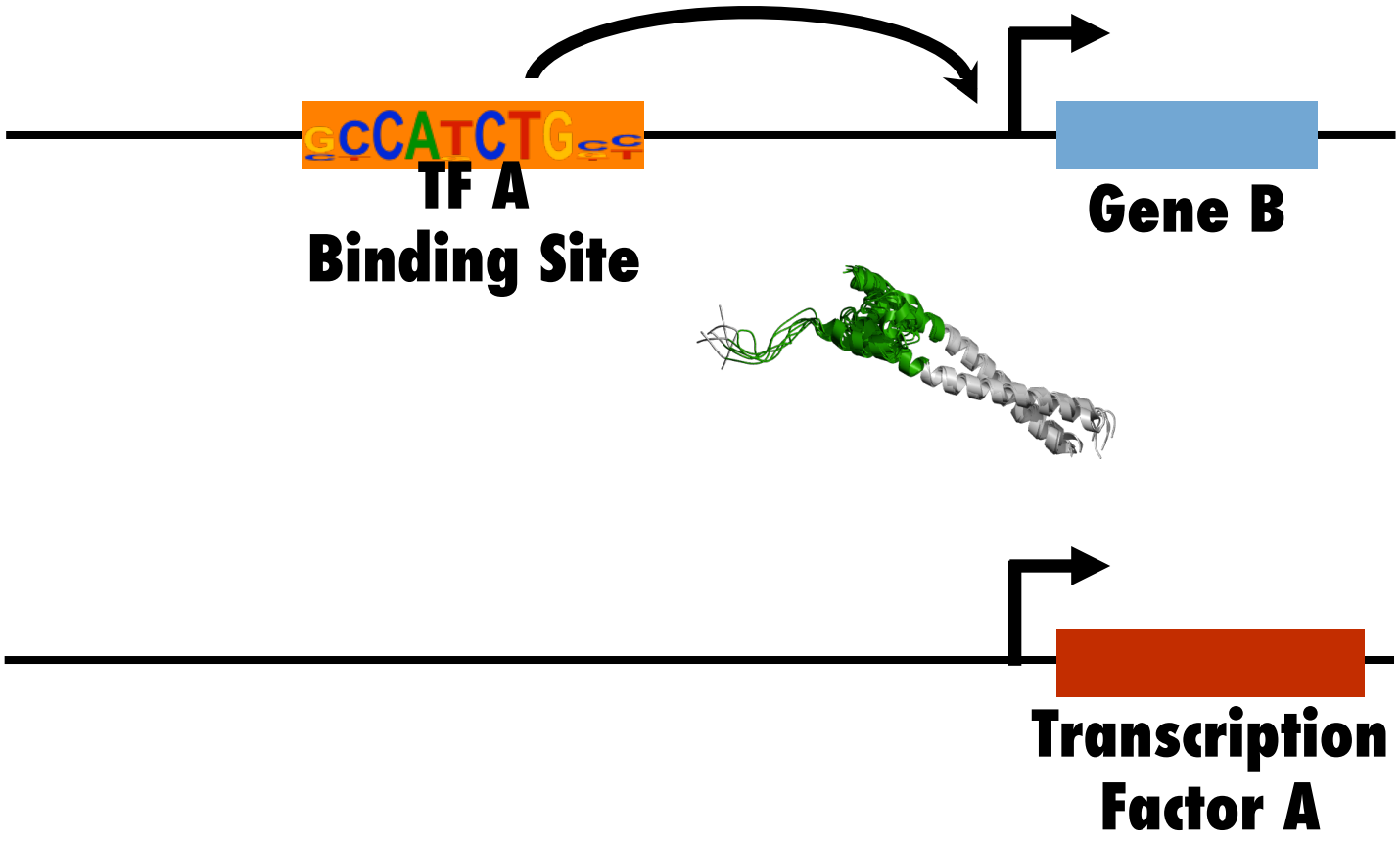
- Short, degenerate DNA sequences recognized by particular transcription factors

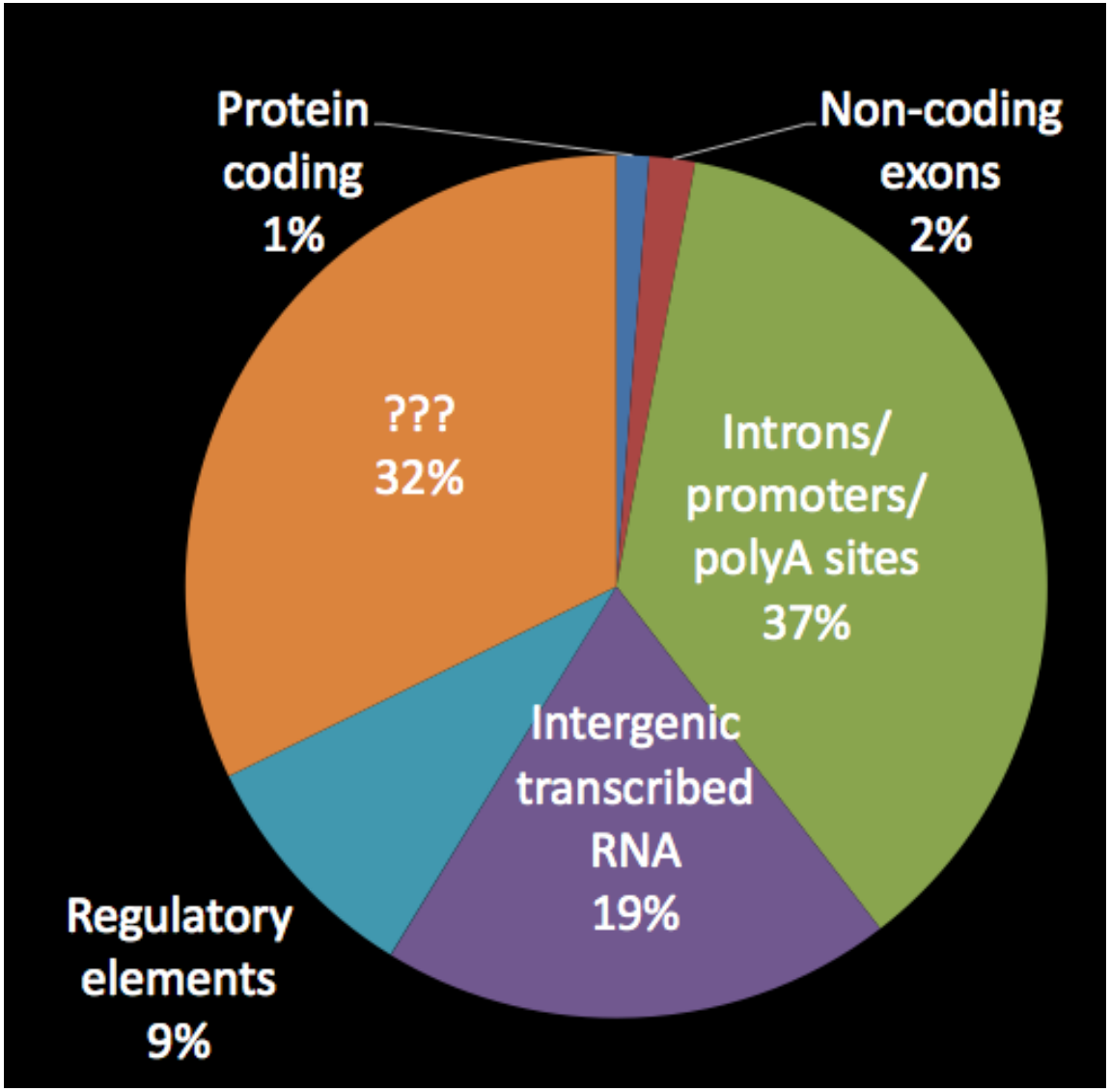
- For complex organisms, cooperative binding of multiple transcription factors required to initiate transcription

Binding Sequence Logo



Transcription Regulation



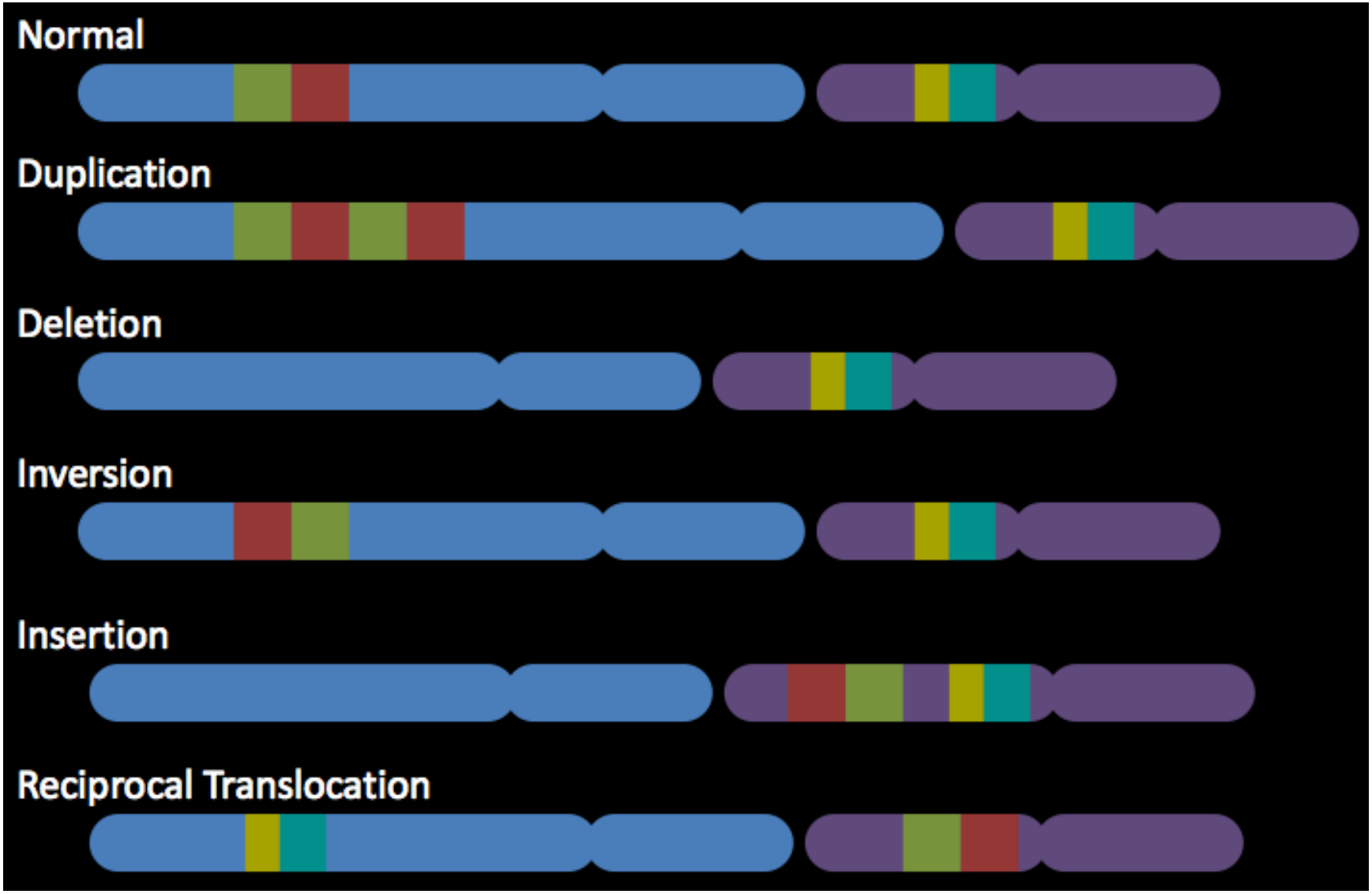

Protein coding 1%
Non-coding exons 2%
??? 32%
Introns/ promoters/ polyA sites 37%
Intergenic transcribed RNA 19%
Regulatory elements 9%

Q: What if the transcription/ translation machinery makes mistakes?

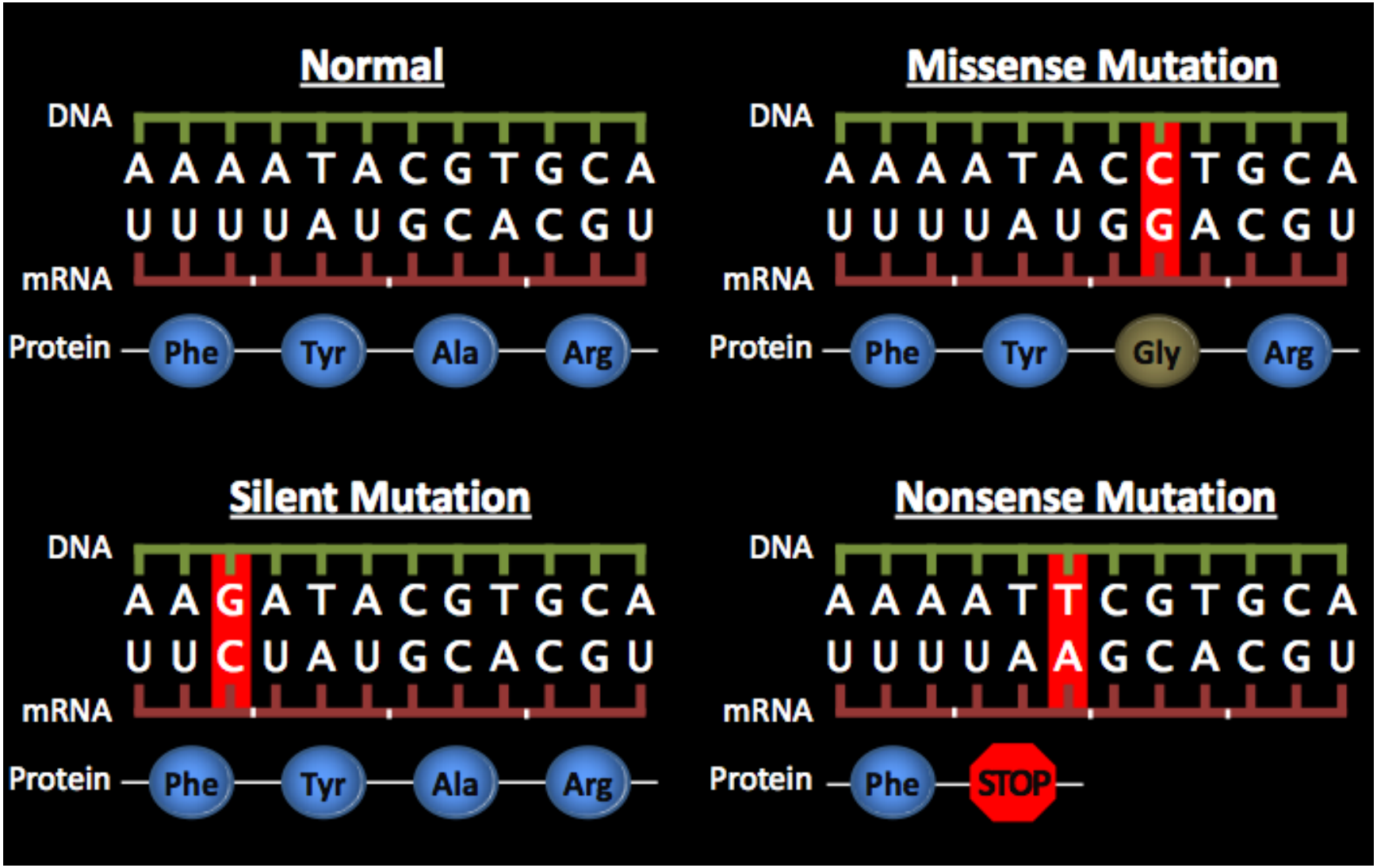
Q:What is the effect in coding regions?

Evolution = Mutation + Selection

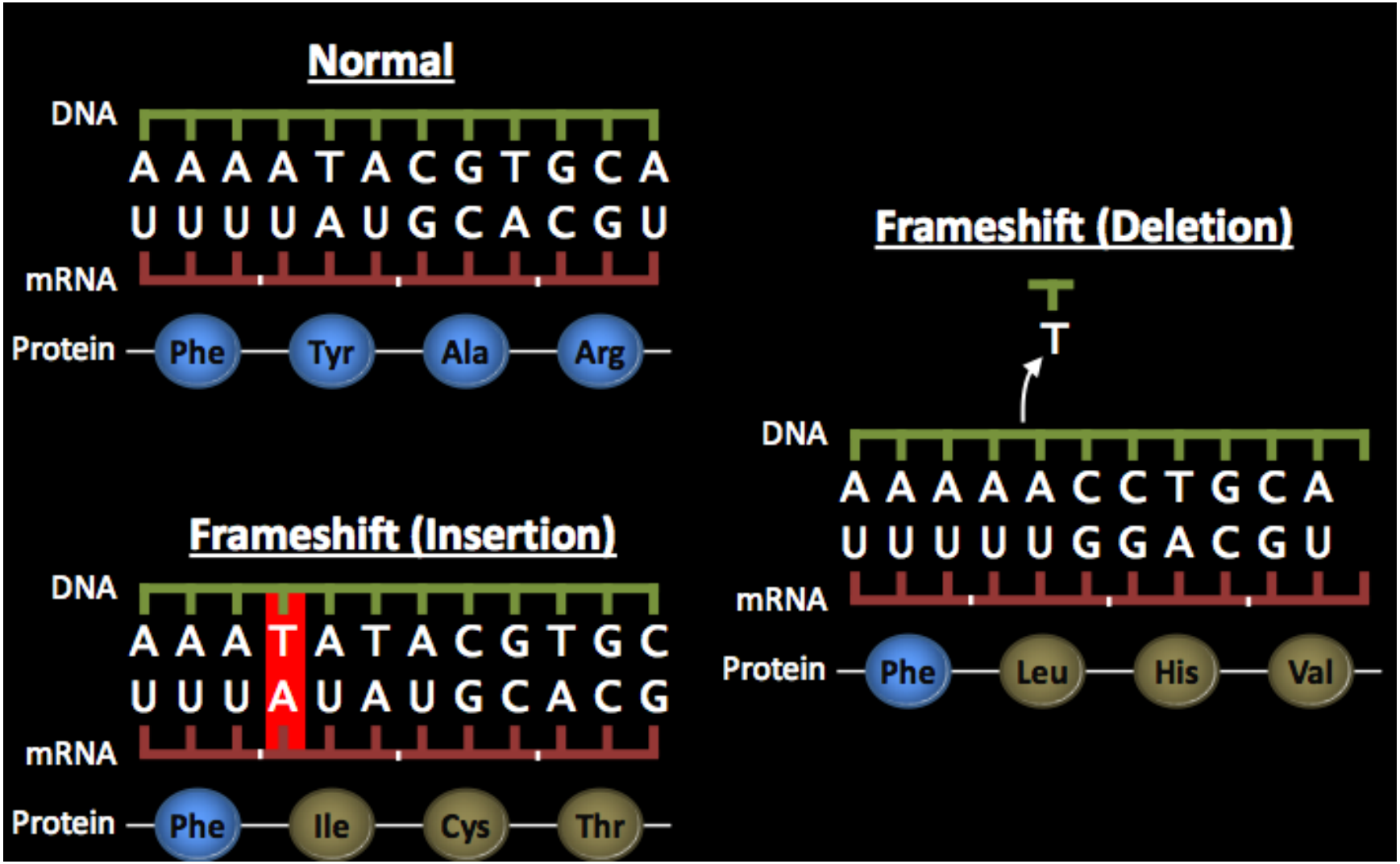
Structural Abnormalities



Single Nucleotide Changes



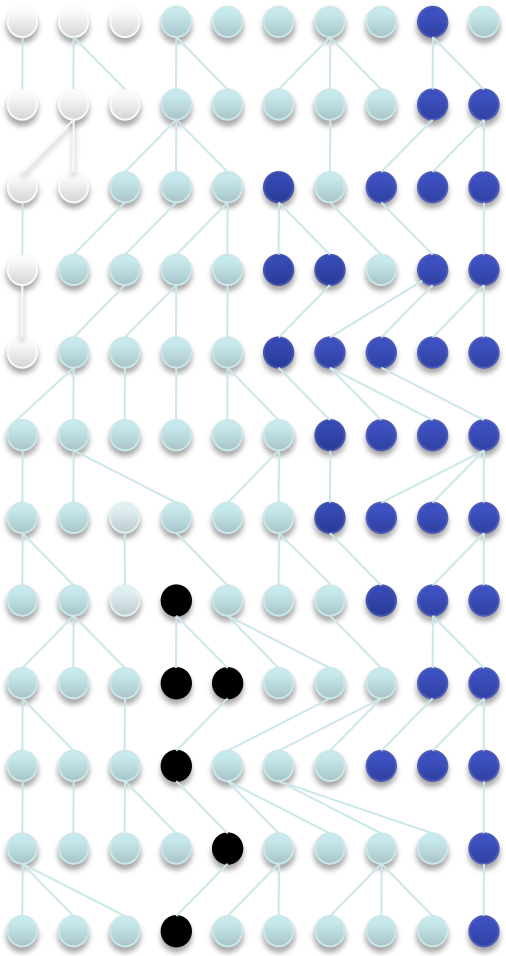
Single Nucleotide Changes



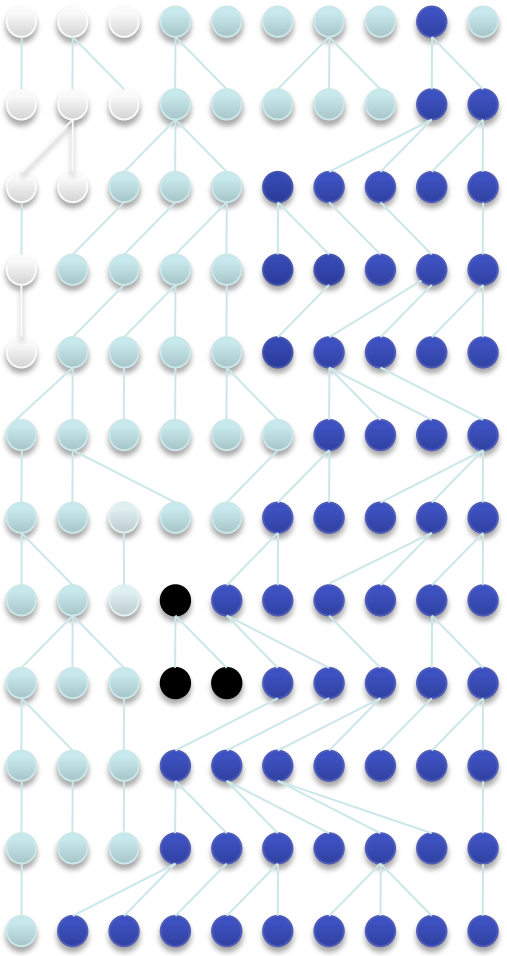
Evolution = Mutation + Selection

Selection

time



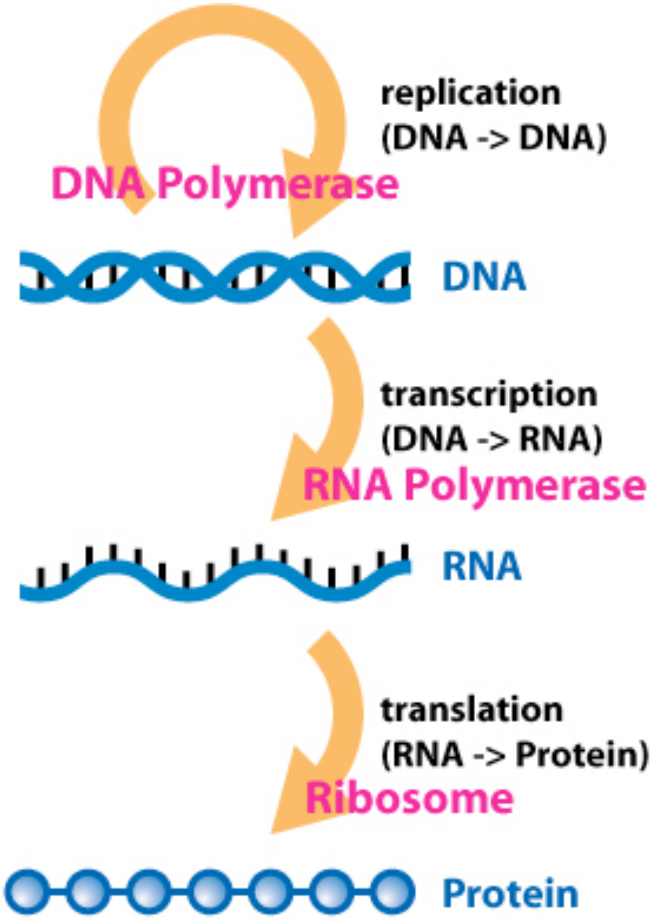
Harmful mutation



Beneficial mutation

Evolution = Mutation + Selection

Summary



Evolution = Mutation + Selection

Summary

- All hereditary information encoded in double-stranded DNA
- Each cell in an organism has same DNA
- DNA → RNA → protein
- Proteins have many diverse roles in cell
- Gene regulation diversifies protein products within different cells

Further Reading

- See website: cs173.stanford.edu

Extra Slides

Gene Regulatory Region