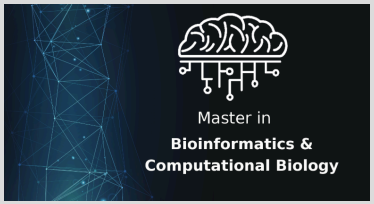
Algorithms for Bioinformatics

2018/2019

Molecular Biology Concepts

Pedro G. Ferreira
pgferreira@fc.up.pt

Master in Bioinformatics and Computational Biology
[dCC] FCUP



Outline

- Molecular Biology concepts
  - Macro perspective
  - Cells
  - Genetic Information: nucleic acids, transcription and translation
  - Genes: structure and regulation
  - Genomes
  - Alternative Splicing
- Exercises

The Macro Perspective: Evolution and Kingdoms

All the living creatures are a result of millions of years of evolution.
But what is evolution?

Evolution has three main components:

- **Inheritance**
  - Passage of characteristics from parents to offspring.
  - Determines most of the structure and functions of the organism.
  - The amount of variation passed from one generation to the next one is very small.

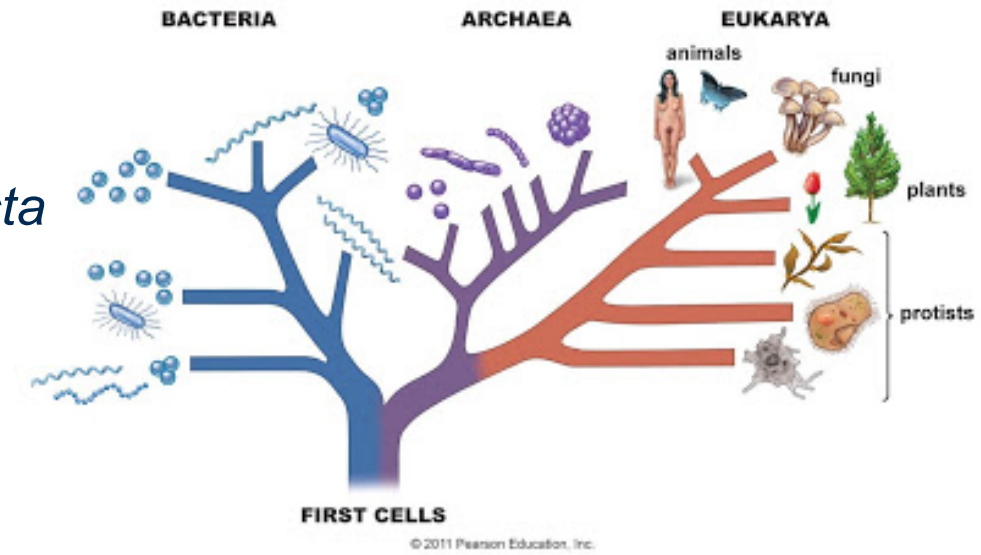
- **Variation**
  - Occurs with mutations, sexual recombination, random changes of genetic material.

- **Selection**
  - Reflects the fitness of the organism to adapt to the medium.
  - This fitness capacity is expressed through reproduction, i.e. parents transmit their fitness to their offspring.

Organization of living things

Three domains: **Bacteria**, **Archea** and **Eukarya**.

- Eukarya is divide in 4 domains:
  Animalia, Plantae, Fungi, Protista



- Virus are sometimes also considered a domain. They need other organisms to survive and reproduce.

- Within each domain organisms are divided according to their cellular structure and composition.

Phylogenetics

Phylogenetics is the division of biology that studies evolutionary divergence and relationship between organisms, based on two important important concepts:

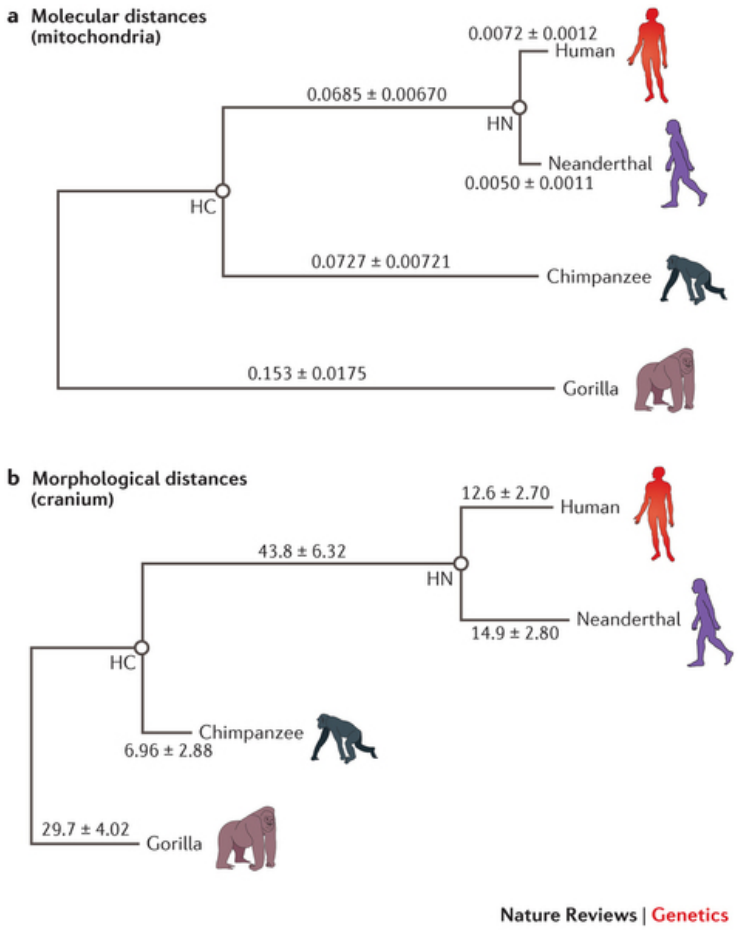
- **Similarity**

  Measures the resemblance and differences between organisms without taking into account any contextualization.

- **Homology**

  Investigates the common ground between the organisms and if they share any ancestral characteristics. Find the point in the evolutionary tree that they started to diverge.

Phylogenetics



Depending on the nature of the characters different evolutionary rates can be observed:

- Molecular sequences have a nearly constant rate in these close species.
- Morphological characteristics evolve in a different way.

Bayesian molecular clock dating of species divergences in the genomics era
Mario dos Reis, Philip C. J. Donoghue & Ziheng Yang
Nature Reviews Genetics 17, 71–80 (2016) doi:10.1038/nrg.2015.8

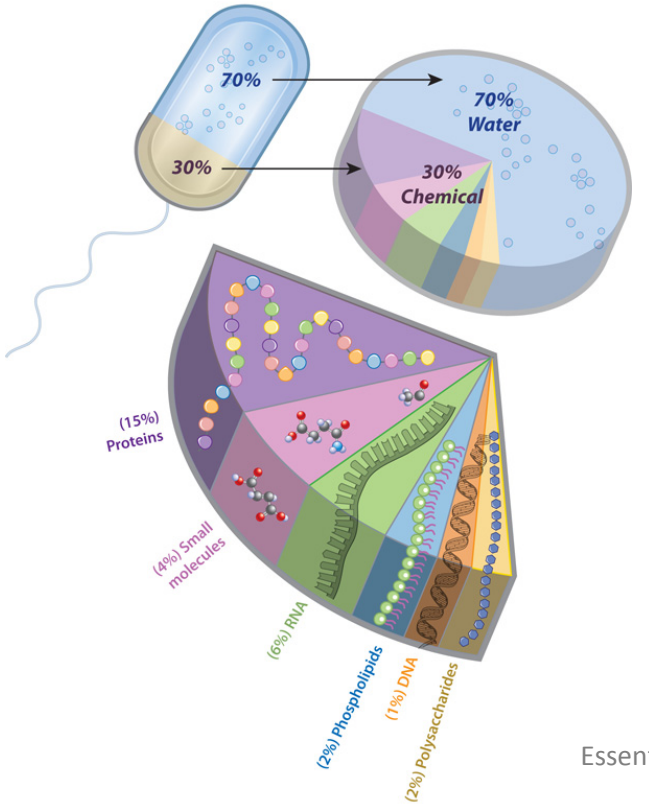
The Micro perspective: The cell

What is Life made of?

All living things are made of Cells.

The cell is the unit of life. Each cell derives from another cell and contains all the necessary information to replicate itself. All cells have some common features:

Composition of a bacterial cell



Its composition (by weight) is:

- 70% water,
- 7% small molecules (amino-acids, nucleotides),
- 23% macromolecules (proteins, lipids, polysaccharides).

Essentials of Cell Biology, Nature Education.

Cell drives Organism organization

Organisms are categorized according to their cell type:

- Prokaryotes
  No nucleus or internal membranes.

- Eukaryotes
  - Nucleus.
  - Internal membranes.
  - Organelles inside the cell that play different and specific roles.

- Unicellular
  Prokaryotes: Bacteria, Archaea.
  Eukaryotes: baker yeast.

- Multicellular
  Eukaryotes: animals, plants, fungi

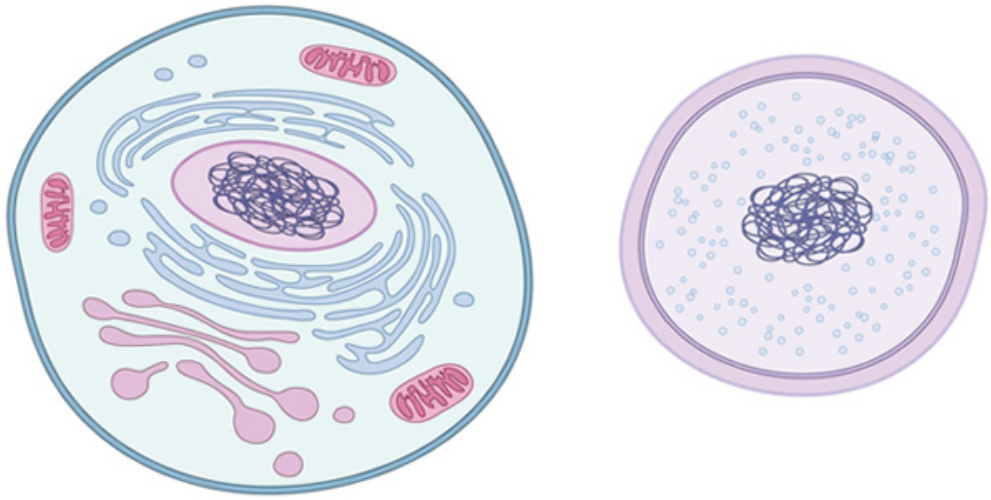
Eukaryotic and Prokaryotic differences

A eukaryotic cell:

- membrane-enclosed DNA, which forms a structure called the nucleus
- additional membrane-bound organelles of varying shapes and sizes

A prokaryotic cell:

- does not have membrane-bound DNA
- lacks other membrane-bound organelles



Essentials of Cell Biology, Nature Education.

What defines a cell?

- **Proteins** perform most of the functions in the cell, they have catalytic and structural functions, from sensors and signaling to promoting chemical reactions (catalysis).
- **Enzymes** are proteins that convert cellular molecules in other types of molecules necessary for the functions of the cell, like generating energy.

DNA and RNA are composed of nucleic acids. Proteins are composed of amino-acids.

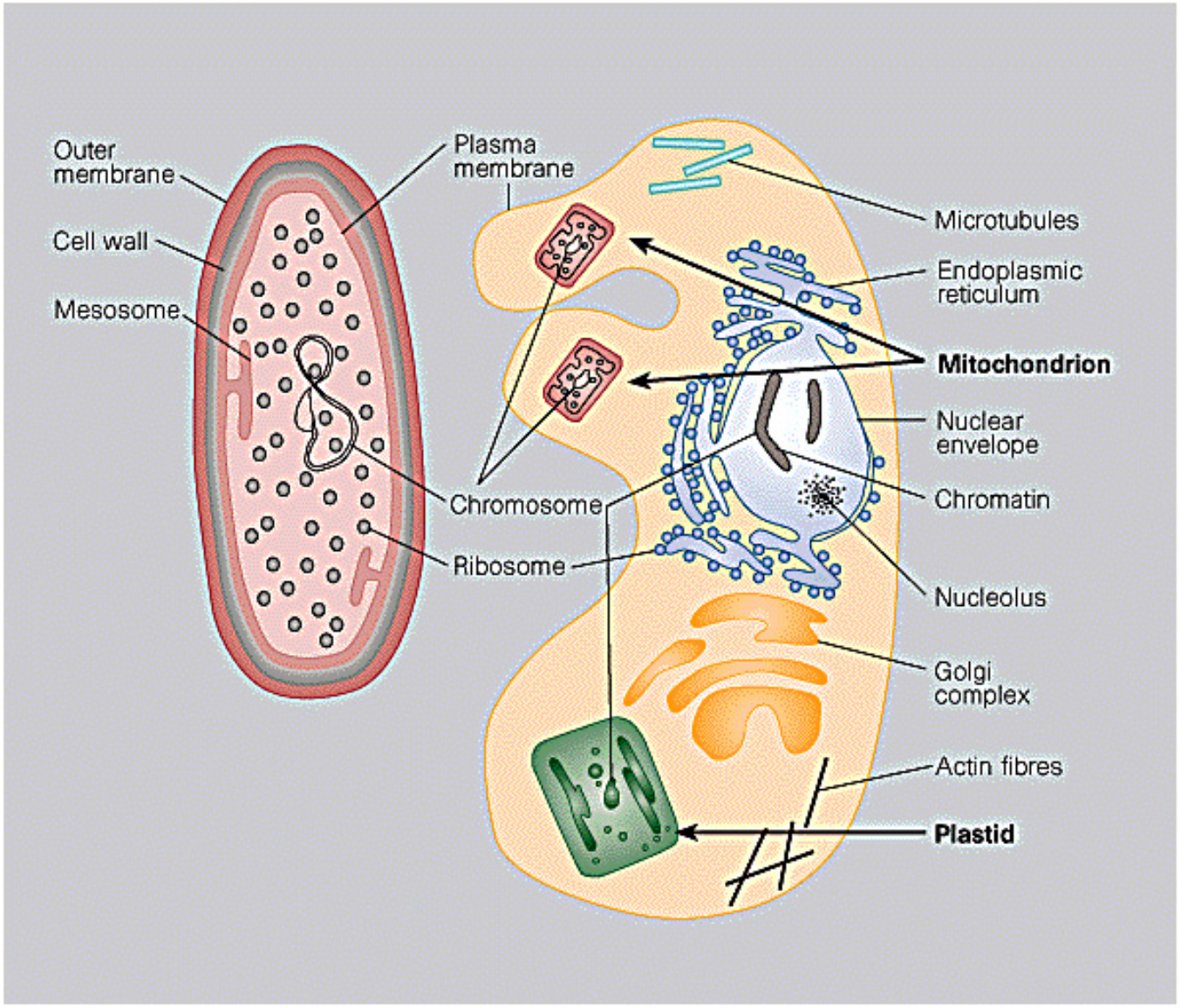
Other important organic molecules:

- **Carbohydrates**, store energy (simple - immediate energy demands, complex - long term storage of energy).
- **Lipids**, make part of the plasma membrane and also store energy and are involved in signaling.

Other components of the cell:

- **Mitochondria** and Chloroplasts are cellular organelles involved in the production of energy.
- **Ribosomes** are large and complex molecules composed by a mixture of proteins and genetic material. Their function is to assemble proteins.

Cell drives Organism organization



Essentials of Cell Biology, Nature Education. 2010

Cell: Conclusions

- Cells are the unit of life. Some cells are organisms by themselves (unicellular organisms); other are part of more complex multicellular organisms.

- Major classes of organic molecules found in all cells:
  - nucleic acids
  - proteins
  - carbohydrates
  - lipids.

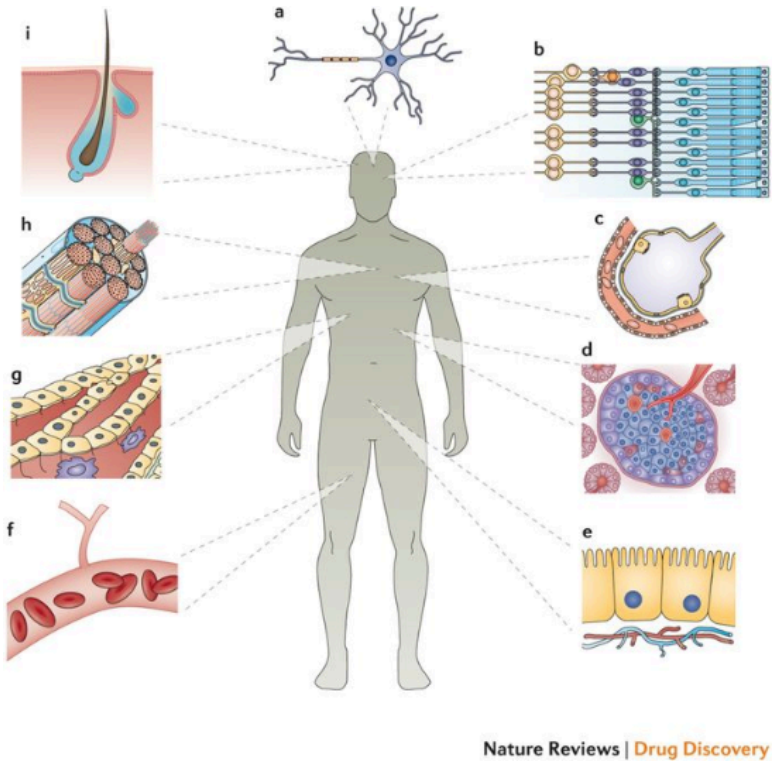
- Two major categories of cells as a result of ancient evolutionary events:
  - prokaryotes, with their cytoplasmic genomes.
  - eukaryotes, with their nuclear-enclosed genomes and other membrane-bound organelles.

- Cells are small and they have a vast variety of shapes and sizes (the human body has > 200 cell types each specialised to carry a particular function or form a particular tissue.

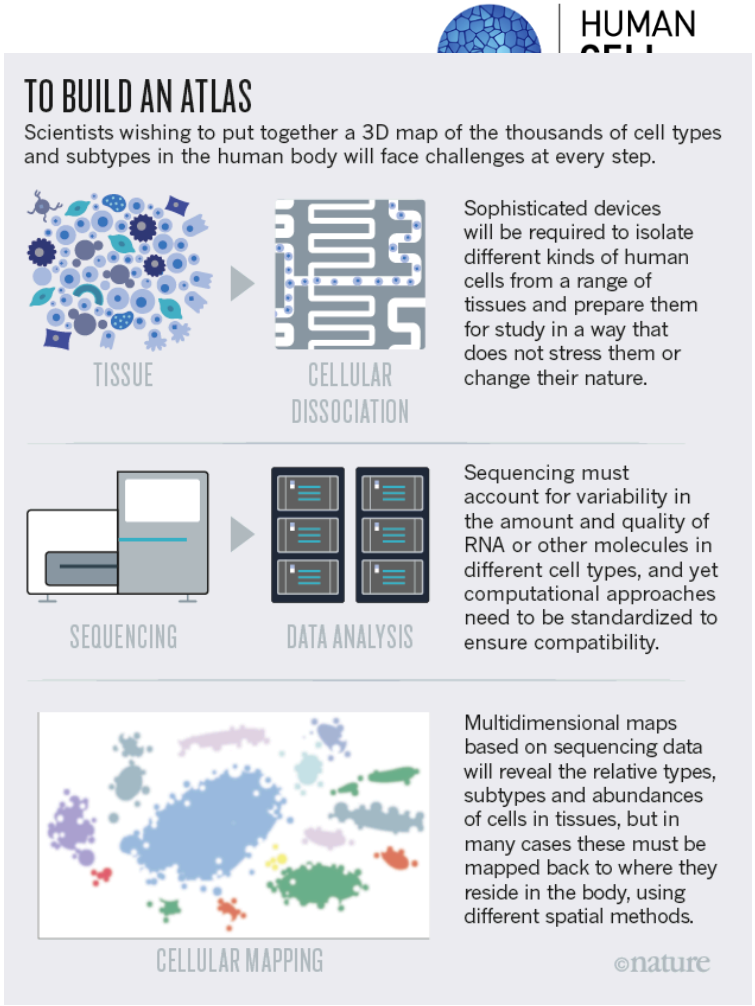
- Cells form tissues that themselves form organs, and eventually entire organisms.

Human Cell Atlas

"To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease."



Current status of pluripotent stem cells: moving the first therapies to the clinic
Erin A. Kimbrel & Robert Lanza
Nature Reviews Drug Discovery volume 14, pages 681–692 (2015)



Information transfer in the cell: nucleic acids

- Both DNA and RNA are polymers composed of four nucleic acid units, called nucleotides or bases.

- Adenyne (**A**) and Guanine (**G**), belong to one group (purynes).
- Cytosine (**C**) and Timine (**T**) and Uracil (**U**), belong to another group (pyrimidine).

- Timine only exists in DNA and Uracil is only found in RNA, the other three bases exist in both.

The **DNA** is composed of **two complementary strands** due to connections established between the bases in both strands.

- **Adenine and Timine** (A == T), connected by two hydrogen connections
- **Guanine and Cytosine** (G === C), connected by three hydrogen connections

- Chains are antiparallel because they are connected in opposite directions

Organization of genetic material

- **Genome**: an organism's genetic material (complete set of DNA).
  - a bacteria contains about 600,000 DNA base pairs
  - human and mouse genomes have some 3 billion.
  - human genome has 24 distinct chromosomes.
  - Each chromosome contains many genes.

- **Gene**: a discrete units of hereditary information located on the chromosomes and consisting of DNA and encode instructions on how to make proteins.
- **Genotype**: The genetic makeup of an organism.
- **Phenotype**: the physical expressed traits of an organism.

Organization of genetic material

Genome organization

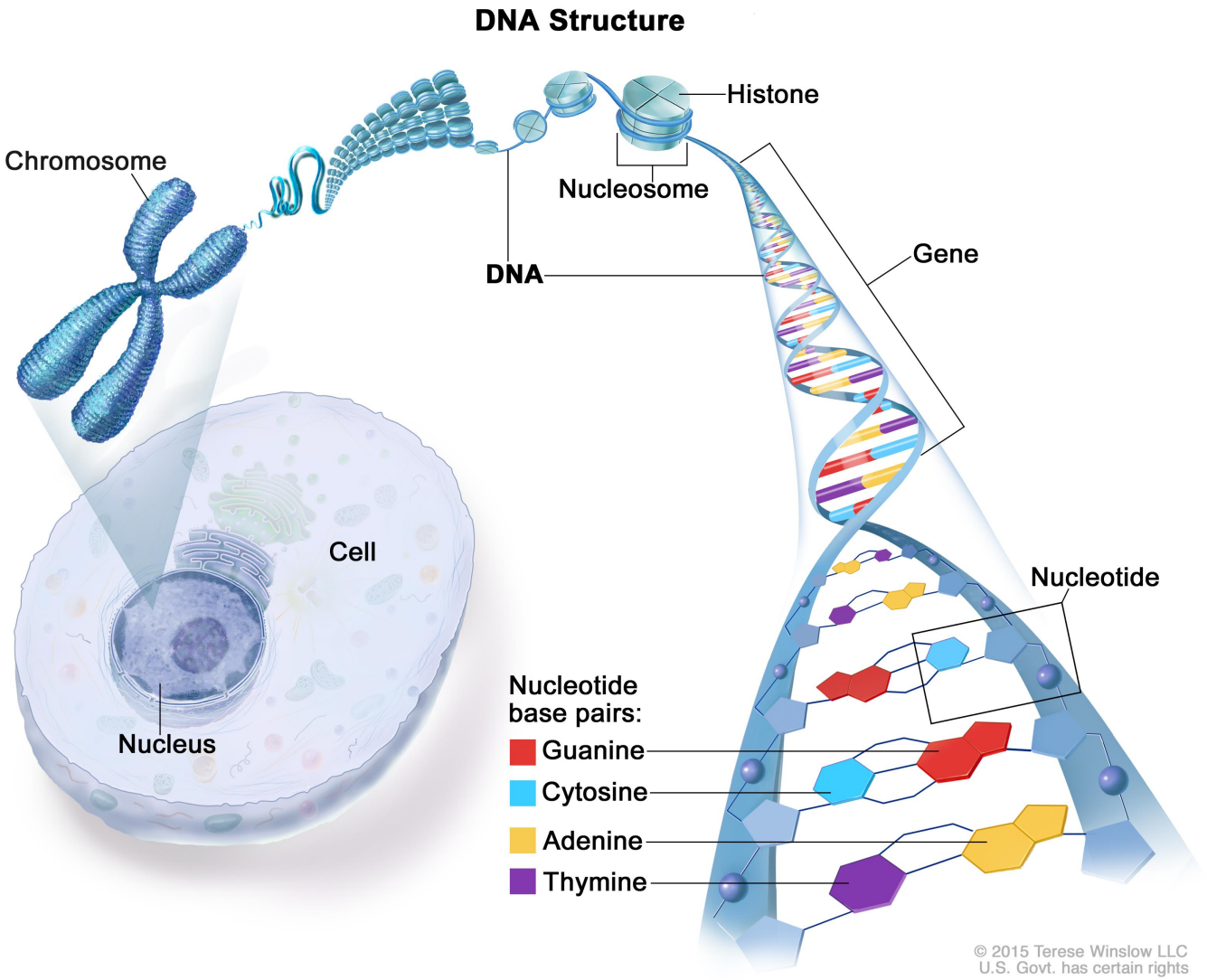
- Prokaryotic
  - exists in the form of a circular chromosome located in the cytoplasm.

- Eukaryotes
  - found in the **nucleus** and
  - tightly packaged into linear **chromosomes**.
  - chromosomes consist of a DNA-protein complex called **chromatin** that is organized into subunits called **nucleosomes**.

This organization allows to:

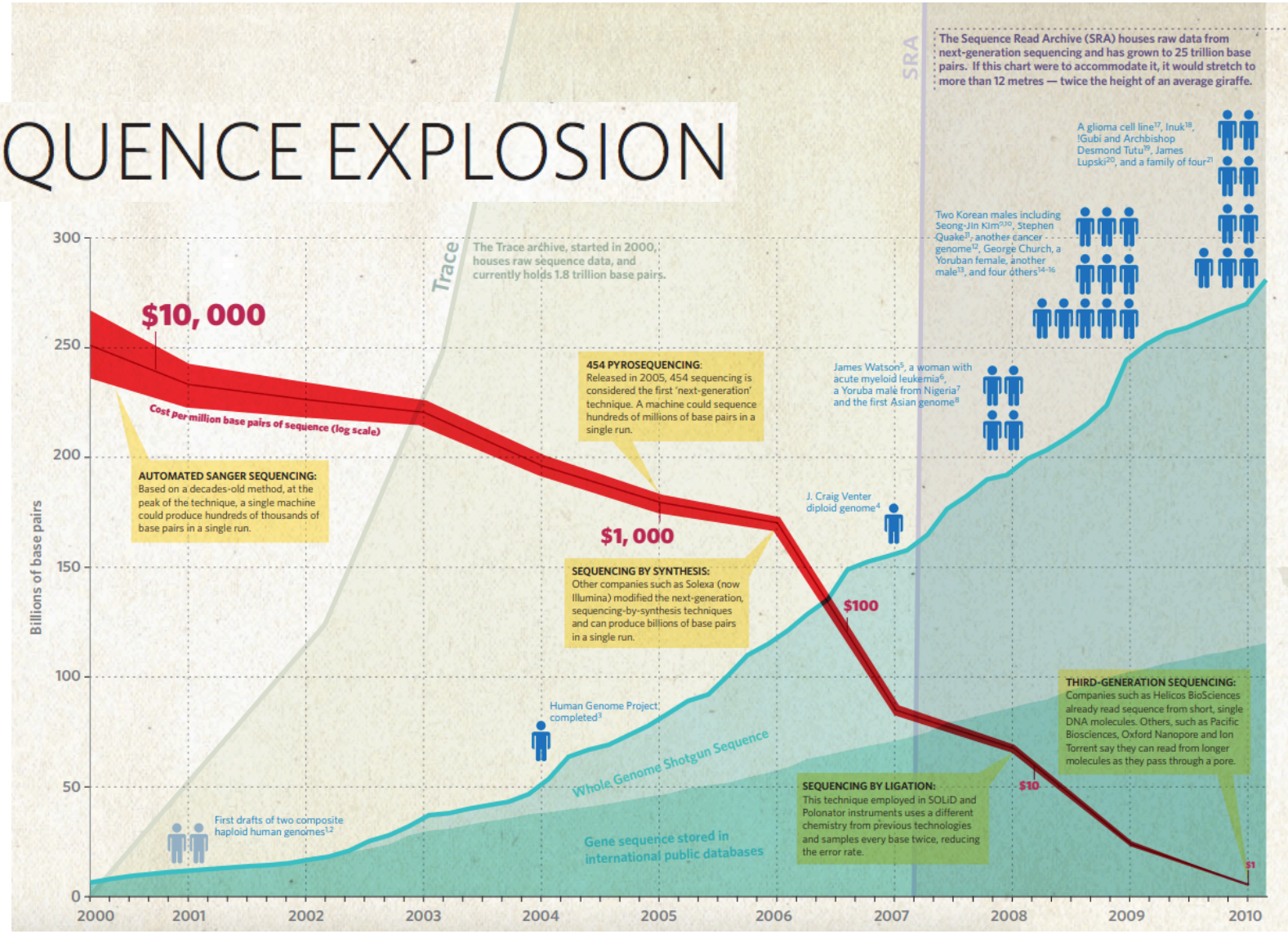
- fit a long DNA molecule in a small space.
- provide regulatory structure for gene expression.

Organization of genetic material



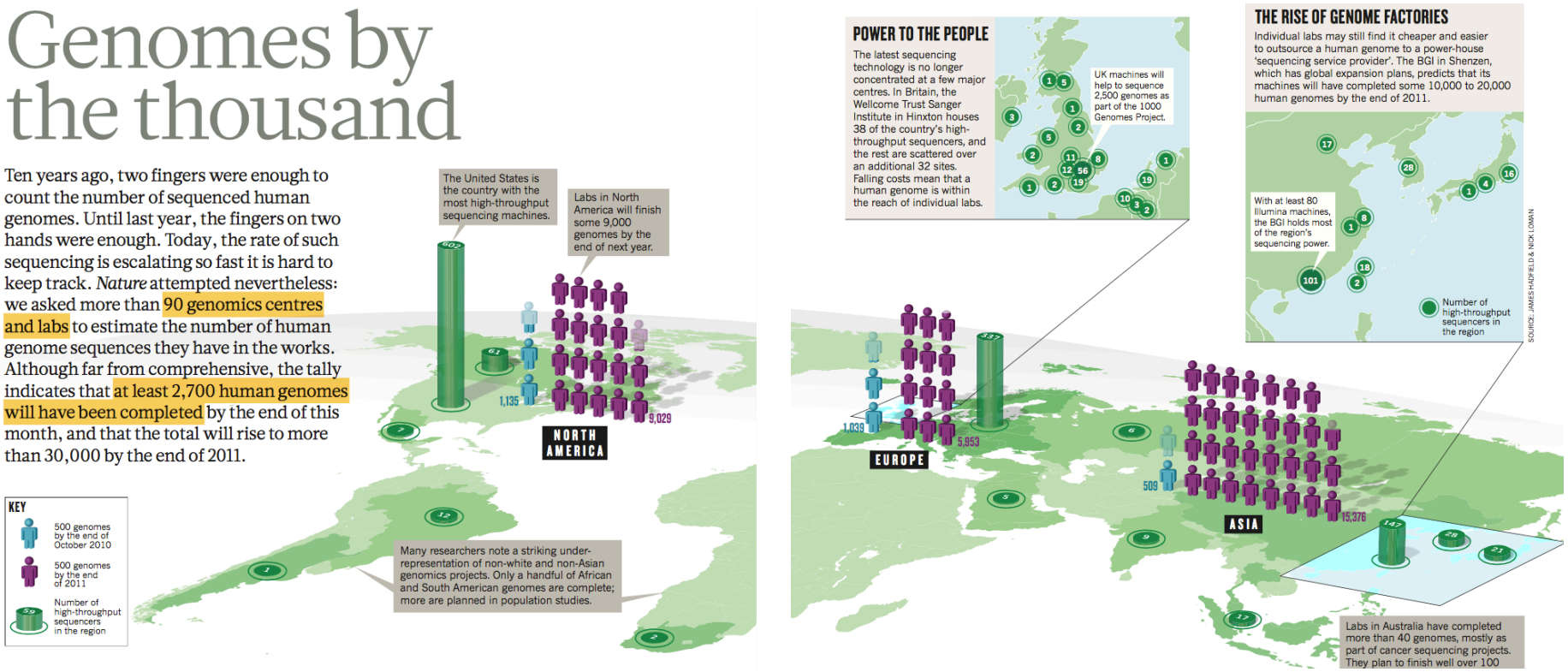
www.cancergenome.gov

Humana Genome Sequencing



Human Genome at ten
Nature, Vol 464 Vol 464, 1 April 2010

Humana Genome Sequencing



Genomes by the thousand
28 October 2010, vol 467 Nature 102

Passing Information to new cells

- Cells divides in two daughter cells and it passes a copy of its DNA to each of its cell.

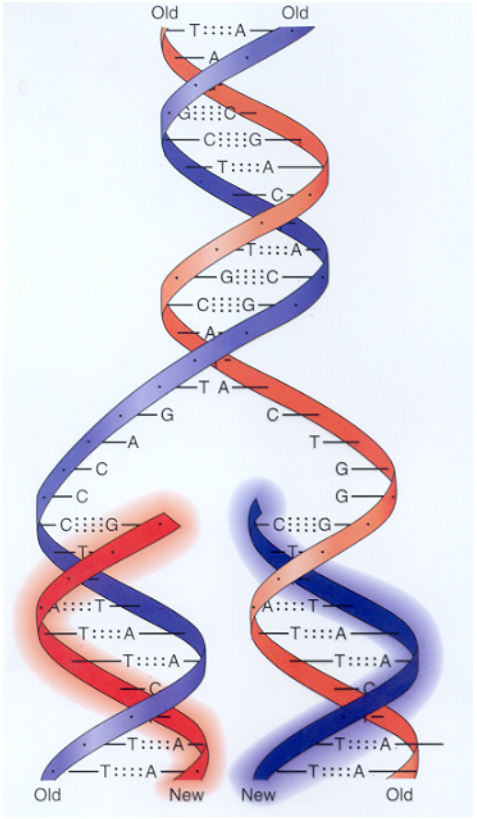
It is important that this is a process as accurate as possible to ensure that the new cells are healthy and inherited the features of the mother cell.

- **DNA replication** is the process of copying the DNA.

In most multicellular organisms, every cell carries the same DNA.

The use of different parts of the information encoded in the DNA results in different cell types.

For instance in nerve cells there is an abundance of proteins called neurotransmitters (send messages to other cells).
In muscle there is an abundance of protein-based filaments (muscle contractions).



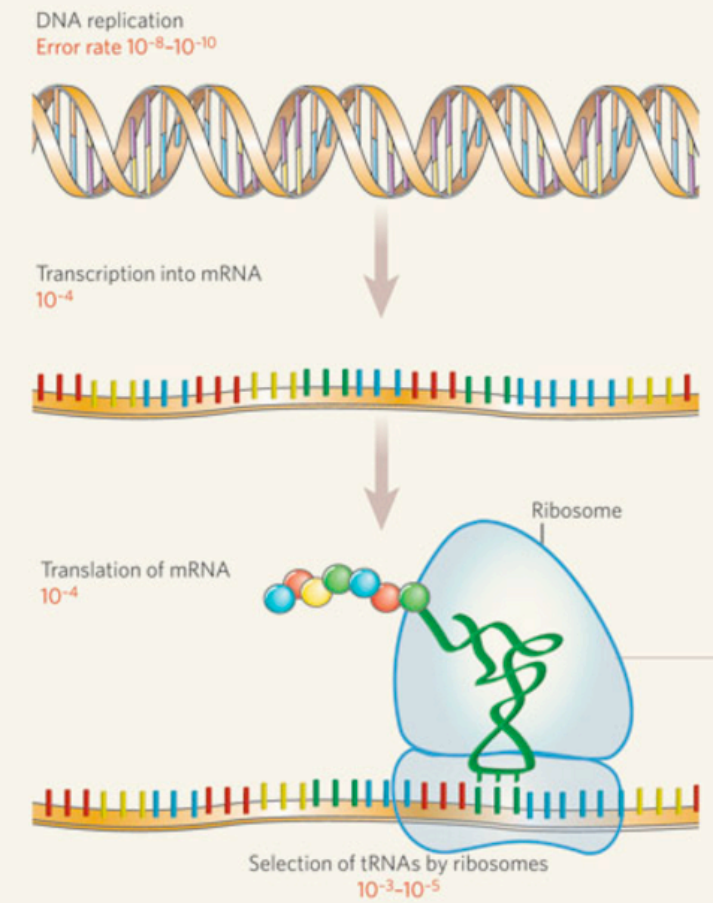
Introduction to chemistry TED Ankar

DNA, RNA and Proteins: Flow of genetic information

- A gene is expressed in two steps:
  1) Transcription: RNA synthesis
  2) Translation: Protein synthesis

- **Transcription** of the genetic message is when the genes in the DNA encode the production of a messenger molecule of RNA.

- **Translation** of the genetic message is when the messenger molecule of RNA encodes the production of proteins.



Adapted from: Molecular biology: Sticky end in protein synthesis
Hervé Roy and Michael Ibba
Nature 443, 41-42 2006.

DNA, RNA and Proteins: Flow of genetic information

DNA

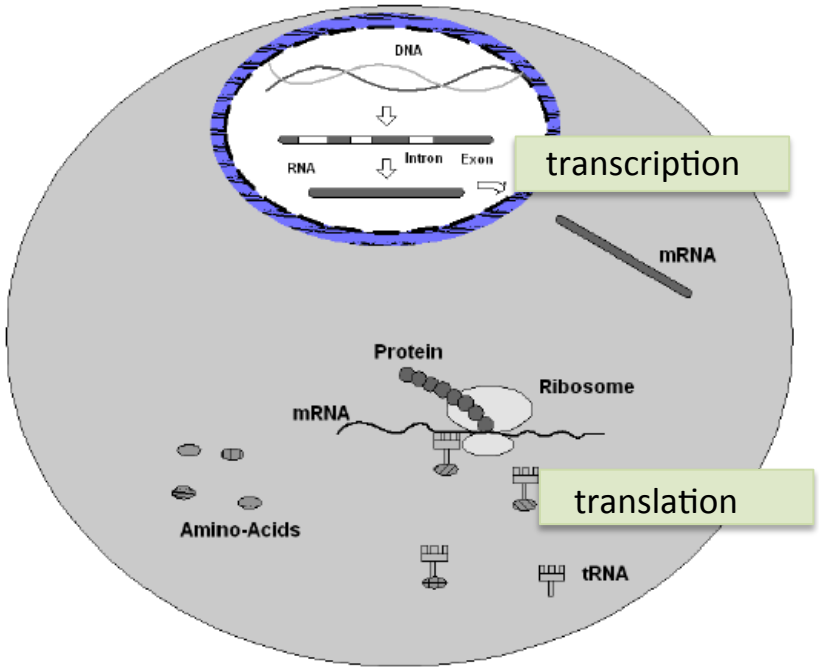
Replication

Transcription

RNA

Translation

Protein



MOLECULAR BIOLOGY FUNDAMENTALS FOR COMPUTER SCIENTISTS
Pedro Gabriel Dias Ferreira

DNA, RNA and Proteins: Flow of genetic information

- DNA, RNA are strings written **in four-letter nucleotide:**
  A C G T for DNA and **A C G U for RNA.**

- **Proteins** are written in **twenty-letter amino acids.**

- The proteins are not directly synthesized by DNA.

- A complementary copy of one of the DNA strands in a uni-dimensional sequence of nucleic acids, the messenger RNA (mRNA) is used.

- The mRNA is the intermediary between the nuclear DNA and the ribosomes where occurs the protein synthesis.

Genetic code

- Each amino acid is coded by 3 (triplet) nucleotides called codon.
- Thus four nucleotides can encode $4^3 = 64$ possible triplets.
- 64 triplets is more than the needed for the existing 20 amino acids (AAs) and therefore most AAs are encoded by more than one codon.

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gin	Arg	A
	Leu	Pro	Gin	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met(Start)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Amino-acid abreviations
Ala = Alanine
Arg = Arginine
Asp = Aspartic Acid
Asn = Asparagine
Cys = Cysteine
Glu = Glutamic Acid
Gln = Glutamine
Gly = Glycine
His = Histidine
Ile = Isoleucine
Leu = Leucine
Lys = Lysine
Met = Methionine
Phe = Phenylalanine
Pro = Proline
Ser = Serine
Thr = Theorine
Trp = Tryptophan
Tyr = Tyrosine
Val = Valine

Genetic code

Main characteristics of the genetic code:

- Constitutes a **standard language** that is common in every cell from the simpler to the more complex organisms.

- **Redundancy** or degeneracy are properties of the genetic code that have to be with the fact that some amino-acids are coded with codons in which only differ in the last nucleotide. This can be a very efficient code-correction mechanism because if some errors occur in the DNA copy it will have no impact in the amino-acid sequence.

- Three of the codons are designate to encode the end of the protein sequence and are called **stop codons**. They are: **UAA**, **UAG** and **UGA**.

- The codon **AUG** has a double function encoding a **start codon** and the amino acid methionine.

Protein synthesis

- Protein synthesis can be described as the **mapping** of a sequence of **codons** into a sequence of **amino acids**.
- This process starts an enzyme called RNA polymerase binds to a determined region of the DNA helix and starts the parsing of the bases sequence in a certain codon.
- Codons are composed of three nucleotides. A nucleotide sequence can have three interpretations, depending where starts the sequence parsing.

Protein synthesis

- For example, the mRNA sequence TGTCGTAGTAATTCG can be read in TGT-CGT-AGT-AAT-TCG (Cis-Arg-Ser-Asn-Ser), TGTC-GTA-GTA-ATT-CG (Val-Val-Val-Met) and TGTCG-TAG-TAA-TTC-G (Ser-Stop-Stop-Fen).

- These three **ways of parsing** the sequence are called **reading frames**. A reading frame with a sufficient given size and with no stop codons is called **Open Reading Frame**.

- It is also possible to read from the second strand of the double-helix, thus making a total of **6 possible reading frames**.

Summary

- Cellular DNA contains instructions for building the various proteins the cell needs to survive.

  1) To manufacture these proteins, specific genes within its DNA must first be **transcribed** into molecules of mRNA;

  2) these transcripts must be **translated** into chains of amino acids,

  3) **fold** into fully functional proteins.

- All cells in a multicellular organism contain the same set of genetic information (**Genome**).

- The differences in the abundance of the RNA (**Transcriptome**) determines the cell specificity.

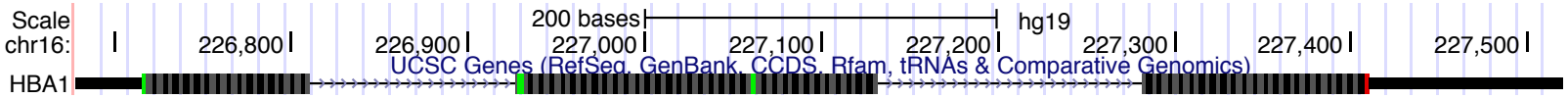
Gene Structure

Gene: a discrete unit of hereditary information located on the chromosomes and consisting of DNA that encodes instructions on how to make proteins.

- In Prokaryotes (e.g. bacteria) genes are a contiguous stretch of DNA.
- In eukaryotes (from yeast to human), a gene consists of a combination of coding segments (exons) interspersed by by non-coding segments (introns).
- In transcription the pre-mRNA contains the sequence of exons and introns.
- Splicing is the process of cutting out the introns and concatenating the exons to form the mature RNA.

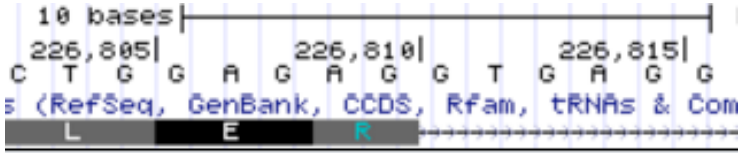
Gene Structure

- Gene: Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA.
- Location: The human alpha globin gene cluster located on chromosome 16 spans about 30 kb
- Function: Involved in oxygen transport from the lung to the various peripheral tissues.



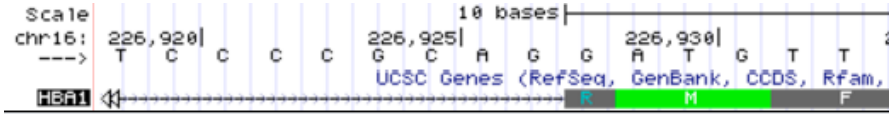
Exons and Introns

- Exons are defined by short and typically conserved intron/exon borders.
- Intron/Exon borders include signal sequences that mark the end of the exon and the start of the intron (5'splice site or donor site); and end of the intron and start of exon (3'splice site and acceptor signal).
- Additional signals found in the introns (e.g. polypyrimidine tract) and exons also help to recognize the intron/exon border.

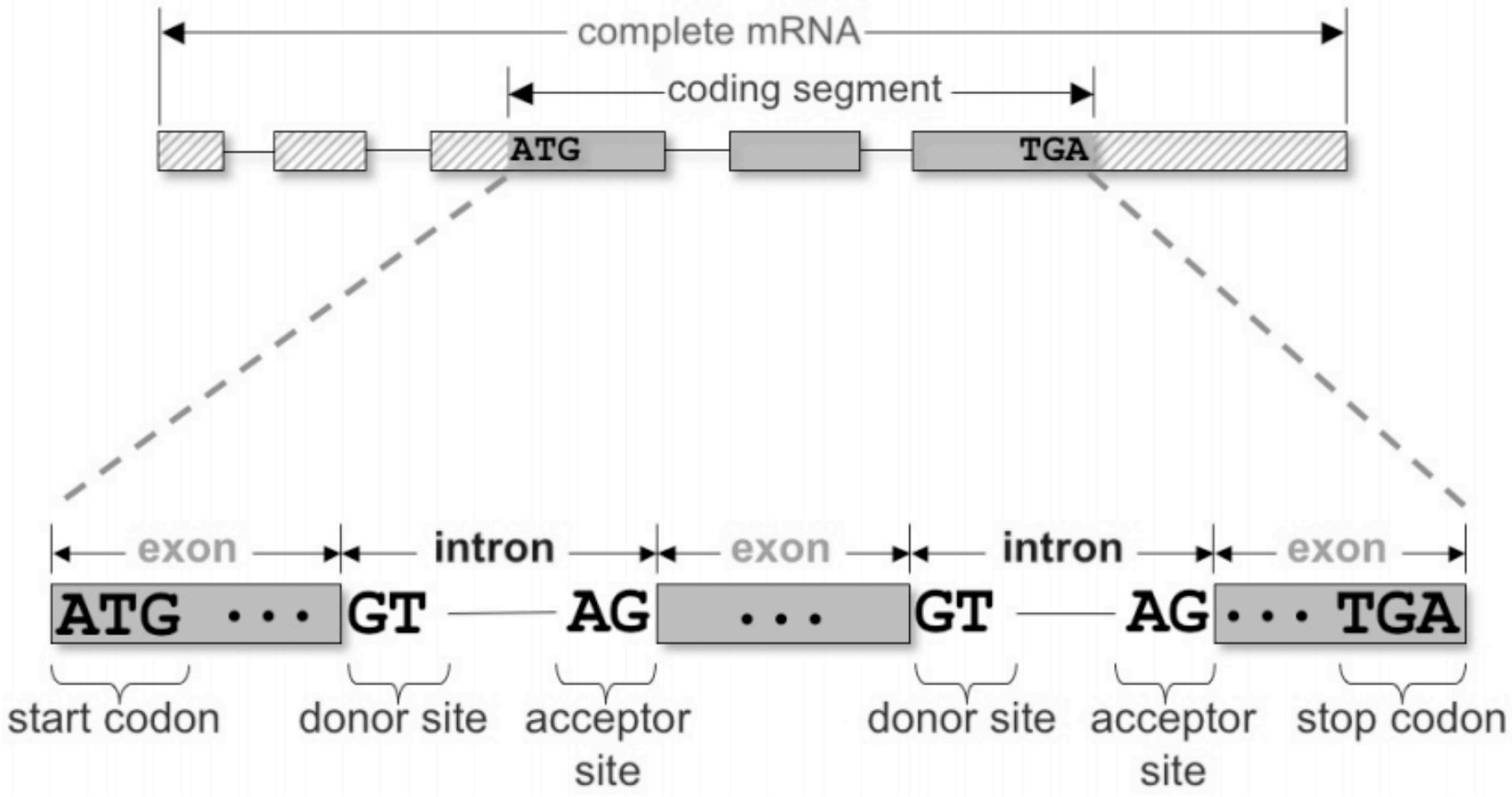


Donor Splice Site or 5'Splice Site

Acceptor Splice Site or 3'Splice Site



Exons and Introns



William H. Majoros, Uwe Ohler
Advancing the State of the Art in Computational Gene Prediction
Proceedings of the Workshop on Knowledge Discovery and Emergent Complexity in Bioinformatics (KDECB / Benelearn'06), Lecture Notes in Bioinformatics, Springer-Verlag, 2007

Gene Structure


```
>HBA1 genomic DNA Gene + 100bps Upstream and Downstream
tccaggccgcgccccgggctccgcgccagccaatgagcgccgcccggccg
ggcgtgcccccgcgccccaagcataaaccctggcgcgctcgcggcccggc
actcttctggtccccacagactcagagagaacccaccatggtgctgtctc
ctgccgacaagaccaacgtcaaggccgcctggggtaaggtcggcgcgcac
gctggcgagtatggtgcggaggccctggagaggtgaggctccctcccctg
ctccgacccgggctcctcgcccgcccggacccacaggccaccctcaaccg
tcctggccccggacccaaaccccacccctcactctgcttctccccgcagg
atgttcctgtccttccccaccaccaagacctacttcccgcacttcgacct
gagccacggctctgcccaggttaagggccacggcaagaaggtggccgacg
cgctgaccaacgccgtggcgcacgtggacgacatgcccaacgcgctgtcc
gccctgagcgacctgcacgcgcacaagcttcgggtggacccggtcaactt
caaggtgagcggcgggccgggagcgatctgggtcgaggggcgagatggcg
ccttcctcgcagggcagaggatcacgcgggttgcgggaggtgtagcgcag
gcggcggctgcgggcctgggccctcggccccactgaccctcttctctgca
cagctcctaagccactgcctgctggtgaccctggccgcccacctccccgc
cgagttcacccctgcggtgcacgcctccctggacaagttcctggcttctg
tgagcaccgtgctgacctccaaataccgttaagctggagcctcggtggcc
atgcttcttgccccttgggcctcccccagcccctcctccccttcctgca
cccgtacccccgtggtctttgaataaagtctgagtgggcggcagcctgtg
tgtgcctgagtttttccctcagcaaacgtgccaggcatgggcgtggaca
gcagctgggacacacatggctagaacctctctgcagctggat
```

Gene Structure

>HBA1 Exons in upper case
tccaggccgcgccccgggctccgcgccagccaatgagcgccgcccggccg
ggcgtgcccccgcgccccaagcataaaccctggcgcgctcgcggcccggc
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTC
CTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCAC
GCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGgtgaggctccctcccctg
ctccgacccgggctcctcgcccgcccggacccacaggccaccctcaaccg
tcctggccccggacccaaaccccacccctcactctgcttctccccgcagG
ATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCT
GAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACG
CGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCC
GCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTT
CAAGgtgagcggcgggccgggagcgatctgggtcgaggggcgagatggcg
ccttcctcgcagggcagaggatcacgcgggttgcgggaggtgtagcgcag
gcggcggctgcgggcctgggccctcggccccactgaccctcttctctgca
cagCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTG
TGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCA
CCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCagcctgtg
tgtgcctgagtttttccctcagcaaacgtgccaggcatgggcgtggaca
gcagctgggacacacatggctagaacctctctgcagctggat

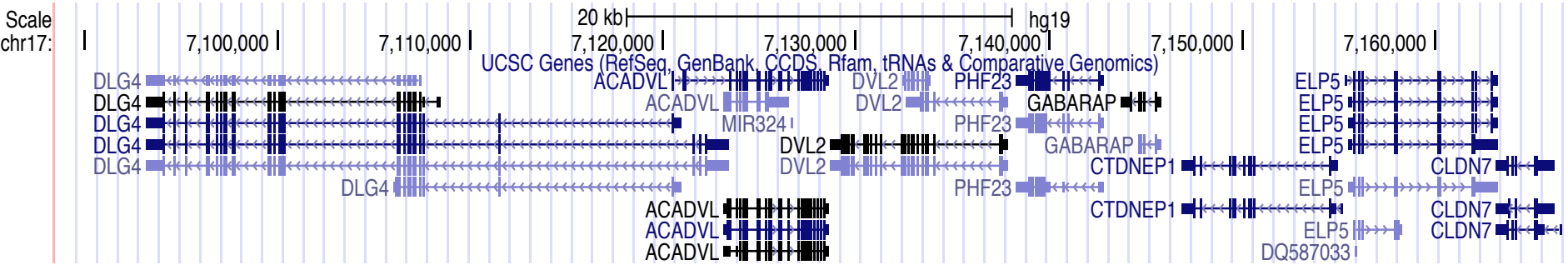
Gene Structure

>HBA1 Coding Sequence in red
tccaggccgcgccccgggctccgcgccagccaatgagcgccgcccggccg
ggcgtgcccccgcgccccaagcataaaccctggcgcgctcgcggcccggc
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTC
CTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCAC
GCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGgtgaggctccctcccctg
ctccgacccgggctcctcgcccgcccggacccacaggccaccctcaaccg
tcctggccccggacccaaaccccacccctcactctgcttctccccgcagG
ATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCT
GAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACG
CGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCC
GCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTT
CAAGgtgagcggcgggccgggagcgatctgggtcgaggggcgagatggcg
ccttcctcgcagggcagaggatcacgcgggttgcgggaggtgtagcgcag
gcggcggctgcgggcctgggccctcggccccactgaccctcttctctgca
cagCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTG
TGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCA
CCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCagcctgtg
tgtgcctgagtttttccctcagcaaacgtgccaggcatgggcgtggaca
gcagctgggacacacatggctagaacctctctgcagctggat

Gene Structure


```
>Protein sequence uc002cfx.1 (HBA1) length=142
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
```

Genes in both strands



Alternative Splicing

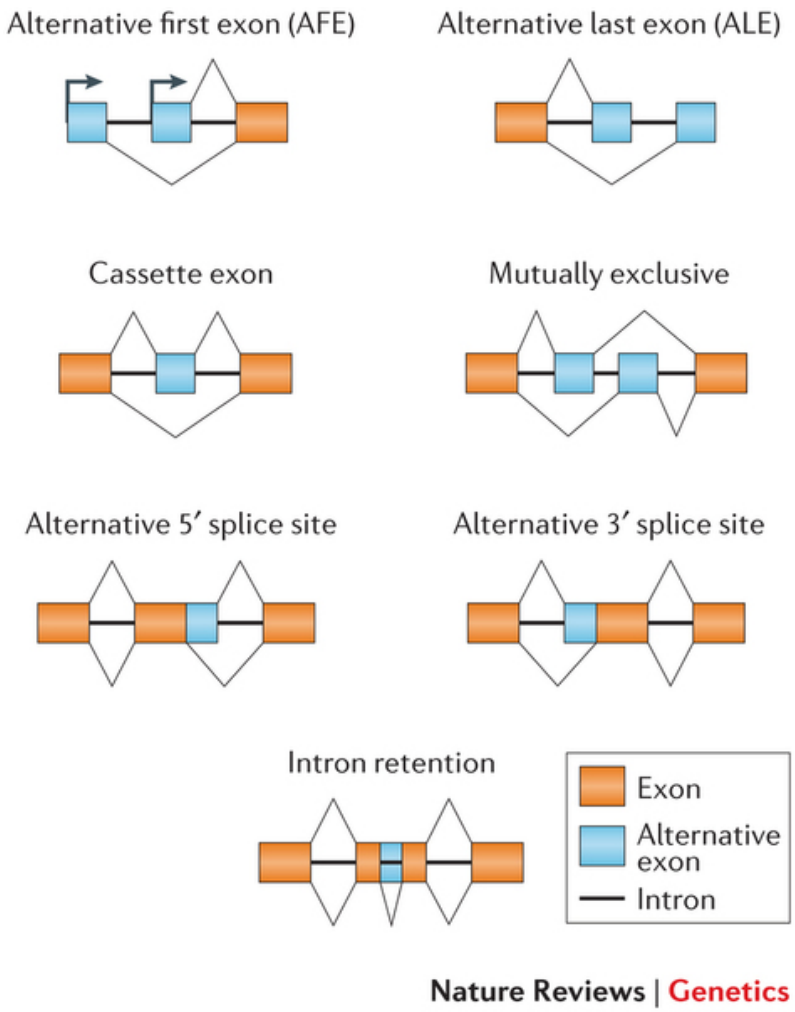
- Alternative splicing (AS), the process in which the exons of the pre-RNAs are spliced in different combinations to produce distinct mRNA that lead to structurally and functionally protein variants.

- This process is most extensively used in higher eukaryotic organisms and leads to macromolecular and cellular complexity.

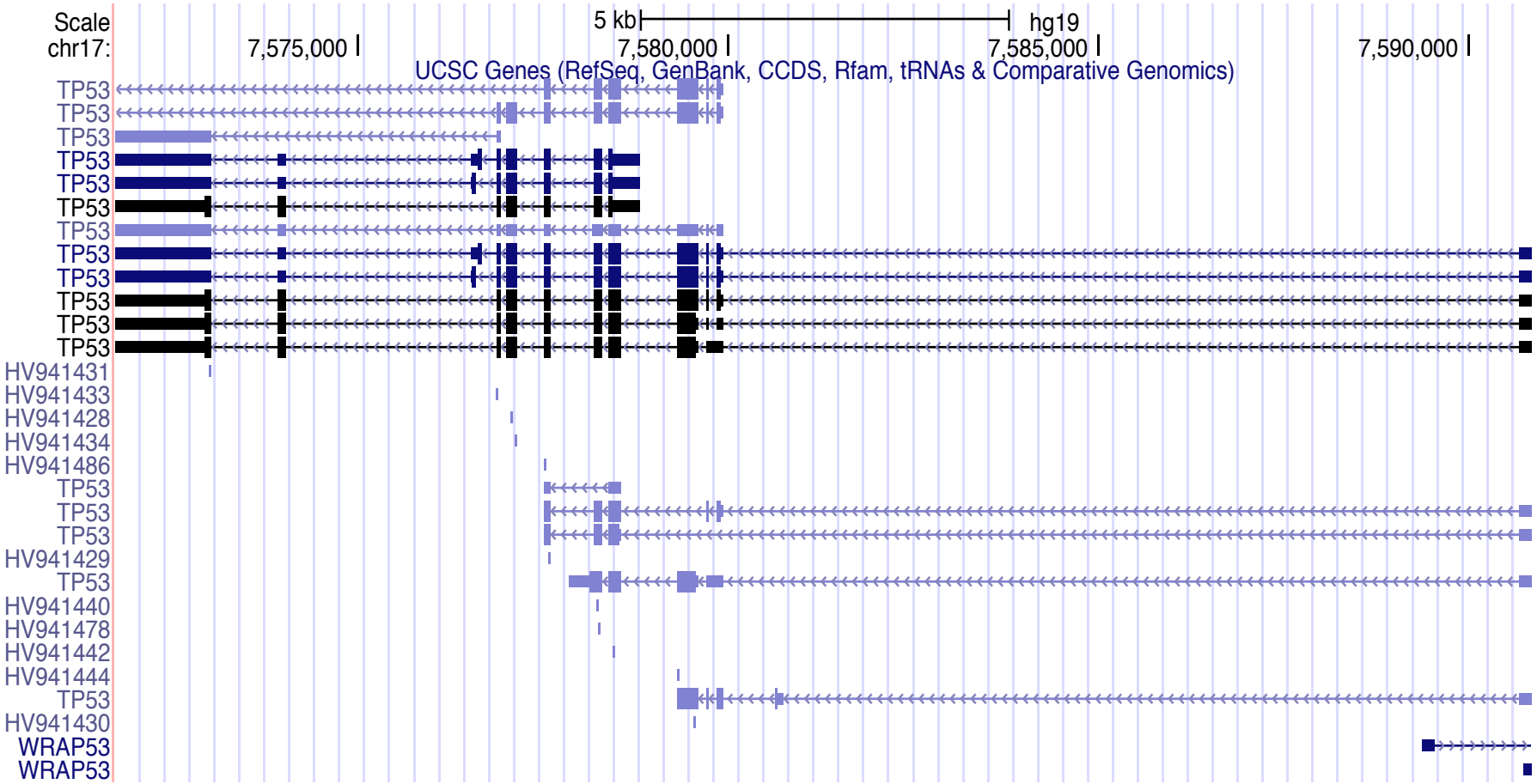
- Recent studies with high-throughput RNA sequencing suggest that alternative splicing is very common in human cells and that 90–95% of human multi-exon genes produce transcripts that are alternatively spliced.

AS is combinatorial

- Combinatorial splicing leads to the generation of of multiple isoforms from a single gene.

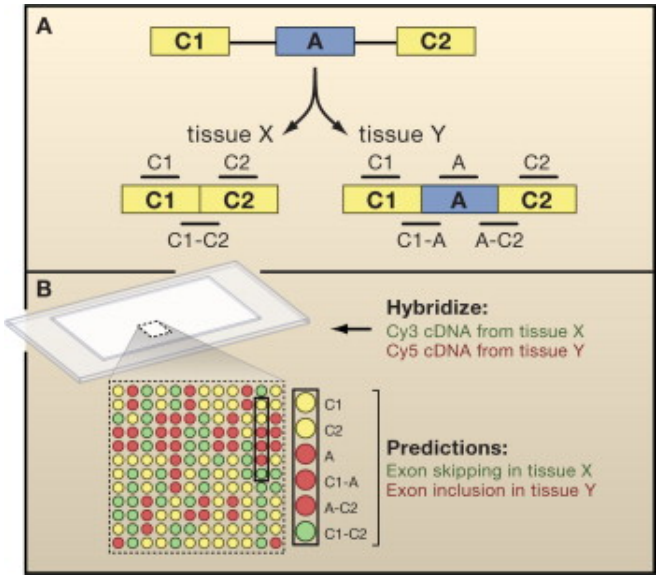


AS in TP53

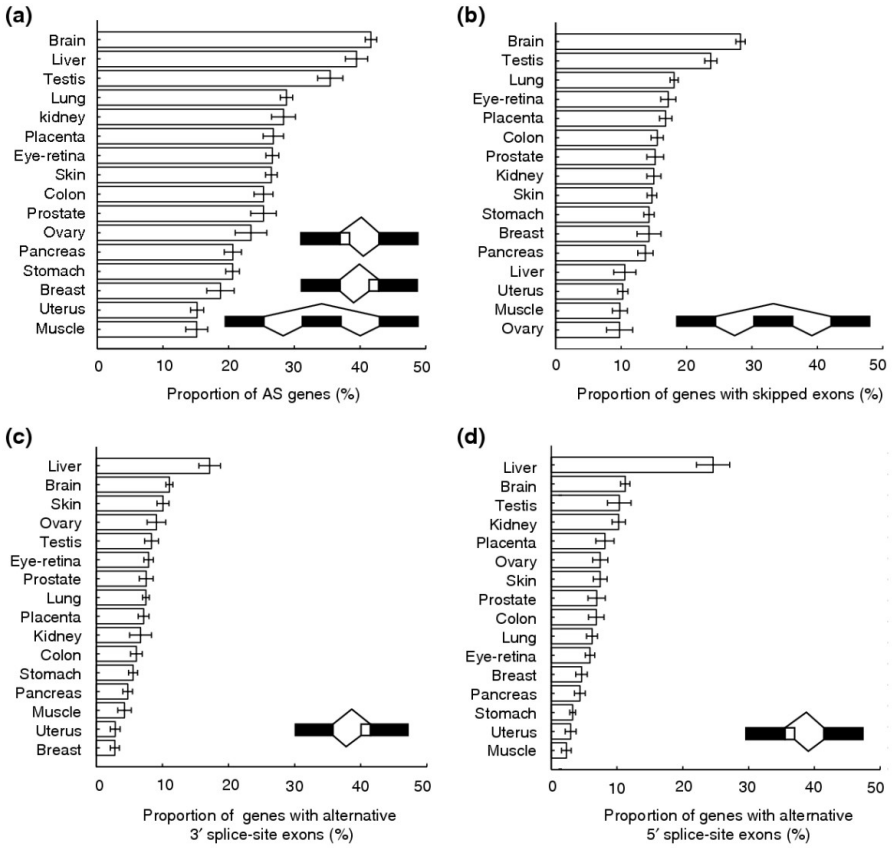

Scale
chr17:
5 kb
hg19
7,575,000
7,580,000
7,585,000
7,590,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
TP53
TP53
TP53
TP53
TP53
TP53
TP53
TP53
TP53
TP53
TP53
HV941431
HV941433
HV941428
HV941434
HV941486
TP53
TP53
TP53
HV941429
TP53
HV941440
HV941478
HV941442
HV941444
TP53
HV941430
WRAP53
WRAP53

AS and Tissue Specificity

- AS is widely used by higher eukaryotes to generate different protein isoforms in specific cell or tissue types.



Benjamin J. Blencowe
Alternative Splicing: New Insights from Global Analyses. Cell 2007



Gene Yeo, Dirk Holste, Gabriel Kreiman and Christopher B Burge. Variation in alternative splicing across human tissues. Genome Biology 2004