Some Basics of Molecular Biology

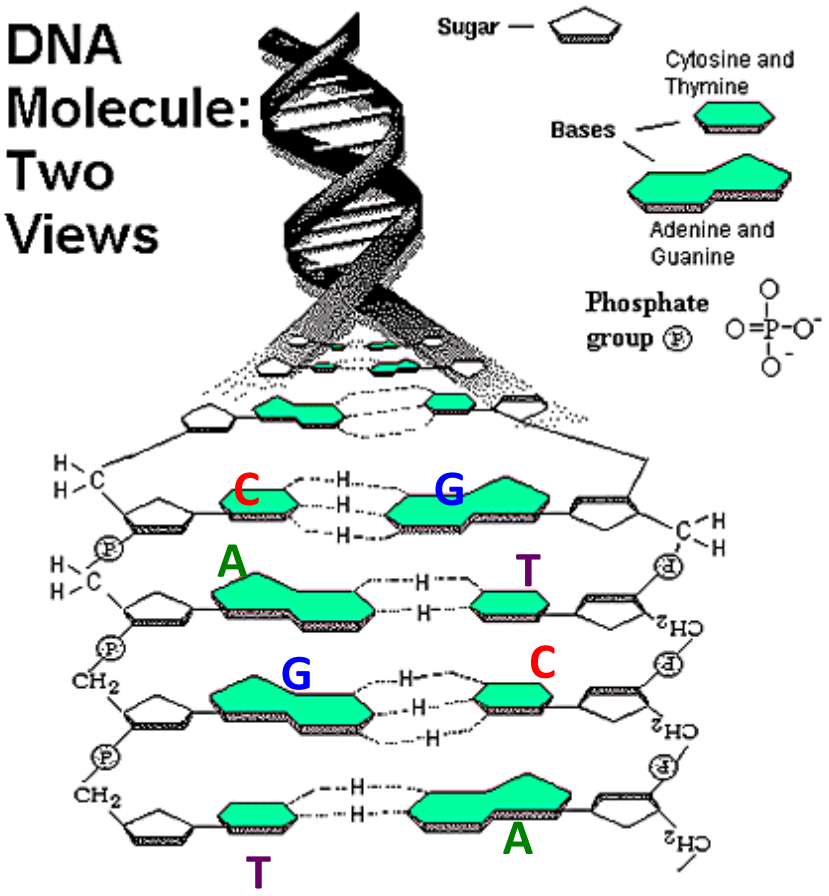
Hubert Rehrauer

Messages

- DNA / genome size /evolution
- genes – concept
- gene regulation
- RNA species; junk
- RNA numbers
- RNA degradation

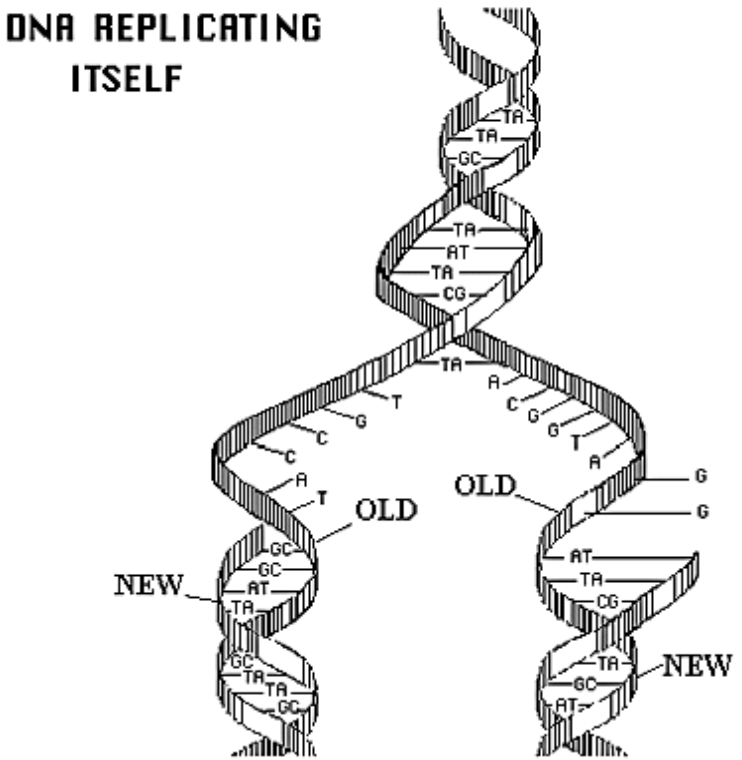
What is DNA?

- A long backbone of sugars with nucleotides attached
  - **Adenine (A)**
  - **Guanine (G)**
  - **Cytosine (C)**
  - **Thymine (T)**
- It can form a self-complementary **double helix**
- In living organisms, the DNA is the carrier of the hereditary information, it is the **source code** of life

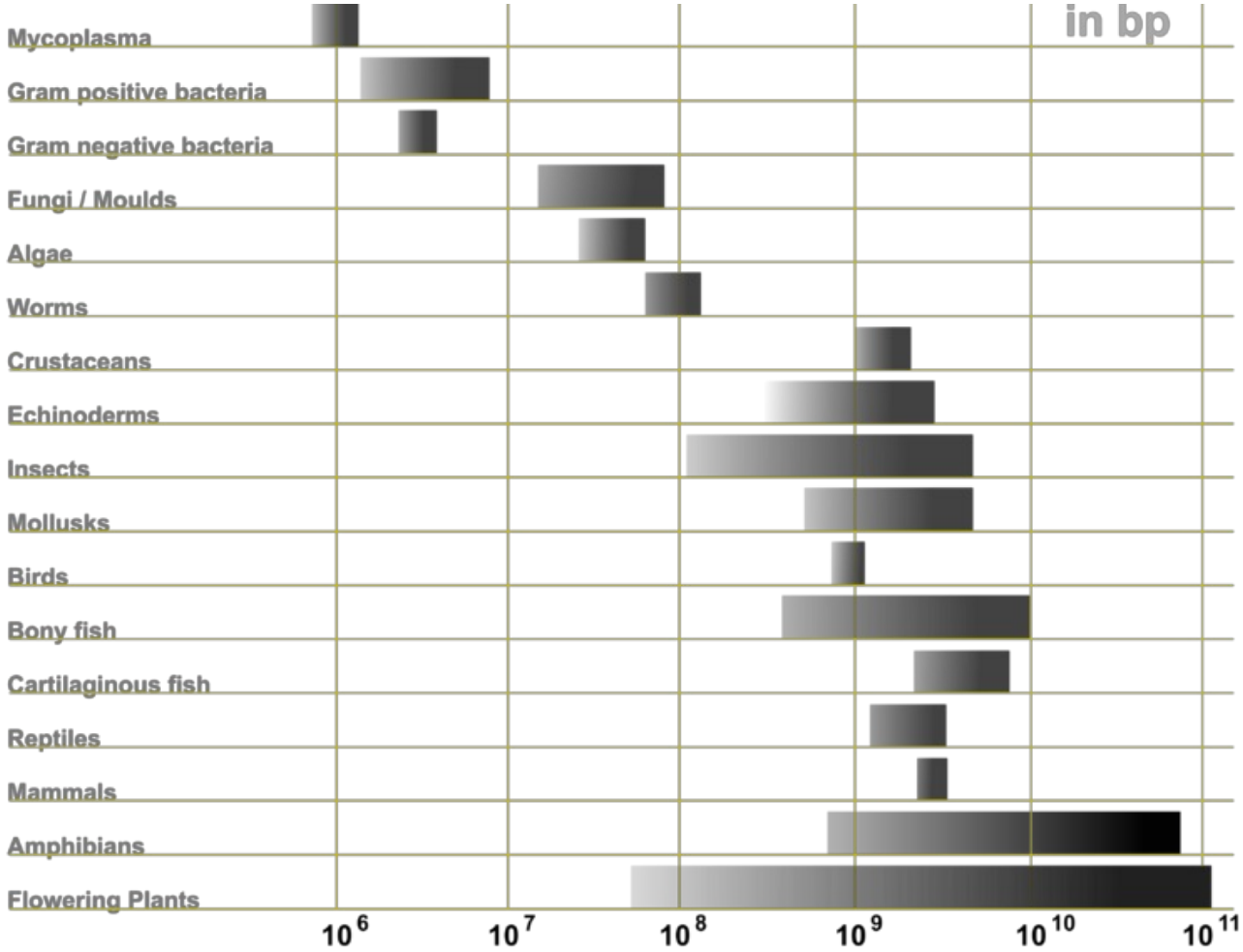


DNA replication

- The helix becomes unzipped and each strand acts as a template for a new complementary strand of DNA



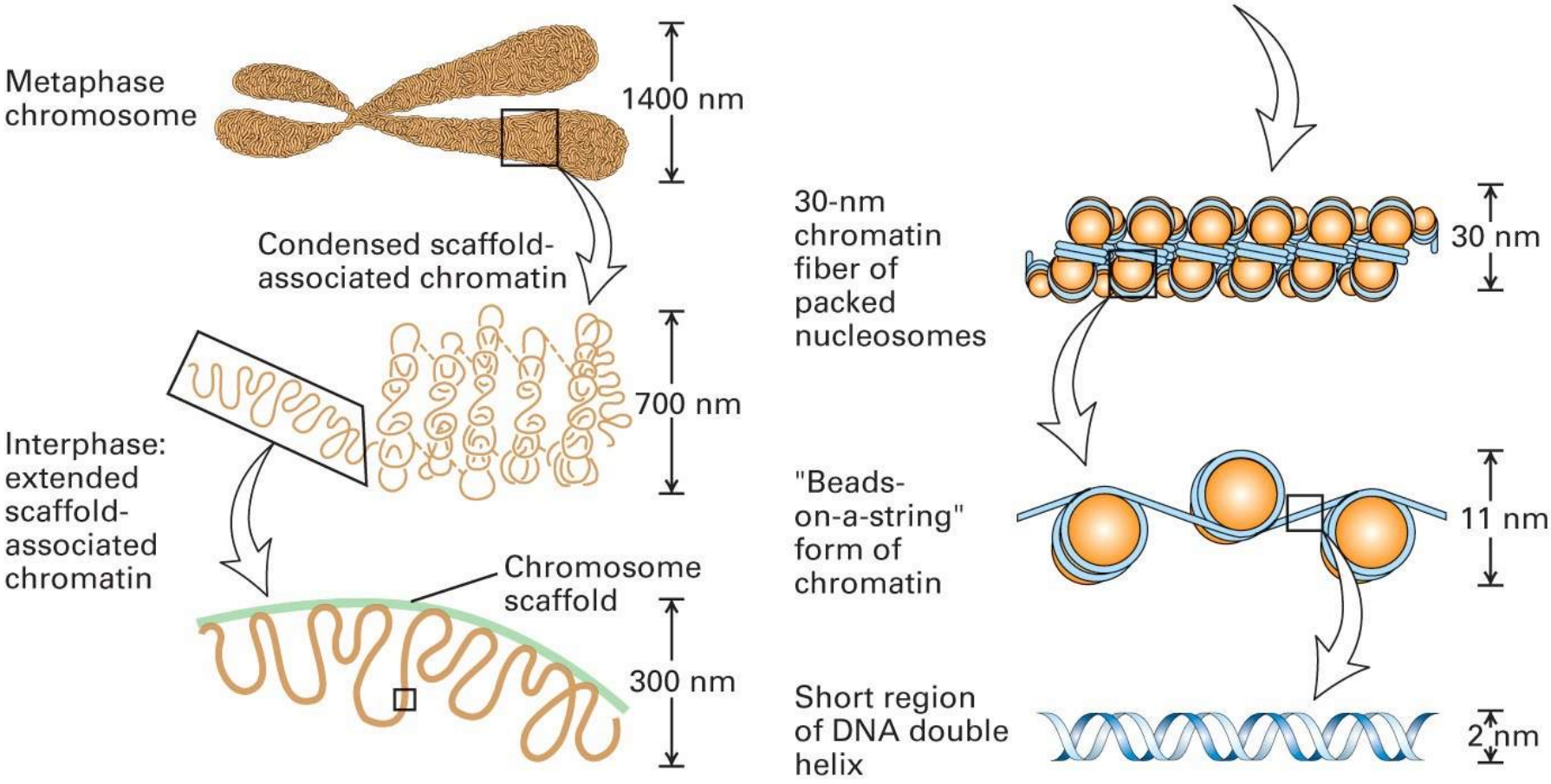
Genome Sizes



The size of the human genome is 3.2 billion base pairs. The length of this DNA string is approx. 2m.

http://en.wikipedia.org/wiki/Genome_size

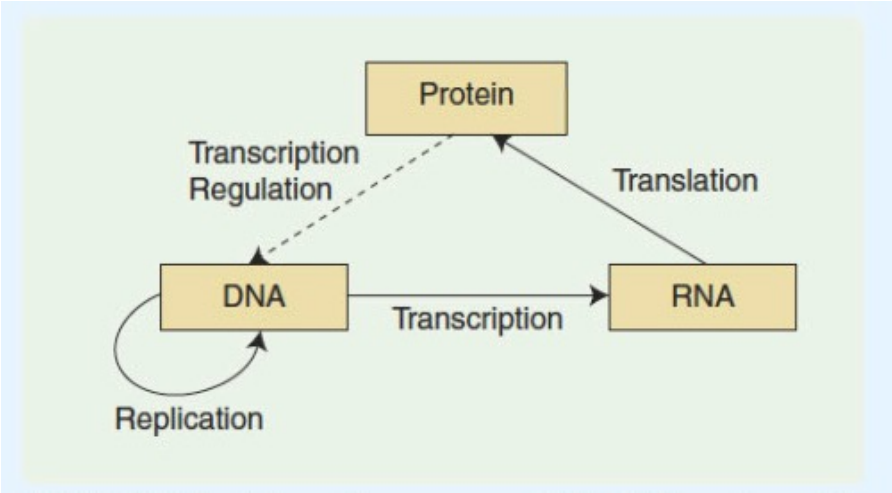
DNA Superstructure



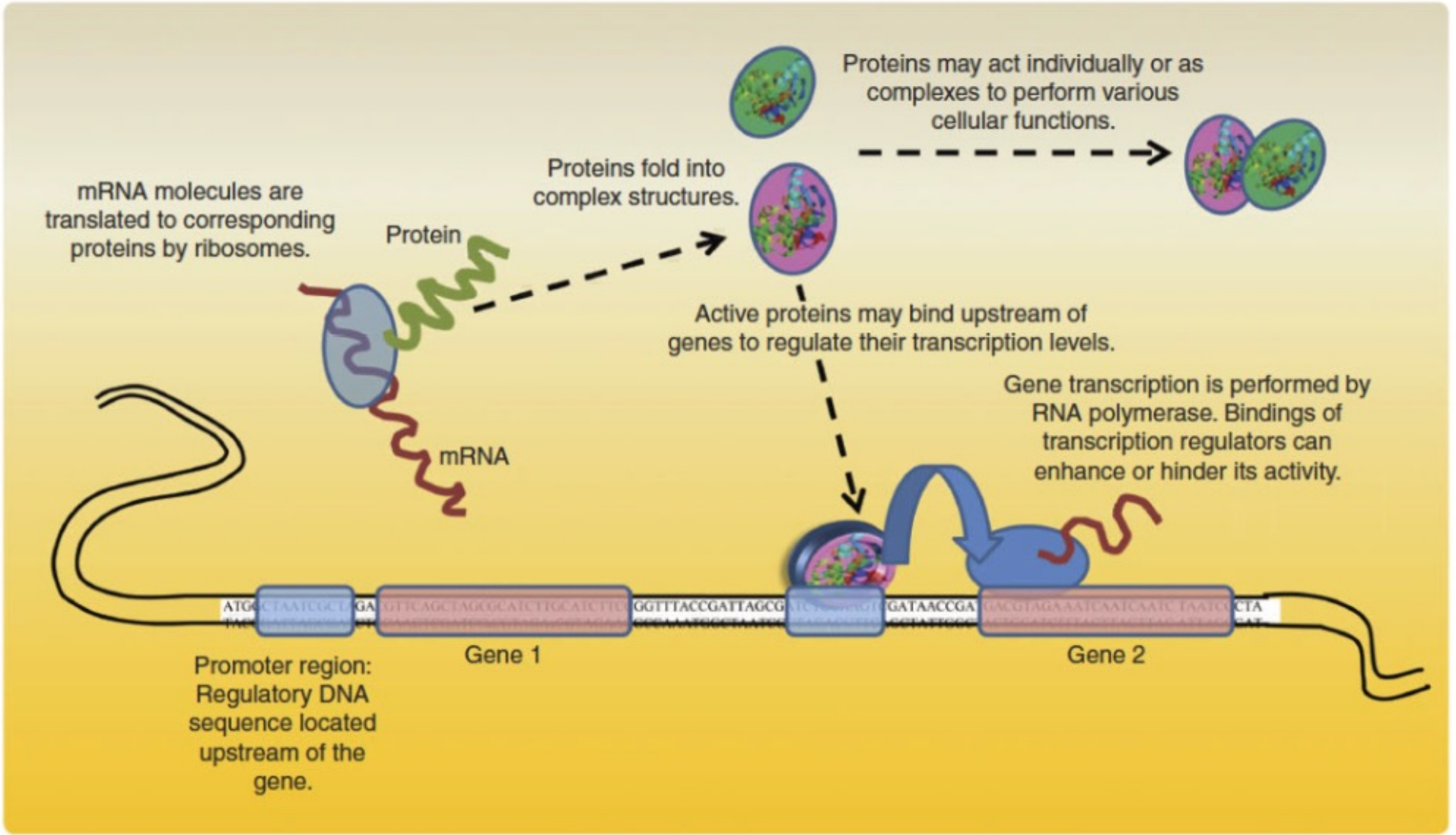
Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

Genes

- A gene is a region of DNA (a locus) that controls a hereditary characteristic
- protein coding genes are transcribed into messenger RNA which is then translated into protein.
- In humans, protein coding genes constitute only ~2% of the human genome (precise number still under investigation)

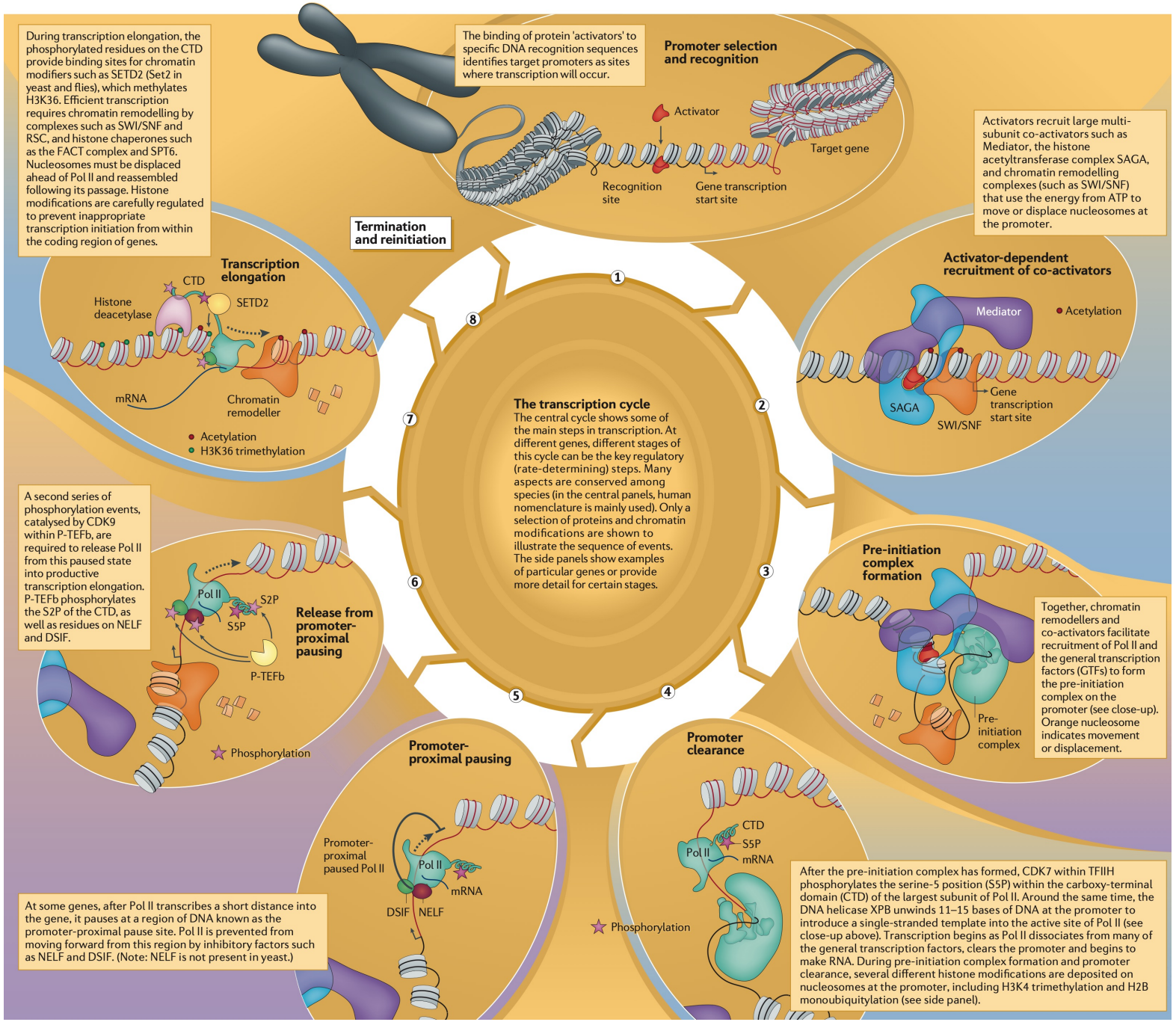


The Central Dogma



Elvitigala et al. 2010

Transcription Cycle



Genome Sizes and Transcriptome Complexity

- larger genomes and complex organisms have emerged by **evolution**
- "**Nothing in Biology Makes Sense Except in the Light of Evolution**" -- Theodosius Dobzhansky, 1973

- genome sizes vary much (in some cases even for closely related species)
- genome sizes seems not strictly correlated with organisms

→ other functional elements vary: non-coding genes, repeats, ….

- number of protein coding genes is relatively constant across species
- C. elegans worm has only ~1000 cells and about the same number of genes like humans

→alternative splicing

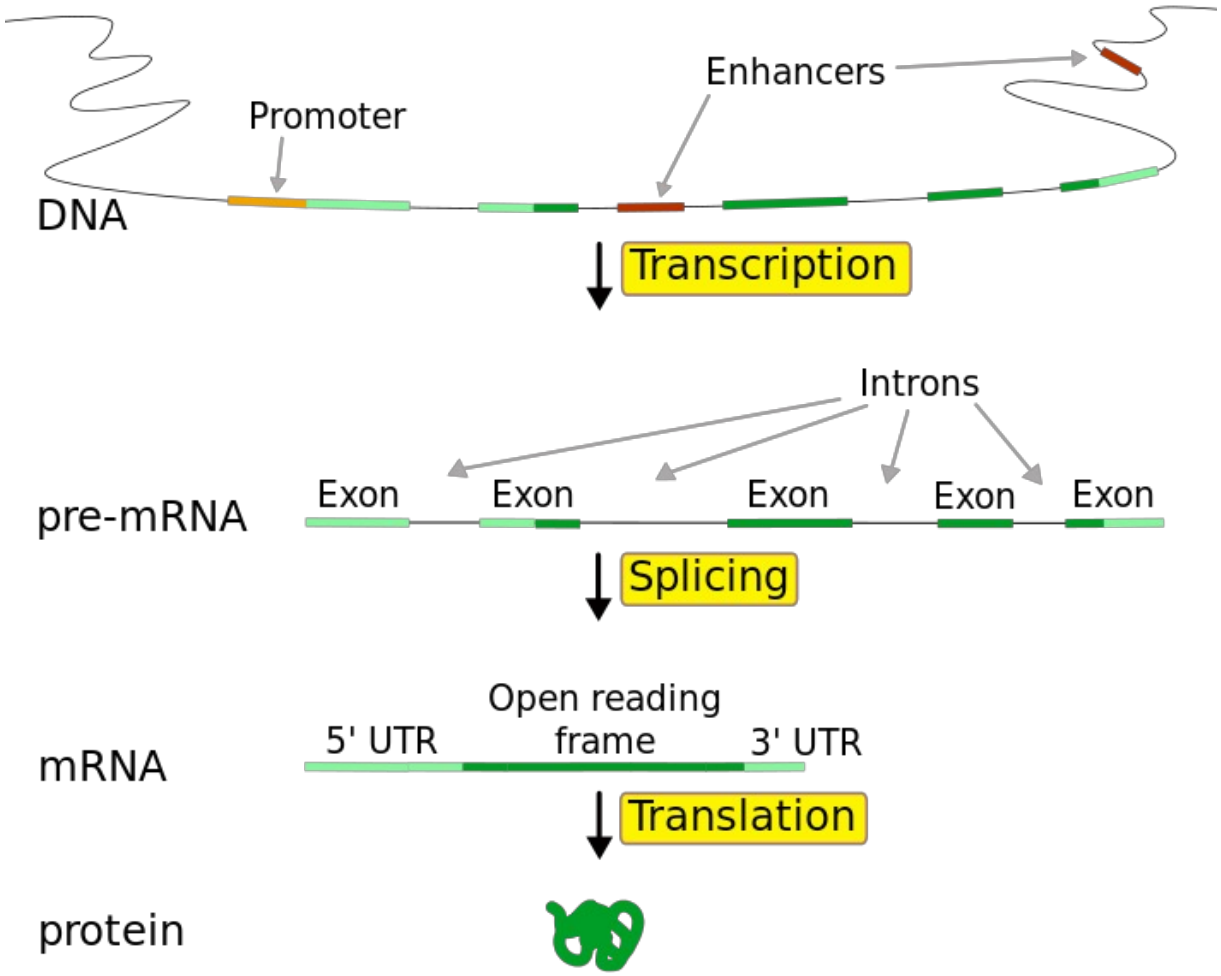
Most of the intergenic DNA is not “junk” DNA but has a function!

Transcription

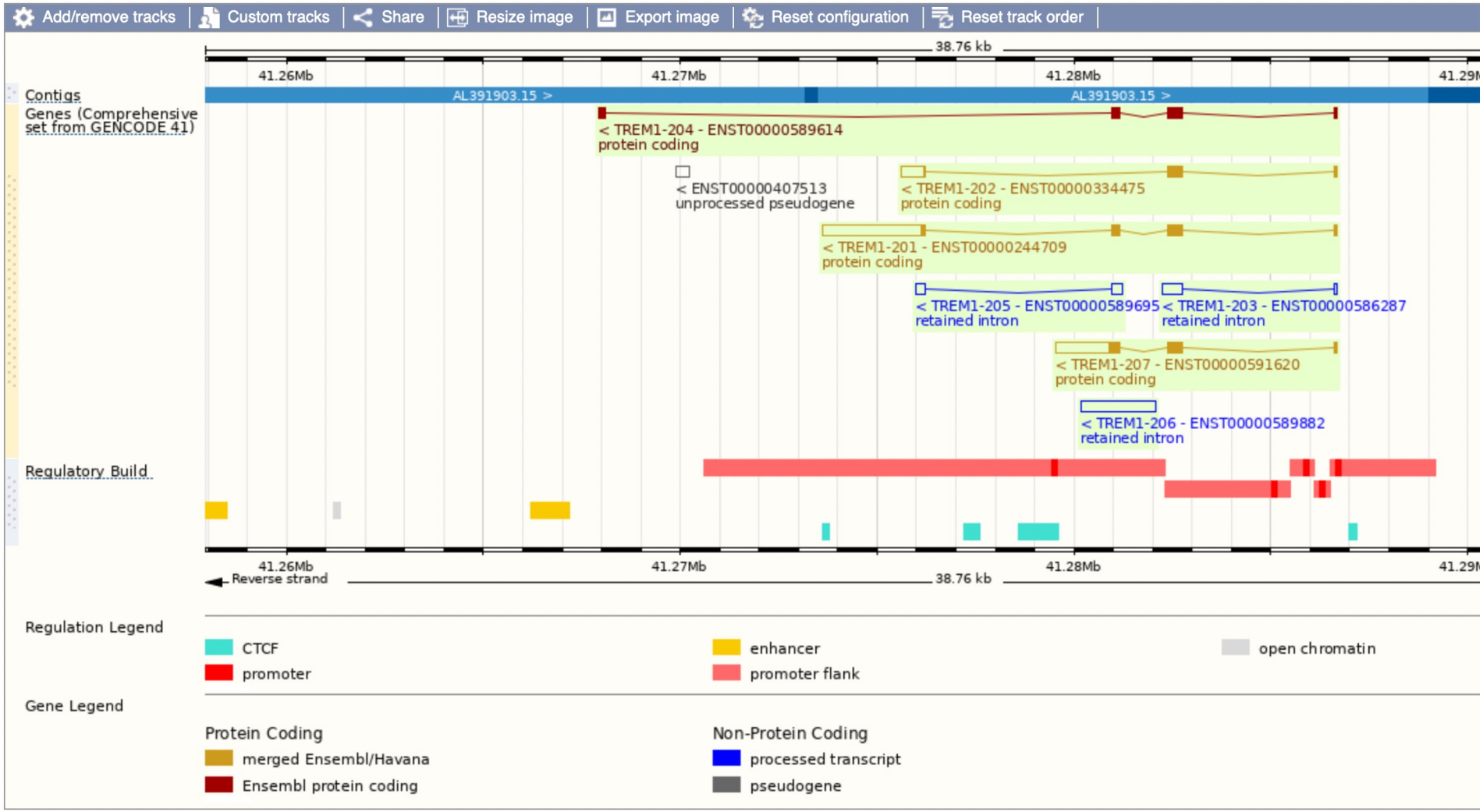
The transcription process generates a messenger RNA molecule from a gene region. RNA is like DNA but

- the sugar-phosphate is different: ribose instead of deoxyribose
- In all places where the DNA has a T the the RNA has a U (uracil)

In higher organisms the protein coding sequences (exons) are interspersed by non-coding sequences (introns) which are spliced out.

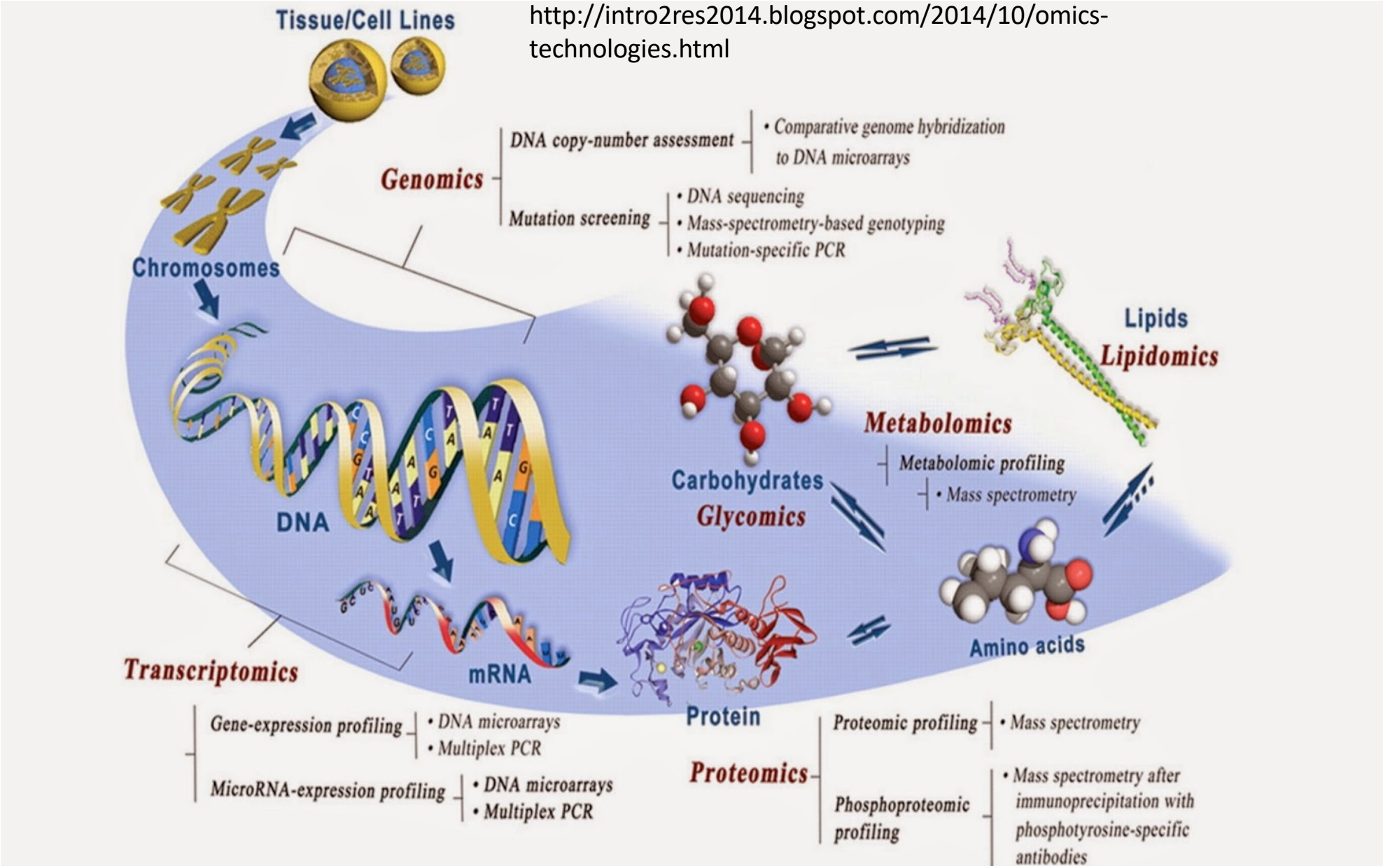


Gene Structure



- introns are much longer than exons
- humans: typical gene length(exonic): 1000 – 2000 bases
- Genome Browsers: UCSC, Ensembl, IGV

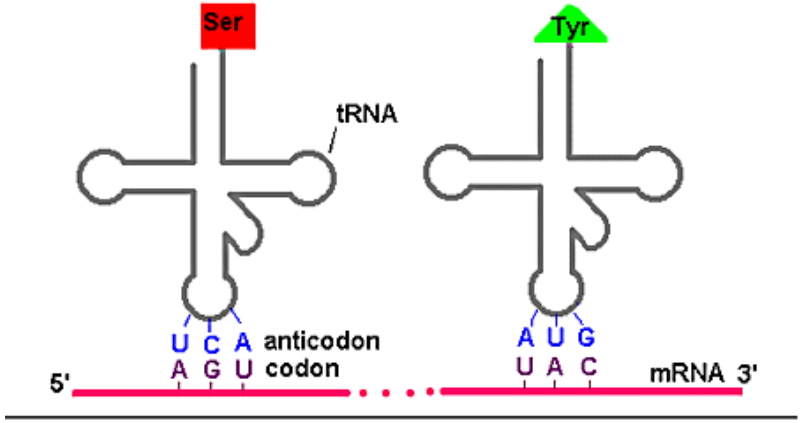
http://intro2res2014.blogspot.com/2014/10/omics-technologies.html



Translation: The Genetic Code

The translation process generates a protein based on the information in the messenger RNA

- A protein is a linear polymer of amino acids linked together by peptide bonds.
- Proteins are the main functional chemicals in the cell, carrying out many functions, for example catalysis of the reactions involved in metabolism.
- Proteins have a complex spatial structure

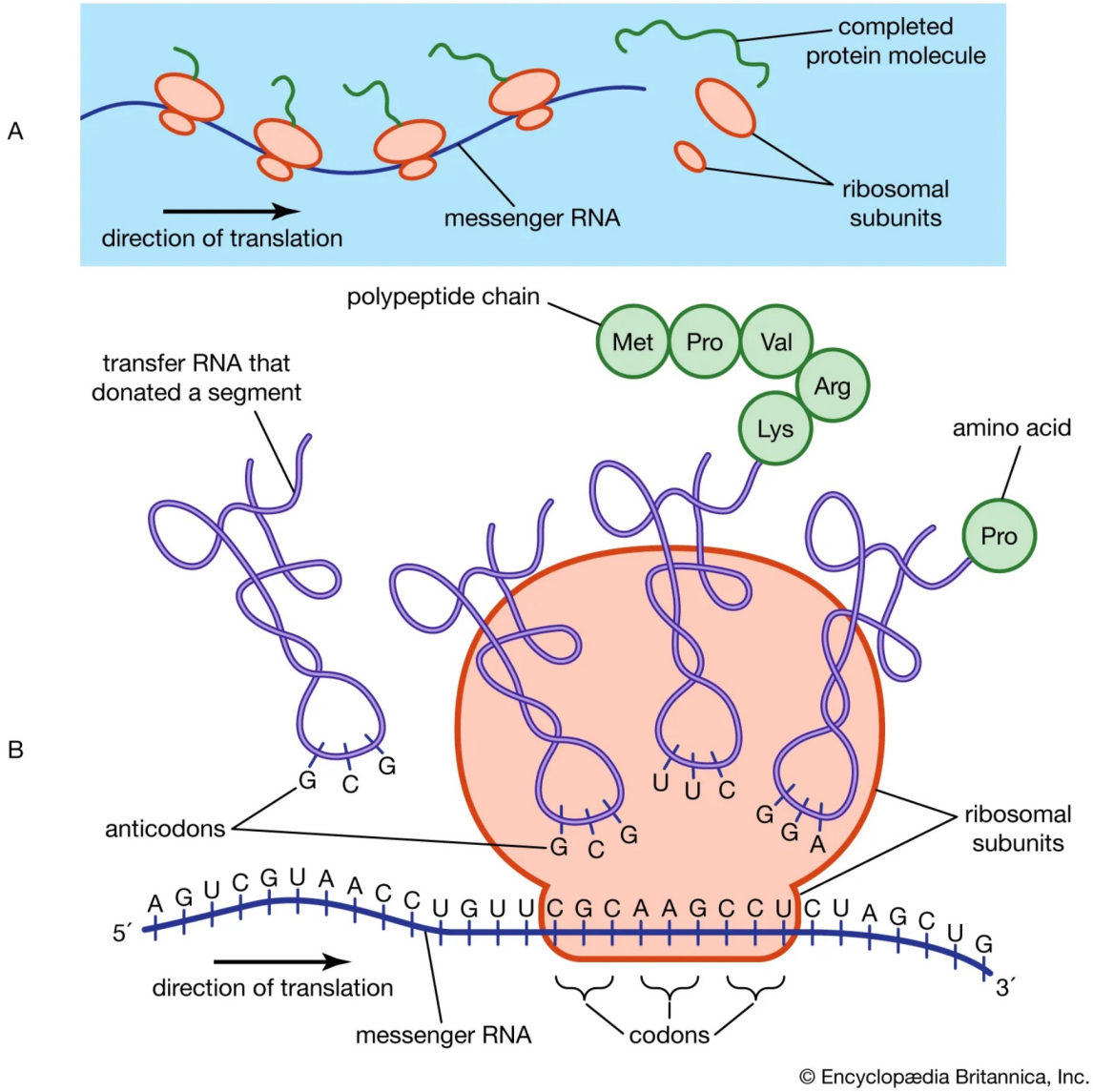


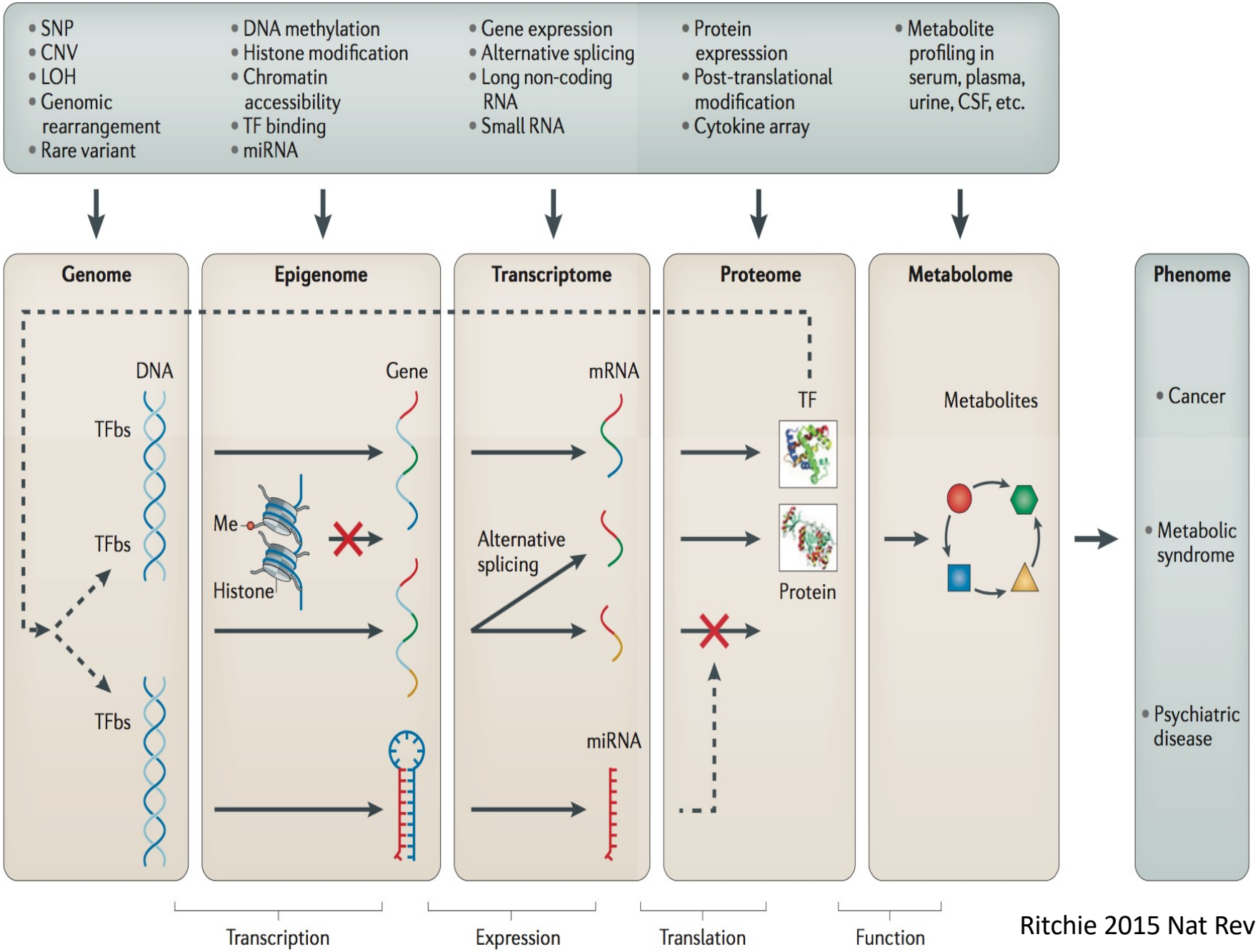
2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

The Genetic Code

Translation

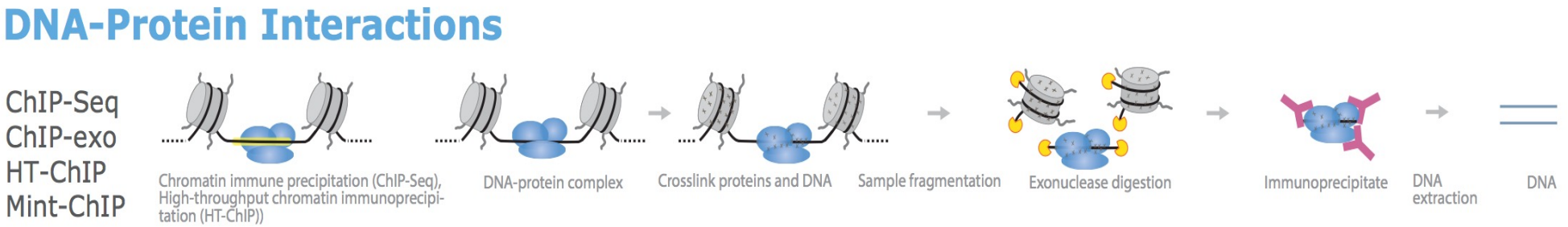




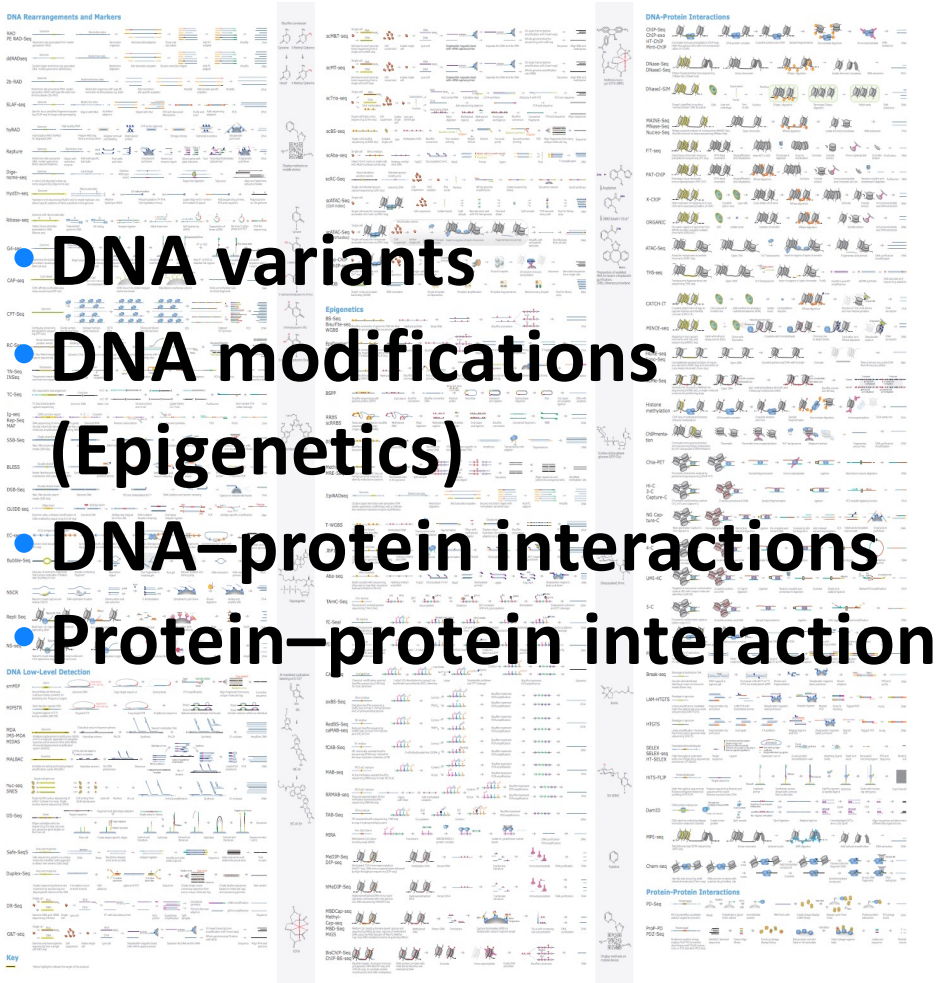
Ritchie 2015 Nat Rev Genetics

NGS Protocols

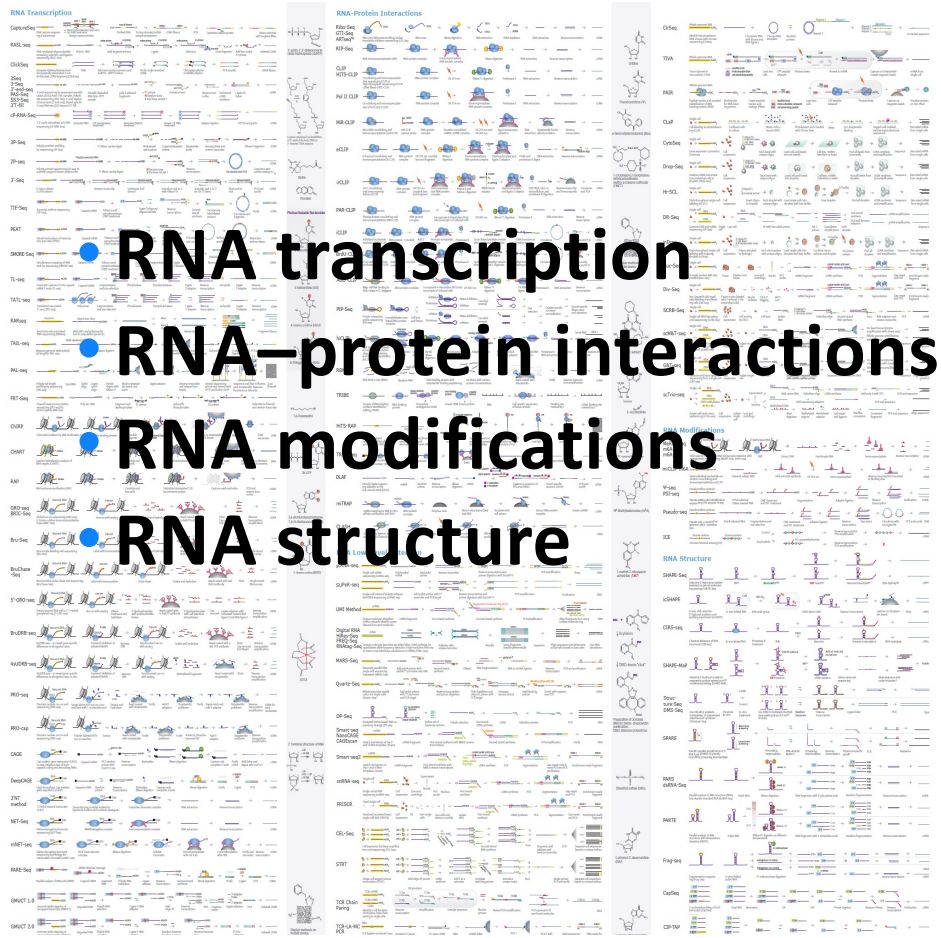
- Example: Preparation of DNA for a ChIP-seq experiment



- The preparation determines how sequenced reads have to be interpreted



- DNA variants
- DNA modifications (Epigenetics)
- DNA–protein interactions
- Protein–protein interactions



- RNA transcription
- RNA–protein interactions
- RNA modifications
- RNA structure

- Big Challenge: targeted molecules may degrade during extraction

NGS Reads

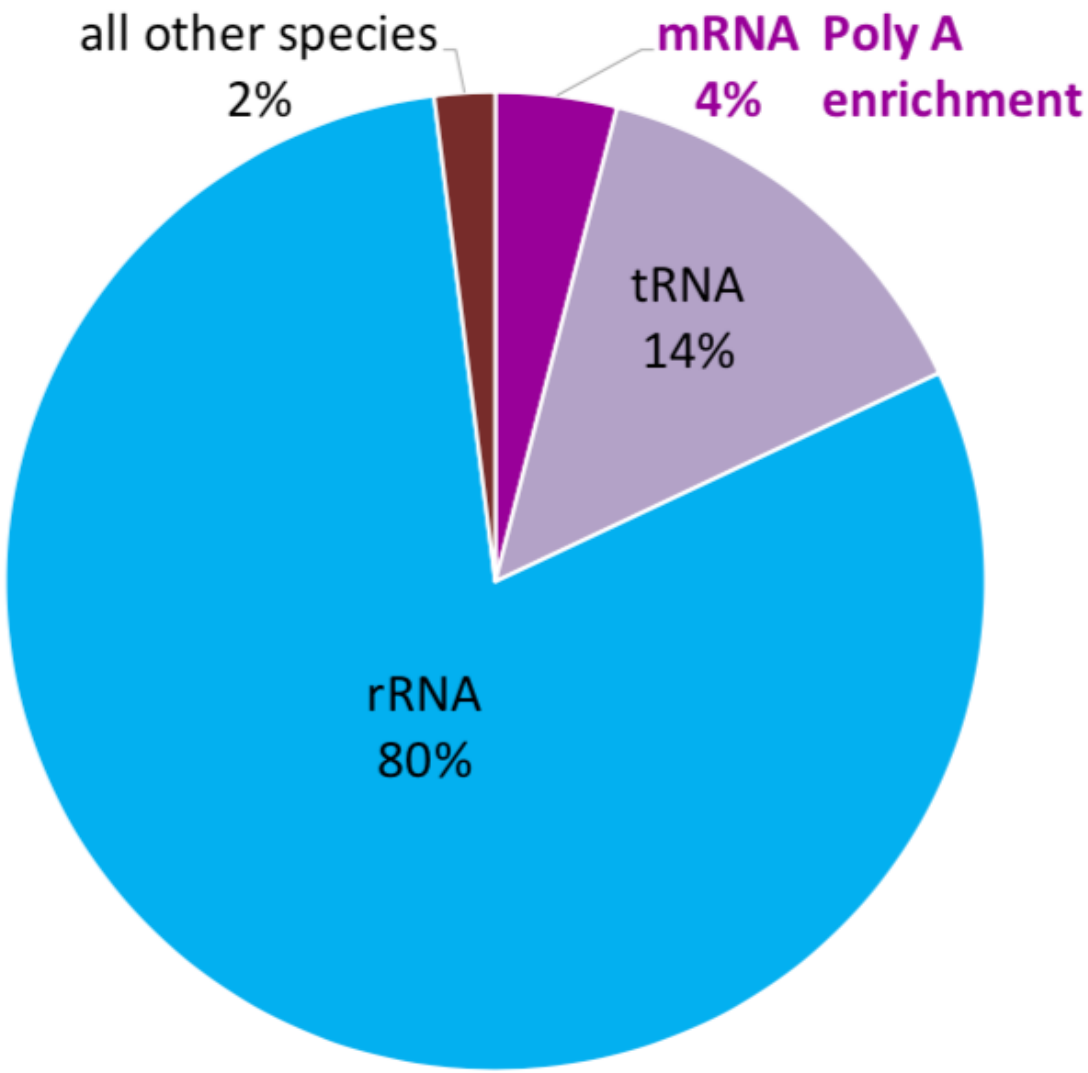
- Extract nucleotide material from specimens
- Select molecules
  - this is the crucial step that determines what information we can learn later from the data
  - Examples: mRNA; mRNA that is currently being translated; DNA positions (promoter regions) where a protein (transcription factor) is bound; cell-free RNA; …
- Add adapters/primers and make the molecules compatible with the sequencer
- Sequencing: Identify the nucleotide sequences

RNA-seq in numbers

- Typical numbers for human cells
- In biology everything is possible!
  Be prepared to encounter cells that have 10x less or 10x more.

- 250’000 – 500’000 mRNA molecules per cell
- ~14% of total RNA is in nucleus
- RNA:DNA in nucleus 1:2
- typical mRNA molecule length ~1900nt

Prevalence of RNA species



- transfer RNA (tRNA) length: 76-90 nt
- ribosomal RNA (rRNA), different subunits
  - 5S: 121 nt
  - 18S: 1869 nt
  - 28S: 5070 nt

mRNA abundance

Abundance class	Copies/cell	Number of different messages/cell	Abundance of each message
Low	5–15	11,000	<0.004%
Intermediate	200–400	500	<0.1%
High	12,000	<10	3%

- A highly expressed gene typically does not contribute more than 3% of all transcripts
- In extreme cases it may be 10 times more
- Large dynamic range

RNA-seq experiment of bulk tissue

- ~ 1 Mio cells averaged
- 100 – 1000 ng of total RNA
- > =100 billion molecules as starting material
- Steps:
  - Poly-A based enrichment oligo(dT) with beads
  - fragmentation
  - PCR Amplification (e.g. ~7 – 20 cycles)
    - 10 cycles: every molecule has ~1000 duplicates
  - sequencing (sampled readout of 20 Mio processed molecules)

- Note: some protocols attach a random sequence before amplification
  - Recommended for cases low input cases, e.g. starting amount < 50ng

PCR amplification

- PCR amplification is assumed to amplify any fragment equally well
- BUT PCR can introduce bias
  - Amplification depends on GC content of fragment (extremely high/low GC content is underrepresented)
  - Shorter fragments are amplified more efficiently
  - Long fragments may be amplified incompletely
  - …

  → Bias depends on number of cycles

  → Bias needs to be kept consistent