Computational Genomics

http://www.csb.pitt.edu/ComputationalGenomics/

Maria Chikina
mchikina@pitt.edu
3064 BST-3

Seyoung Kim
sssykim@cs.cmu.edu
7721 Gates Hillman Complex

Register for Piazza and use it for all communication

Topics

- Sequence analysis
- Gene expression/multivariate data analysis
- Population Genetics
- Systems biology

Class grades

- Problem sets (40%)
- Midterm (30%)
- Project (25%)
- Class participation (5%)

Assignments

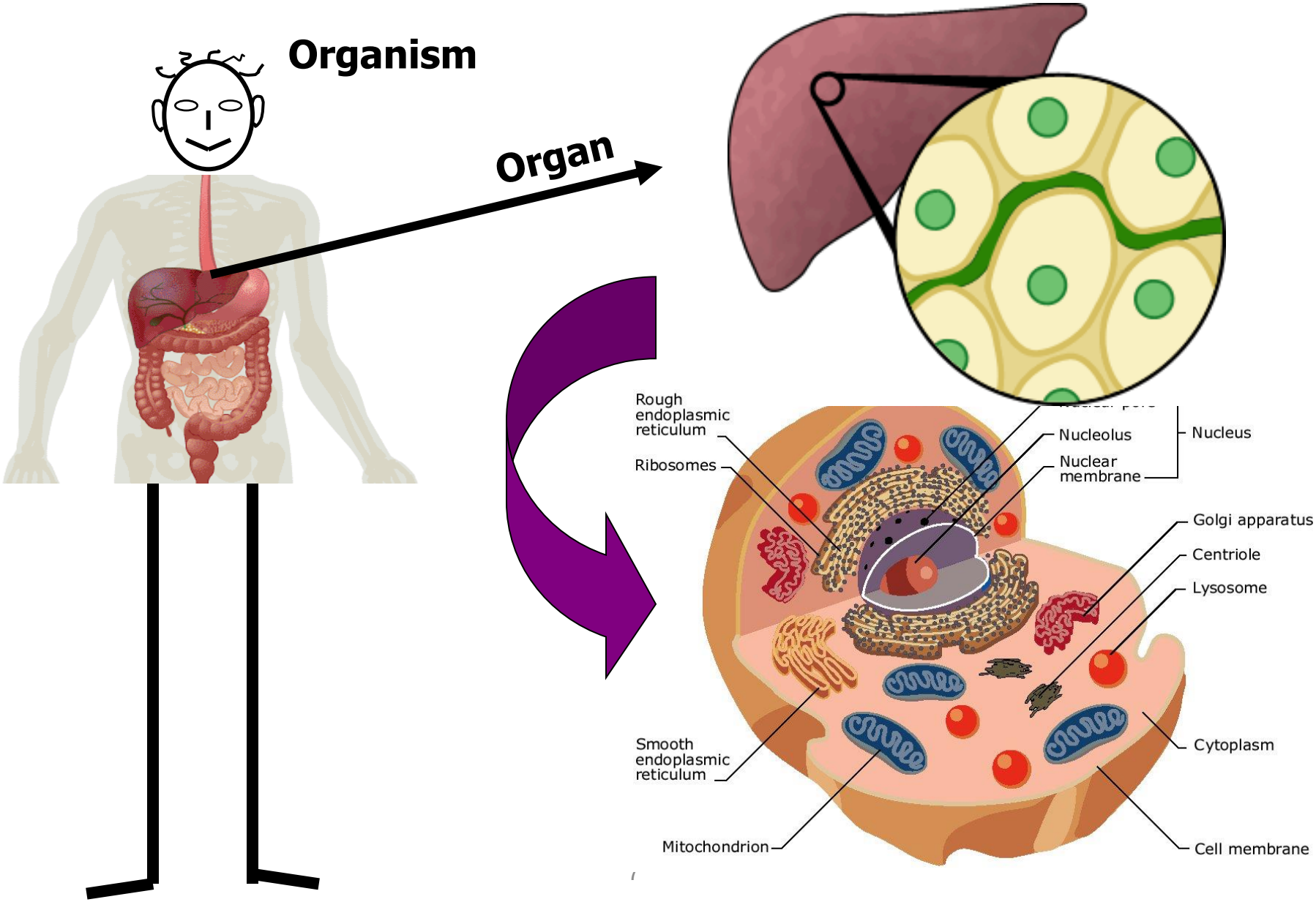
- 4 assignments
- Some programming component
- Languages: python and R

High level and brief intro to molecular biology and genomics

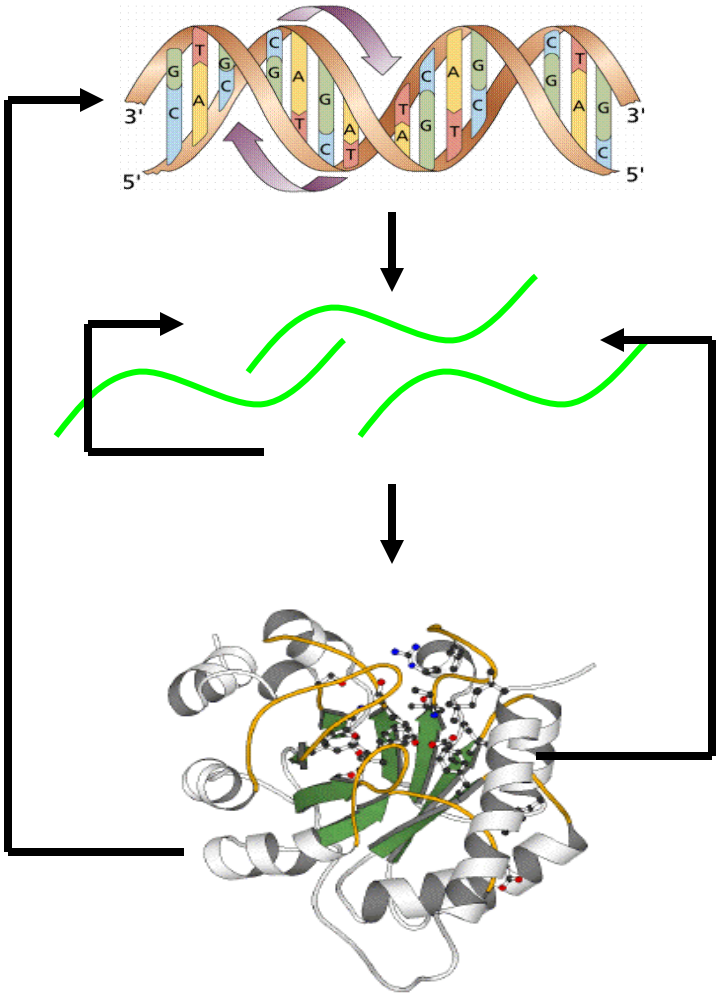
Types of Cells

- Prokaryotes:
  - Bacteria
  - Do not contain compartments-biochemical soup
- Eukaryotes:
  - Plants, animals, humans
  - DNA resides in the nucleus
  - Highly organized cells
  - Yeast is the model eukaryote

Organism, Organ, Cell



Central dogma



DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

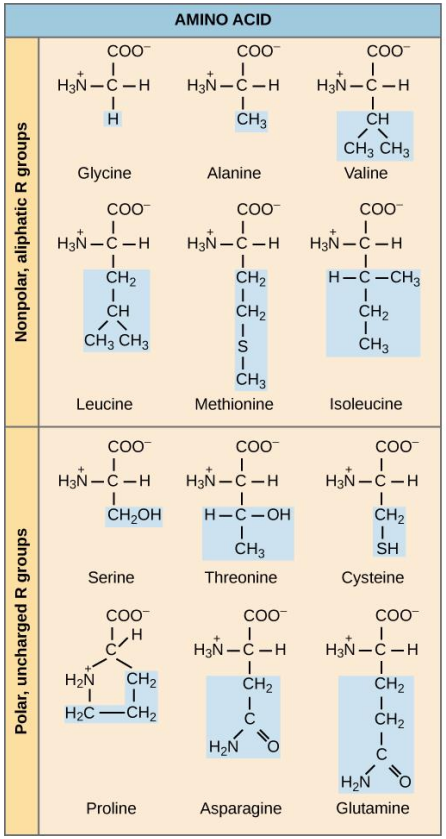
Protein Types and Functions

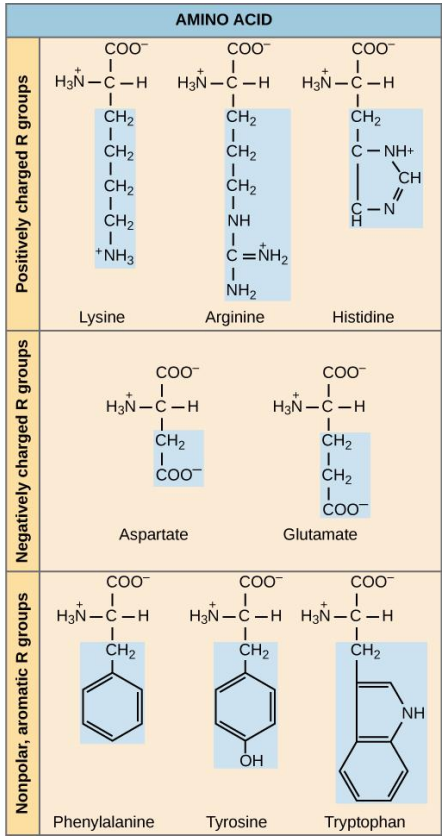
Type	Examples	Functions
Digestive Enzymes	Amylase, lipase, pepsin, trypsin	Help in digestion of food by catabolizing nutrients into monomeric units
Transport	Hemoglobin, albumin	Carry substances in the blood or lymph throughout the body
Structural	Actin, tubulin, keratin	Construct different structures, like the cytoskeleton
Signaling	Hormones, receptors, signal transduction	Protect the body from foreign pathogens
Contractile	Actin, myosin	Effect muscle contraction
Storage	Legume storage proteins, egg white (albumin)	Provide nourishment in early development of the embryo and the seedling

Genes Encode for Proteins

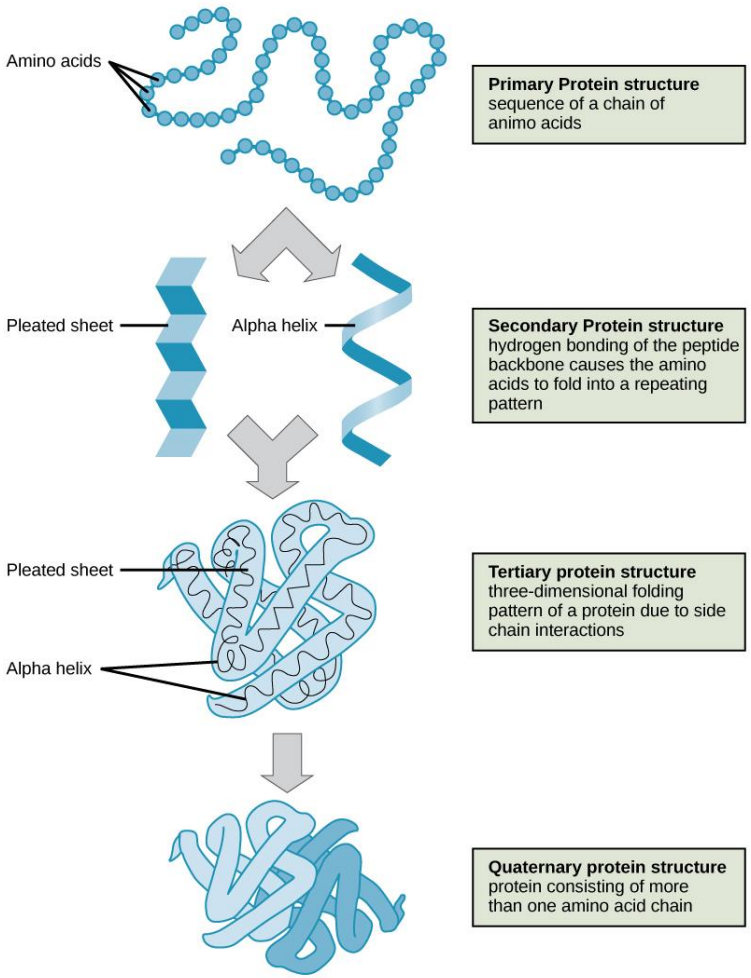
Second Letter

1st letter	U		C		A		G		3rd letter
U	UUU UUC	Phe	UCU UCC UCA UCG	Ser	UAU UAC	Tyr	UGU UGC	Cys	U
	UUA UUG	Leu			UAA UAG	Stop Stop	UGA UGG	Stop Trp	C A G
C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC	His	CGU CGC CGA CGG	Arg	U C A G
					CAA CAG	Gln			
A	AUU AUC AUA	Ile	ACU ACC ACA ACG	Thr	AAU AAC	Asn	AGU AGC	Ser	U C
	AUG	Met			AAA AAG	Lys	AGA AGG	Arg	A G
G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC	Asp	GGU GGC GGA GGG	Gly	U C A G
					GAA GAG	Glu			



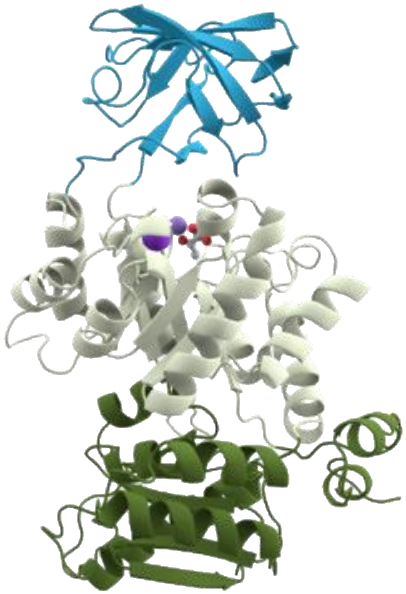


Protein structure

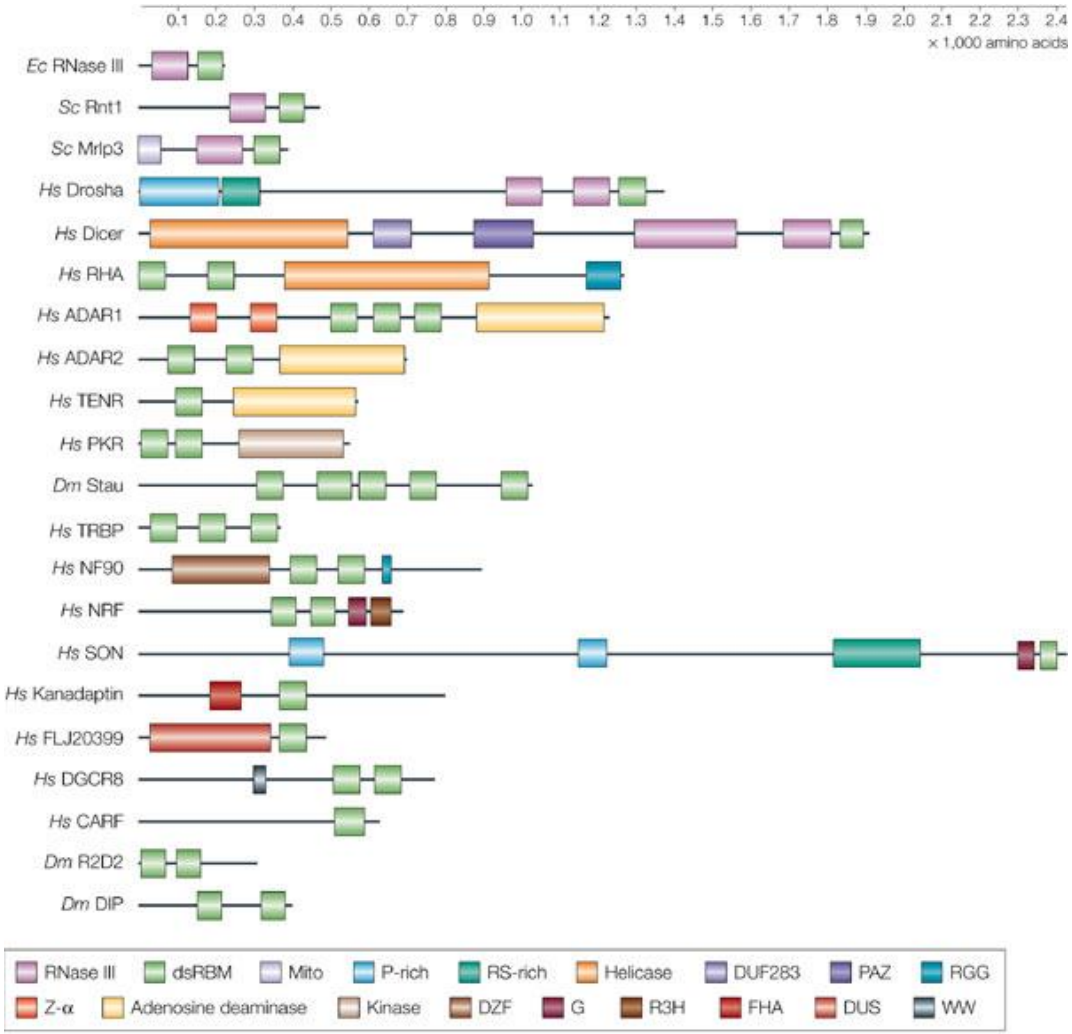


Domains and Motif

- **Domain** is a conserved part of a given protein's tertiary fold and function independently
- **Motif** peptide sequence which also appears in a variety of other molecules.



Domains and motifs



Nature Reviews | Molecular Cell Biology

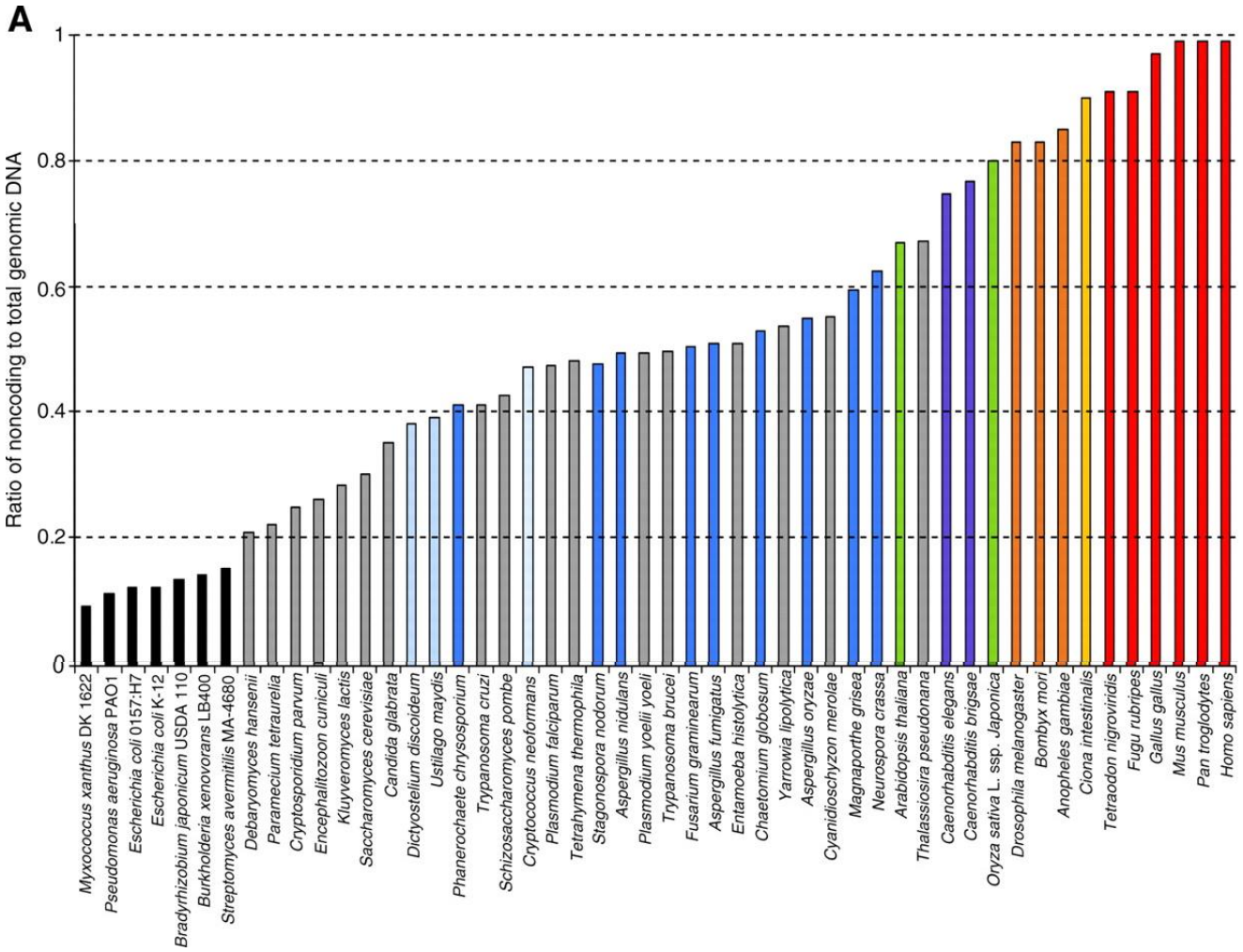
- Enzymatic function
- Interaction with other proteins
- Localization

Genome

- A genome is an organism's complete set of DNA (including its genes).

	Genome size	Num. of genes
E. coli	$.05*10^8$	4,200
Yeast	$.15*10^8$	6,000
Worm	$1*10^8$	18,400
Fly	$1.8*10^8$	13,600
Human	$30*10^8$	25,000
Plant	$1.3*10^8$	25,000

Non-coding fraction of the genome



Multicellular organisms

- (Almost) every cell has the same genome and the same protein coding capacity
- They don't all make the same proteins

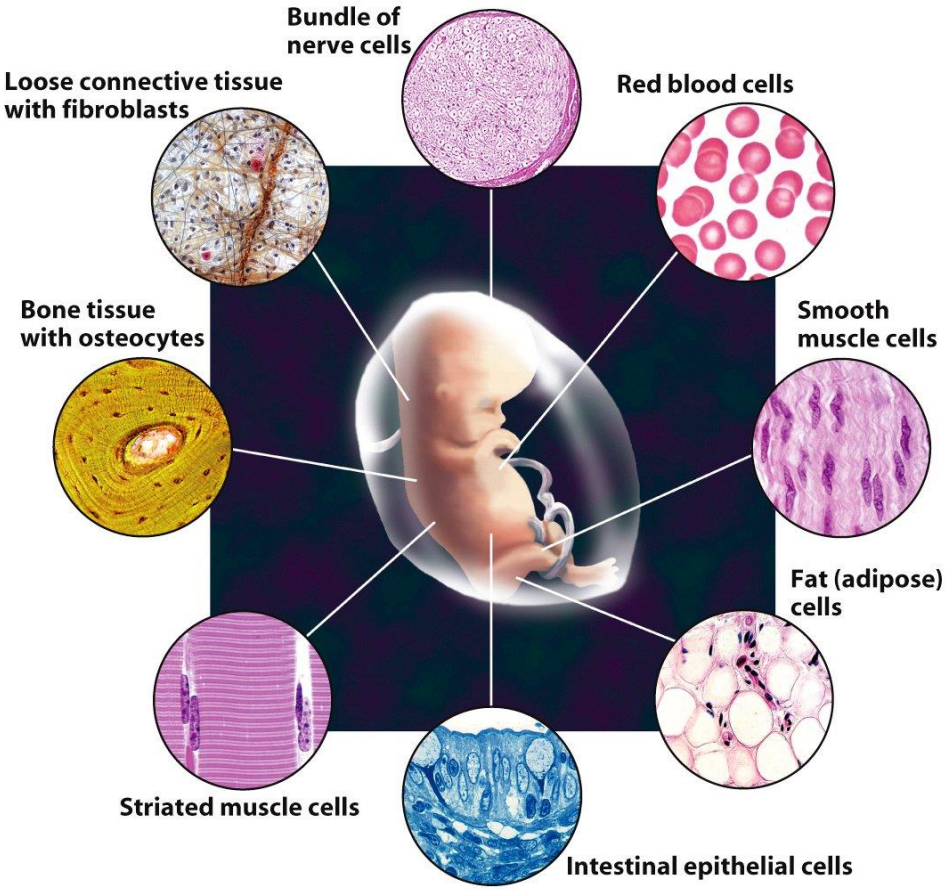


Figure 1-17 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Gene transcription is highly regulated

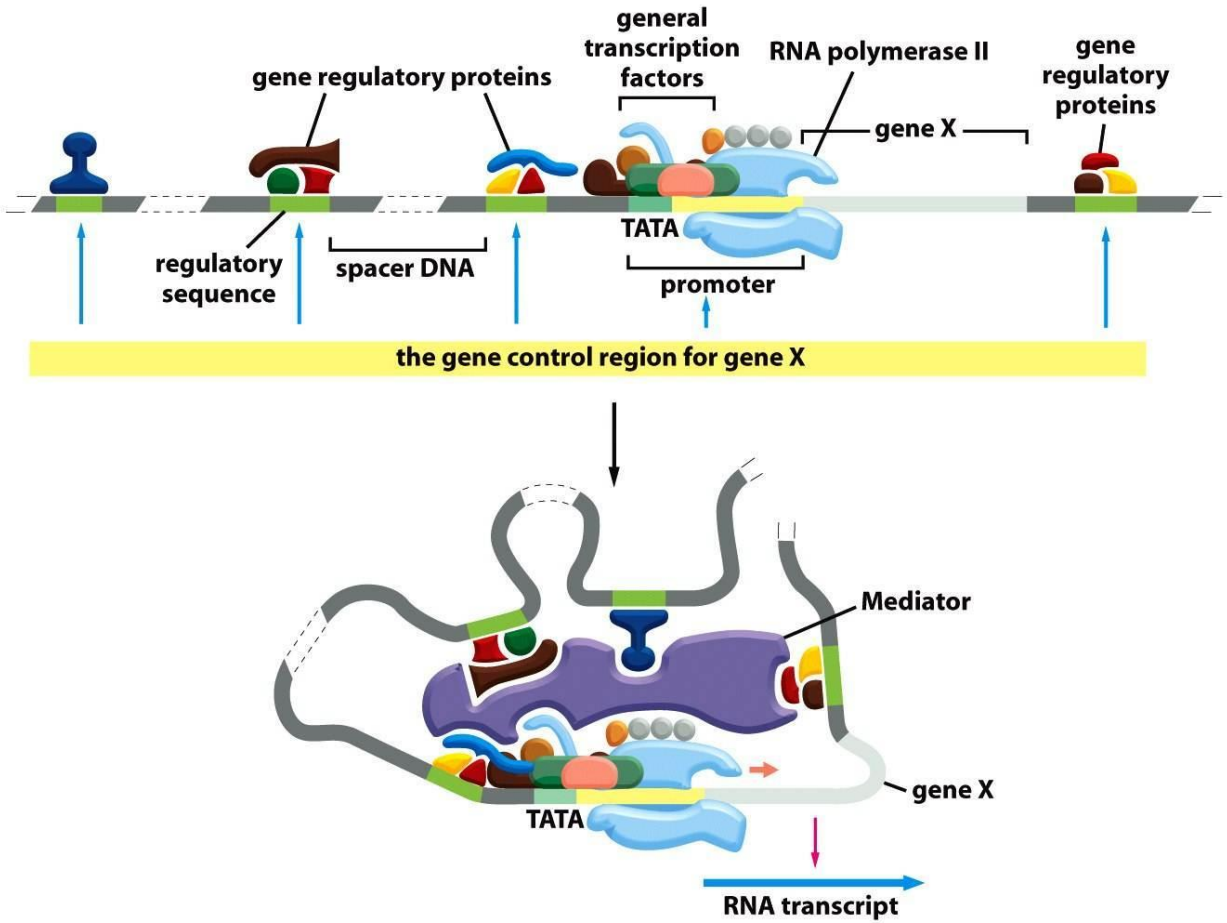


Figure 7-44 Molecular Biology of the Cell 5/e (© Garland Science 2008)

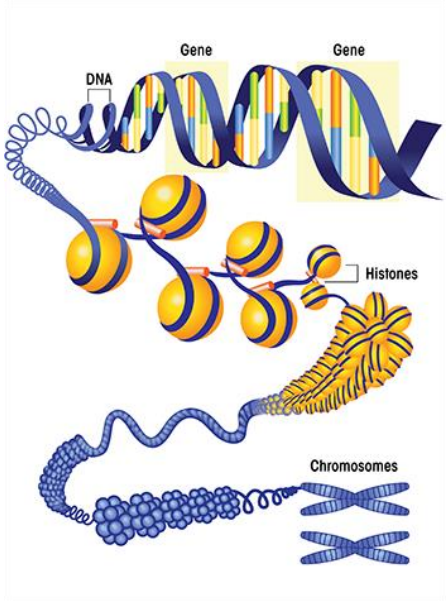
Many regulatory proteins bind in a sequence specific manner

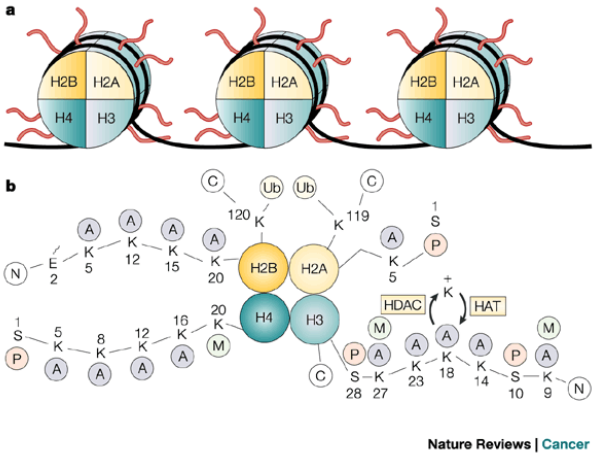
Motif	TF	Motif	TF
SPT10	Spt10	E2F	E2F
TBP	TBP	GC box	Sp1
NEG	NI	HEX	NI
AACCCT	NI	HiNF-D	HiNF-D
Oct-1	POU2F1	IRF-7	IRF-7
CCAAT box	NF-Y	IRF-1	IRF-1
a-CP1	NF-Y	TATA box	TIIFD

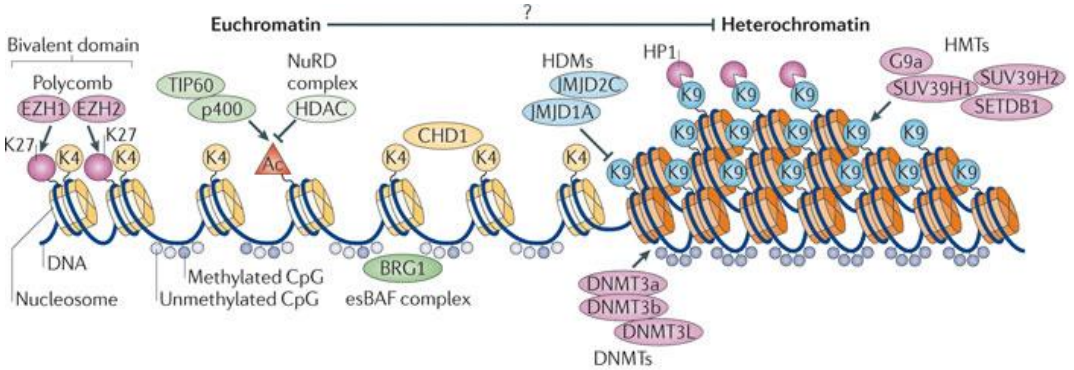
Motifs are easy to find
Challenging to predict functionality

Epigenetics

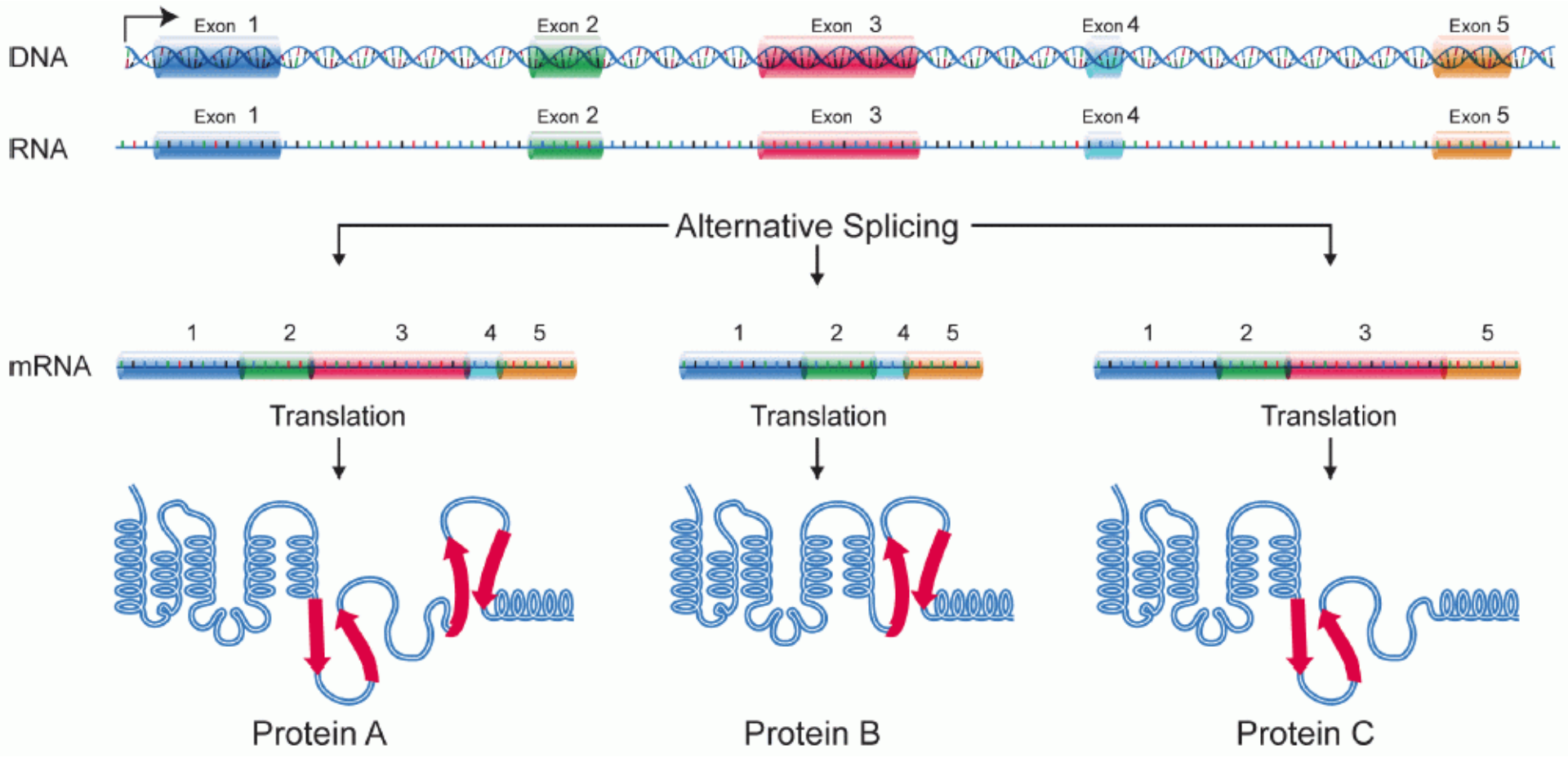
- Beyond/on-top-of genetic
  - Study of (stable/heritable) non-genetic differences in traits
- **Chromatin**: DNA and all the proteins bound it
- Chromatin structure dictates the transcriptional potential of a cell
- The structure is heritable: across cell divisions and sometimes trans-generationally







RNA splicing

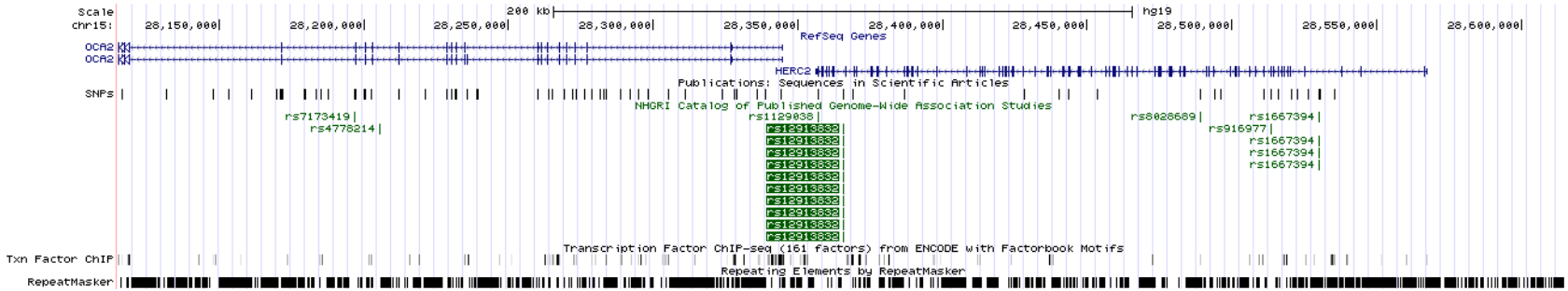


What is a mammalian genome the genome

- Coding segments – code for protein, only about 2%
- The rest:
  - Psuedogenes-genes that have lost their ability to code for protein
  - introns
  - non coding genes
  - Regulatory elements
  - Not mutually exclusive!
- 10-15% is constrained based on conservation

Genome view: OCA2

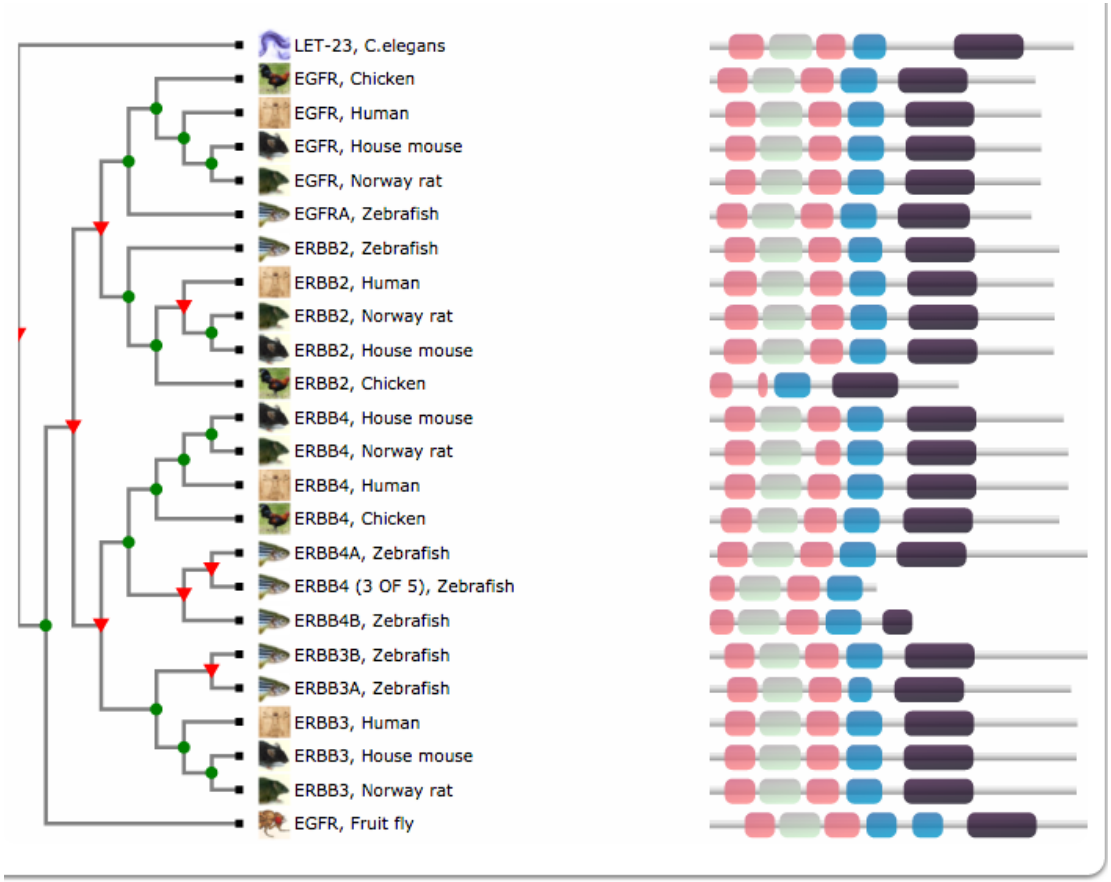
- oculocutaneous albinism II:
  - Involved in melanin production
  - Mutations in this gene cause changes in coloration in both skin and iris
- Eye color can be affected independently of skin color via a regulatory region in a neighboring gene



UCSC genome browser

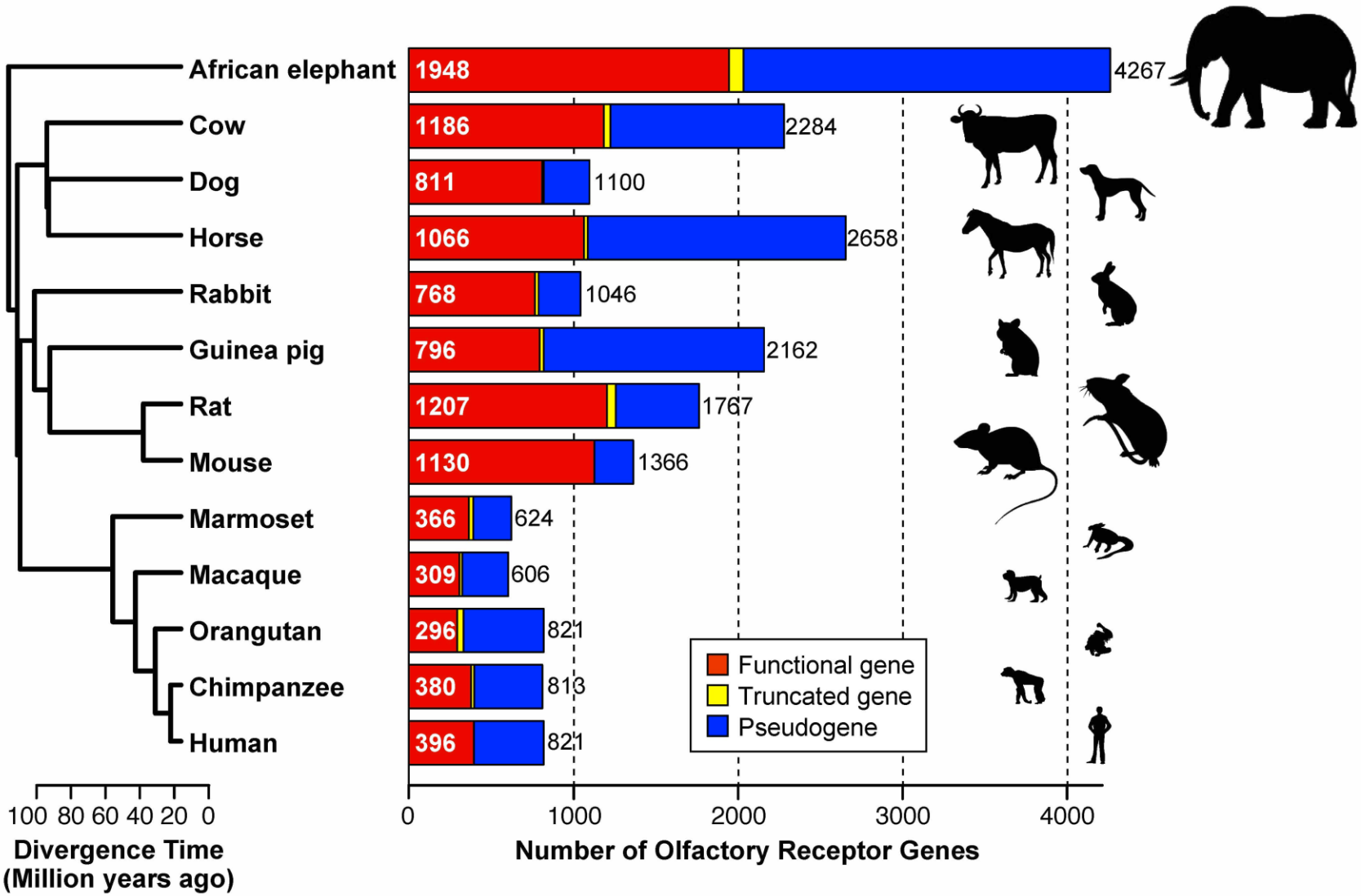
Protein families

Homolog genes related to a second gene by descent from a common ancestral DNA sequence.
Ortholog genes in different species that evolved from a common ancestral gene by speciation.
Paralog Paralogs are genes related by duplication within a genome.



Treefam database

Some families are very large



Genomic data

- Sequence
  - Identify genes/proteins
  - Assign molecular function based on similairy to known proteins, domains, and motifs
  - Examples of functional classes assigned from sequence
    - Digestive enzyme
    - G-coupled protein receptor
    - Kinase
  - Non-coding regions –identify potential TF binding sites
- Functional genomics data
  - Gene Expression-assays genes only
    - quantify which mRNAs are made when/where
    - example: which genes are transcribed only in neurons?
  - Chromatin assays-probe non-genic regions for function
    - How is gene expression regulated biochemically
    - Assay for chromatin structure: open/closed?
    - Find where in the genome a regulatory protein binds

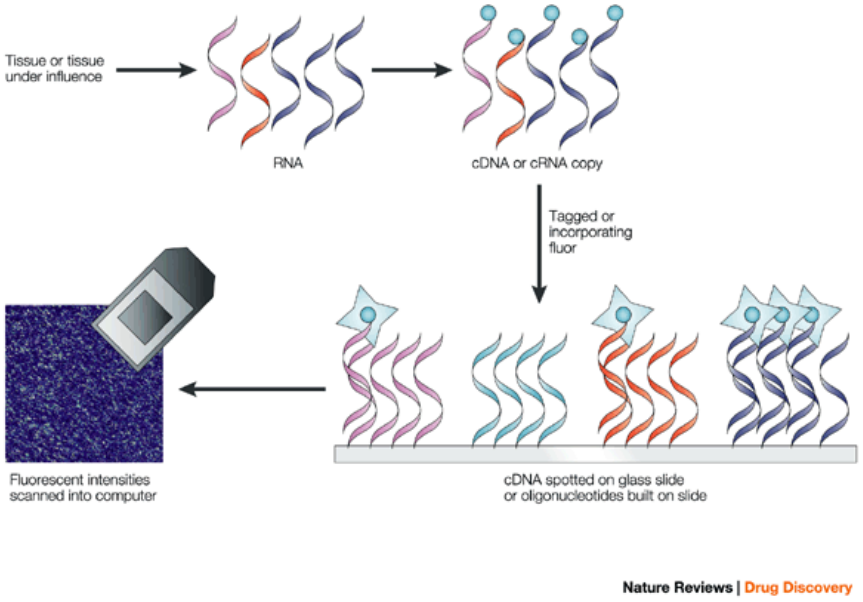
Gene Expression

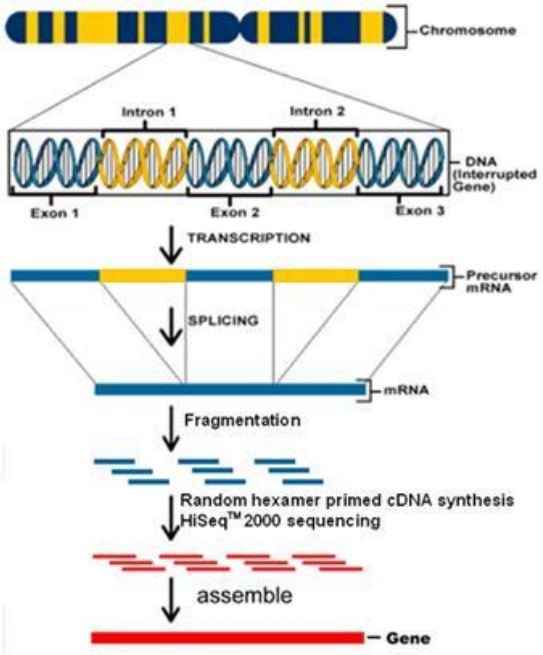
Hybridization and Scanning—Microarrays

- Quantity of mRNA is measured by fluorescence
- Requires complementary probes so the set of genes must be known

RNAseq using next generation sequencing methods

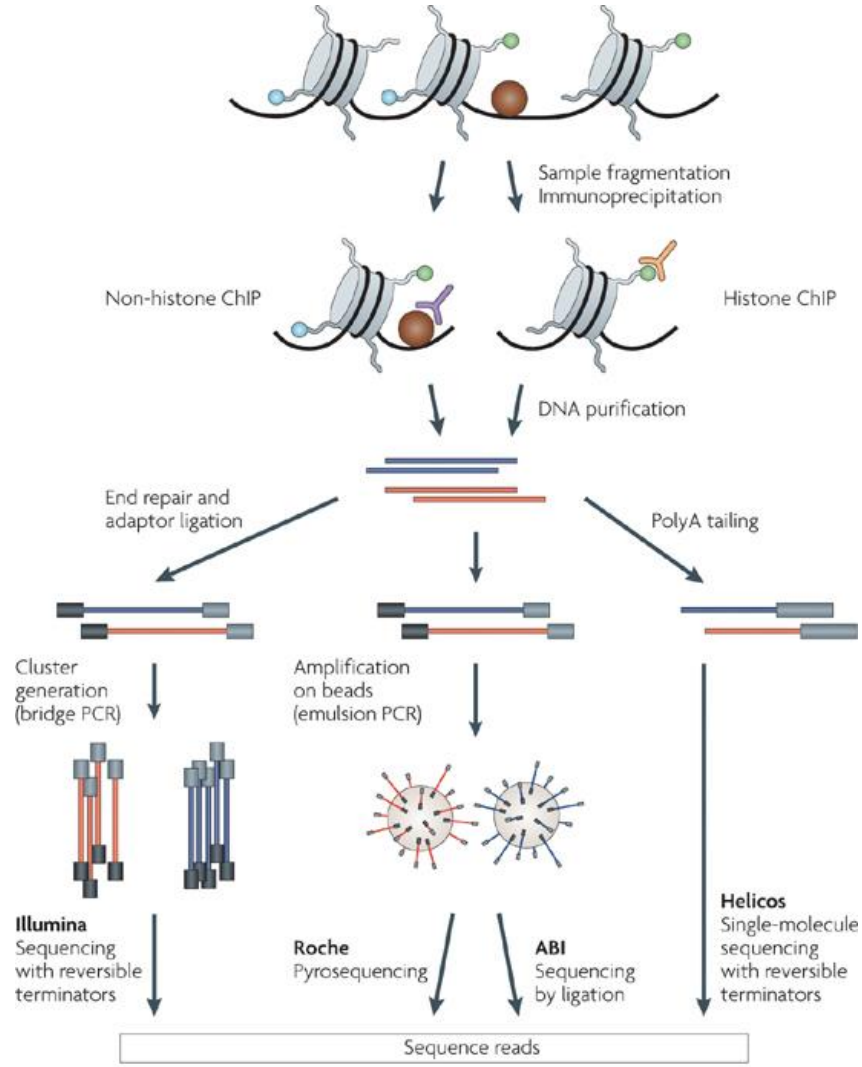
- Quantify by counting reads
- Don't have to know what you are looking for
- Generates much more data





ChIPseq

- **Ch**romatin **I**mmuno **P**recipitation followed by **seq**uencing
- Cross link DNA and proteins bound to it via a covalent bond
- Select the complexes of interest with an antibody
  - Transcription factors
  - Specific histone modification
- Sequence the bound DNA
- Bound regions are found across the genome—maybe be very far from the nearest gene



Nature Reviews | Genetics