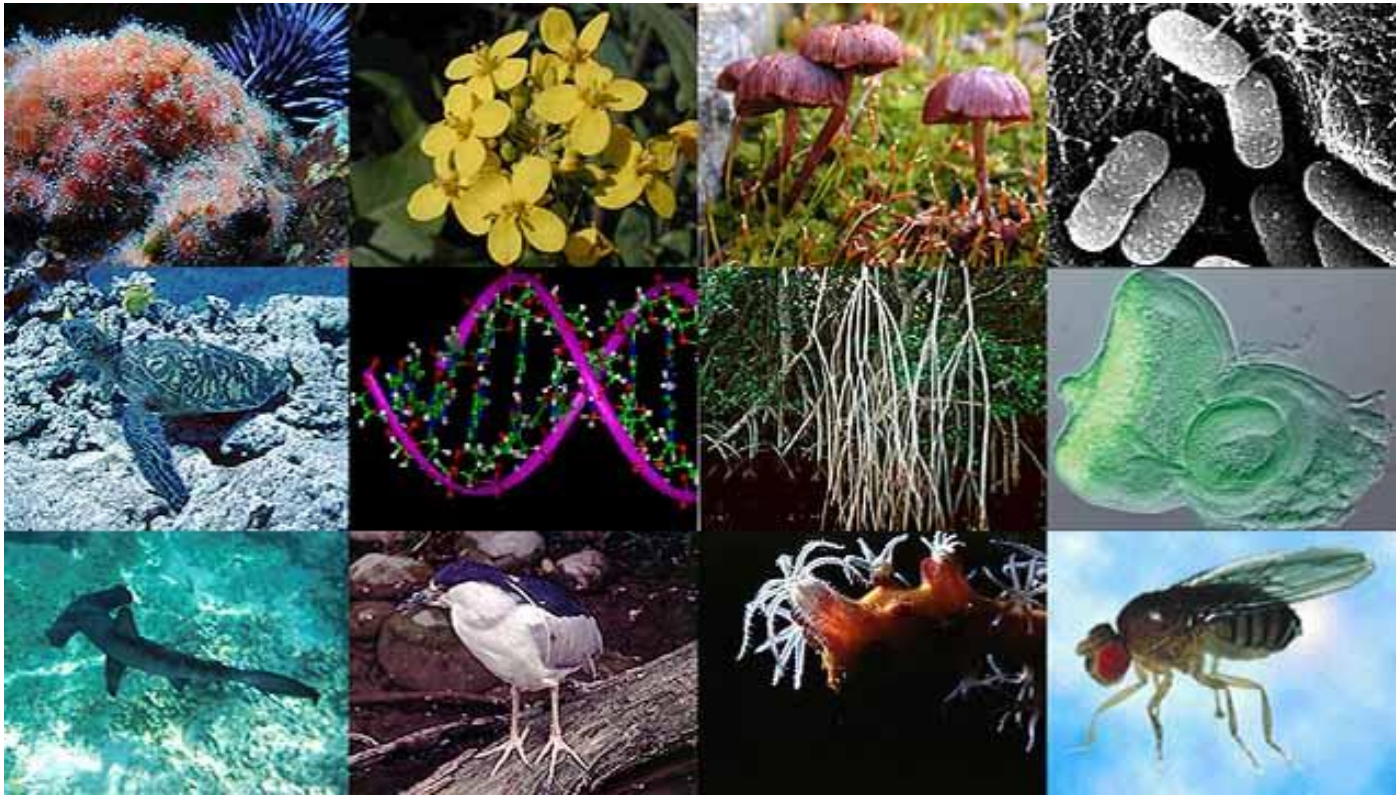
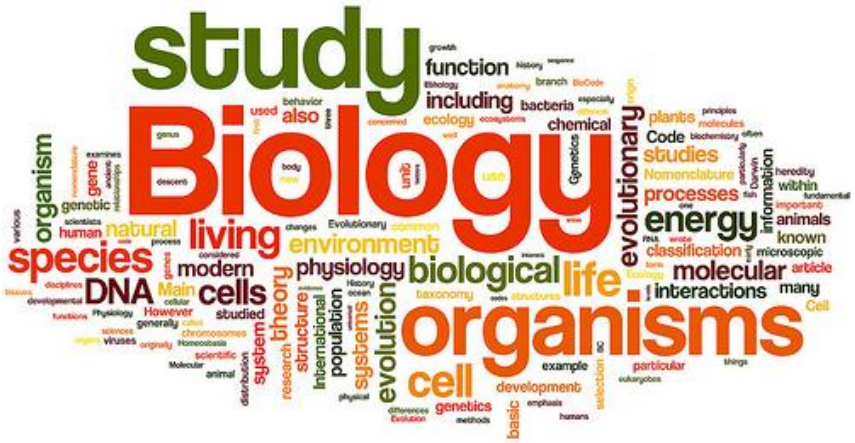
STAT540 Biology Introduction

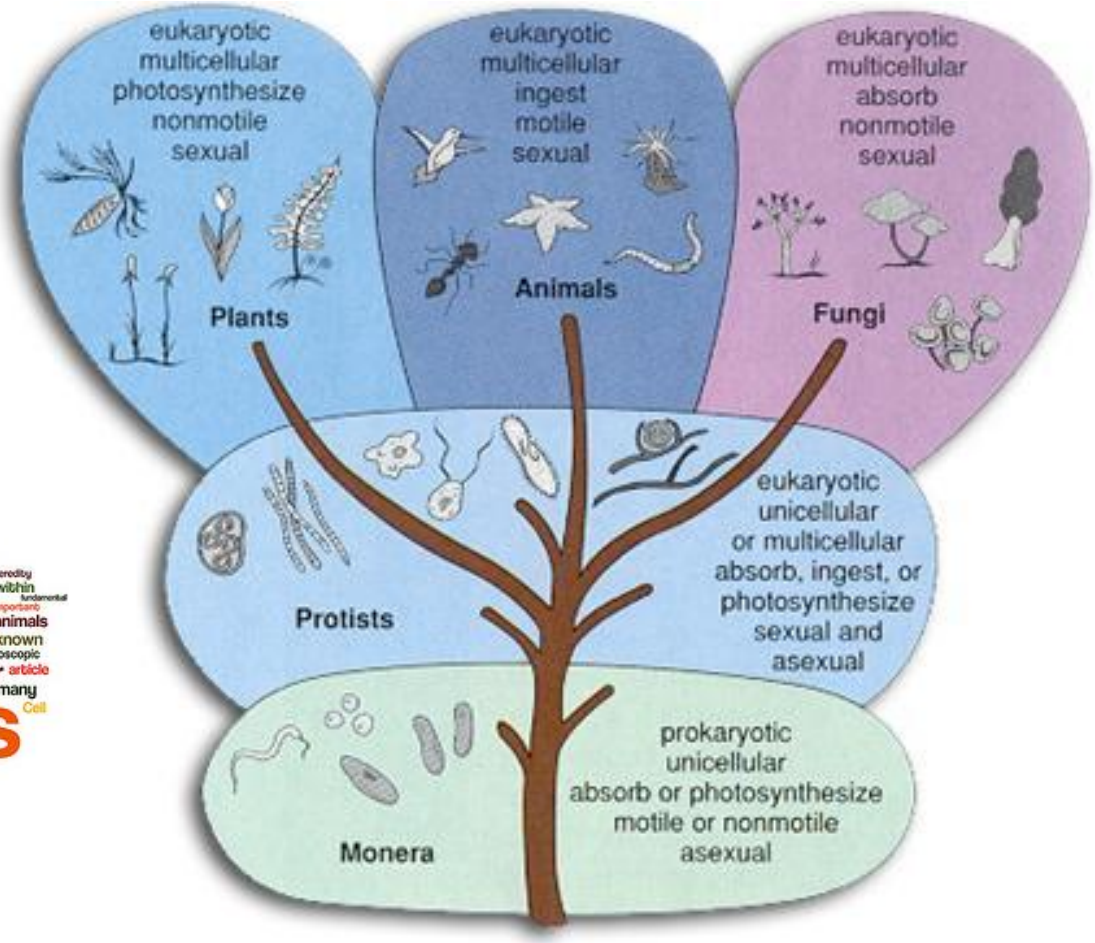
Alice Zhu, Evan Durno



http://www.crystalinks.com/biologycollage.jpg

Biology studies organisms and their interaction with the environment



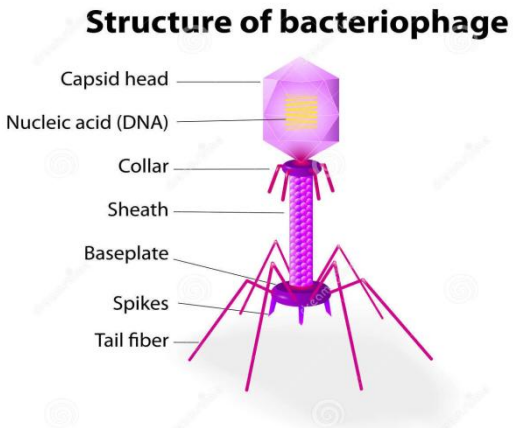


Cells, signal network, DNA/RNA, Proteins

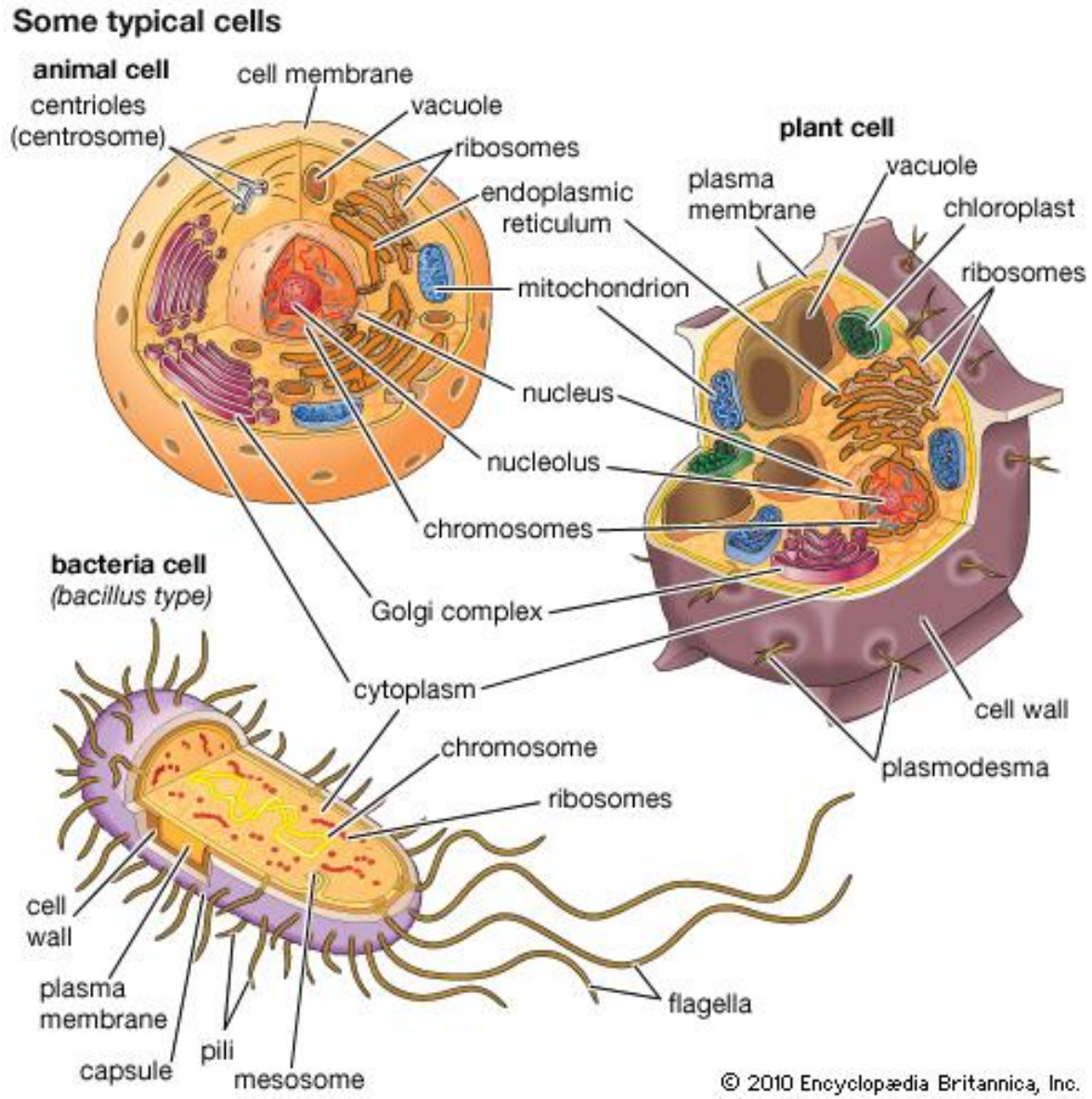
https://farm8.staticflickr.com/7211/6961733008_4e562b21a0.jpg

http://images.tutorvista.com/content/modern-classification/five-kingdom-classification.jpeg

Characteristic compartment layout of cells

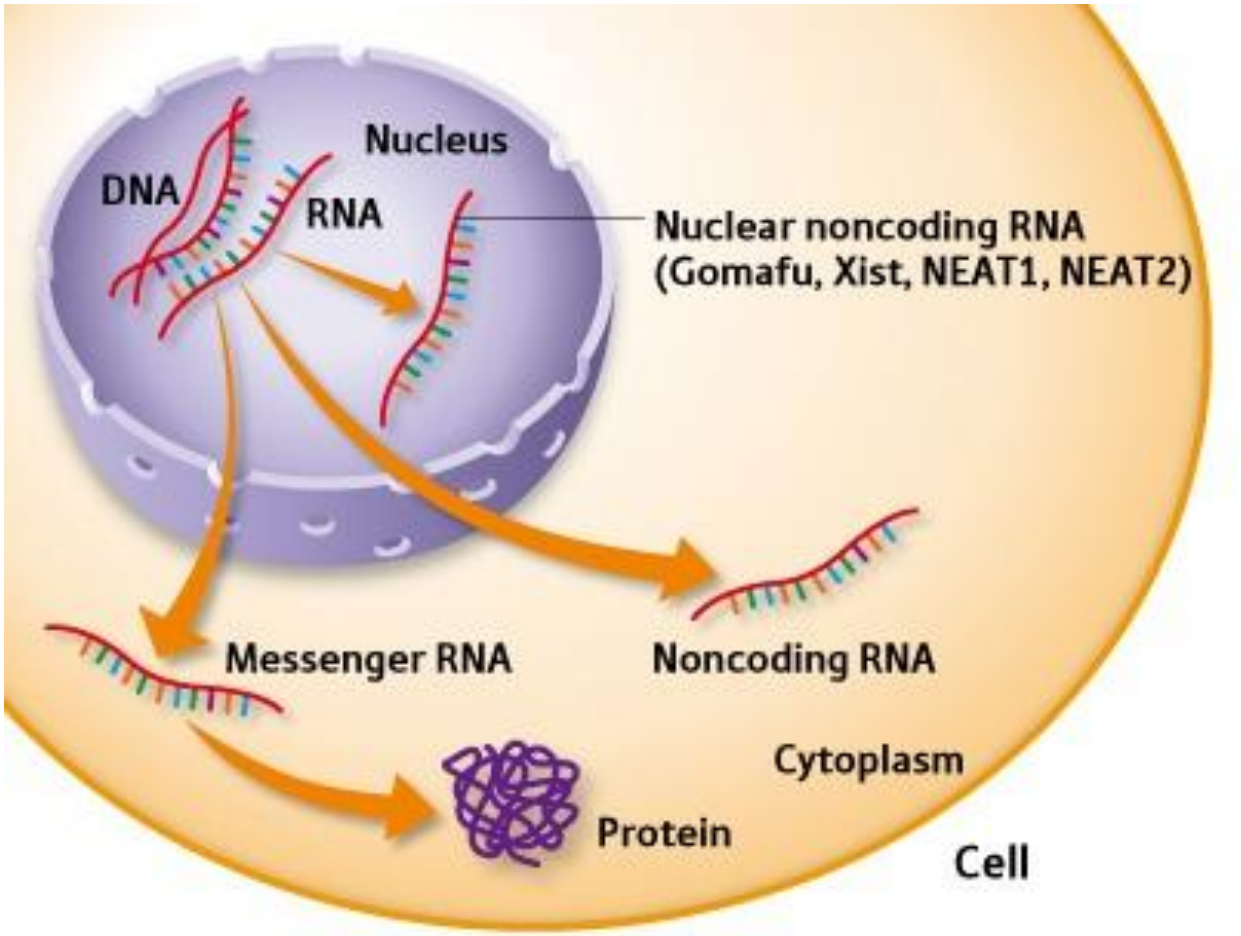


Dreamstime.com



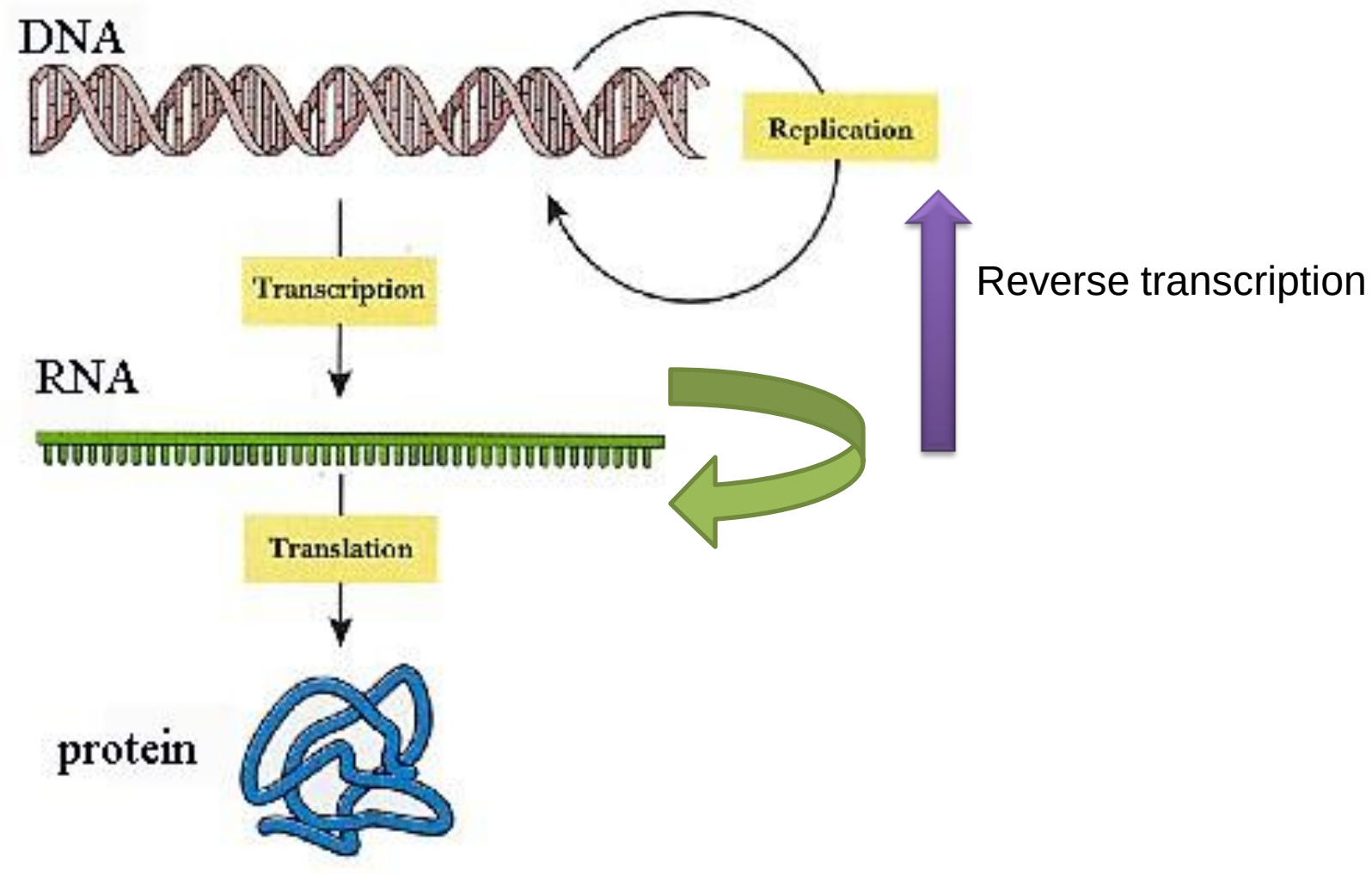
http://media-3.web.britannica.com/eb-media/85/78585-004-BC0F7EE4.jpg

The **Central Dogma**: Flow of heritable information



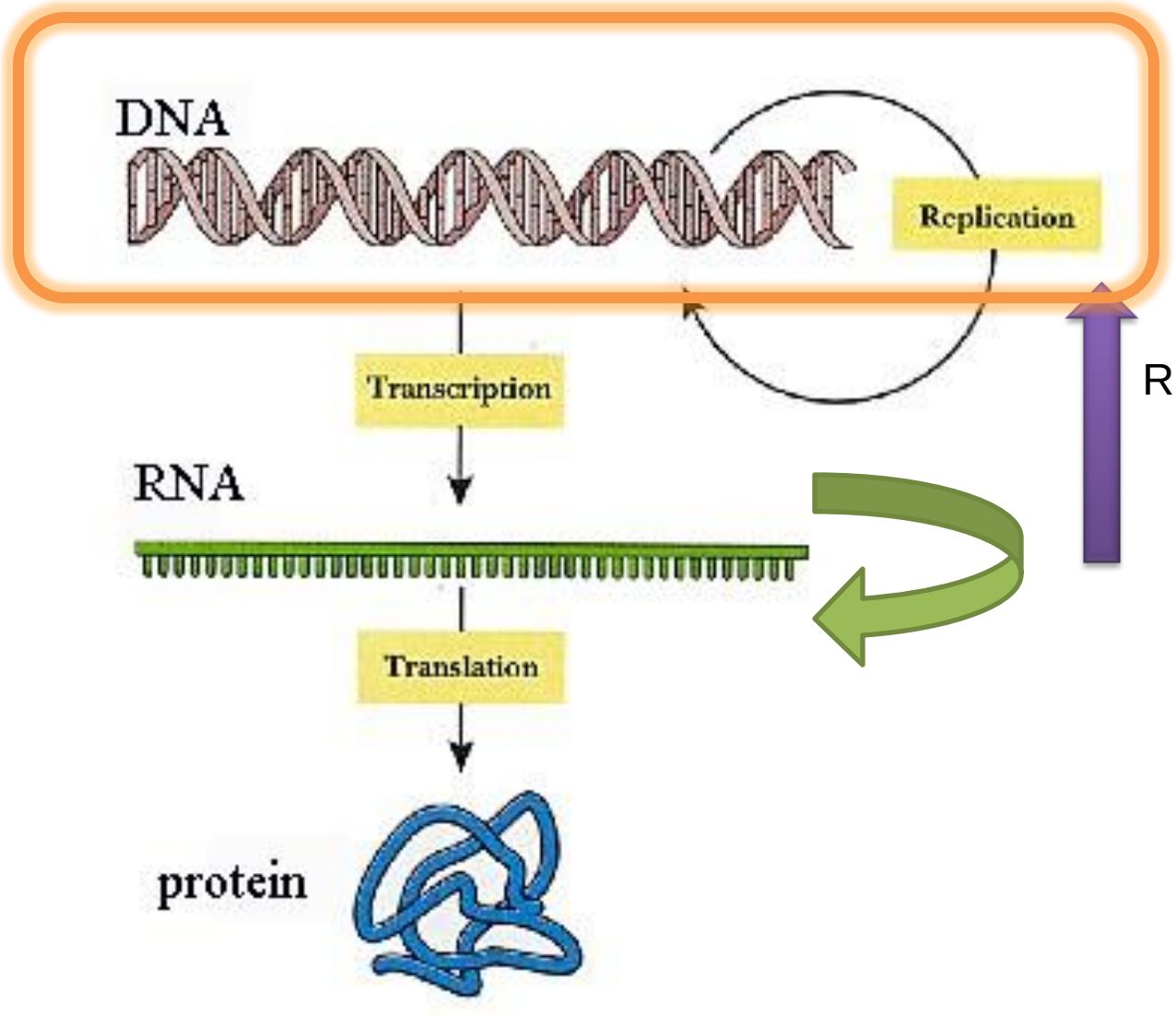
(left) http://www.rikenresearch.riken.jp/eng/frontline/5920

The **Central Dogma**: Flow of heritable information



http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

The **Central Dogma**: Flow of heritable information



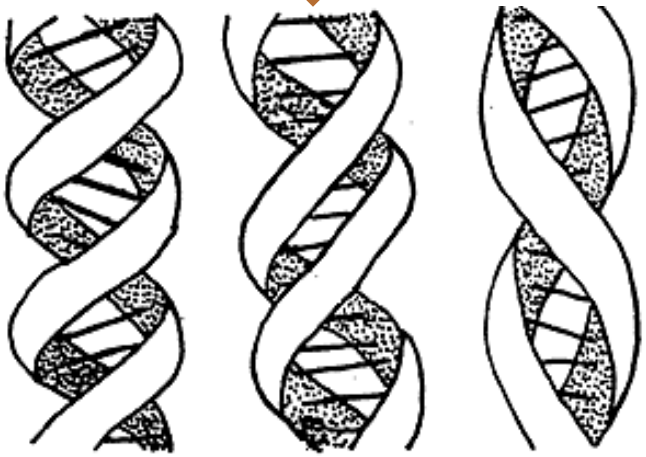
DNA

Reverse transcription

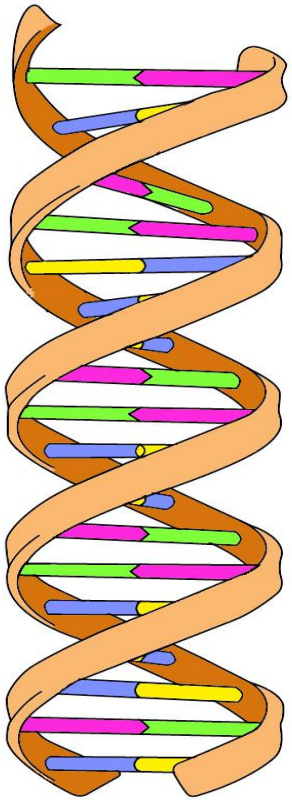
http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

DNA has 3 forms of double helix, and is composed of nucleotides

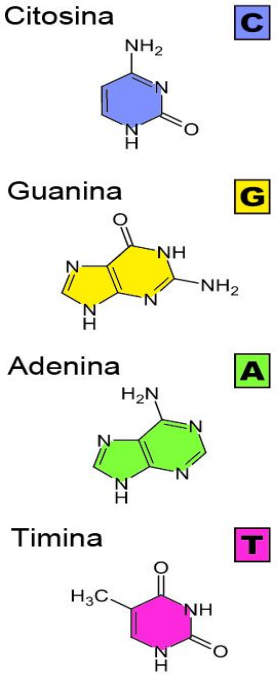
Neutral PH,
Physiological [salt]



Form	A	B	Z
helical twist	right	right	left
base pairs per turn	11	10	12
occurrence	RNA, DNA	DNA	DNA
rise/bp	0.24nm	0.34nm	0.37nm



DNA
Acido Desossiribonucleico

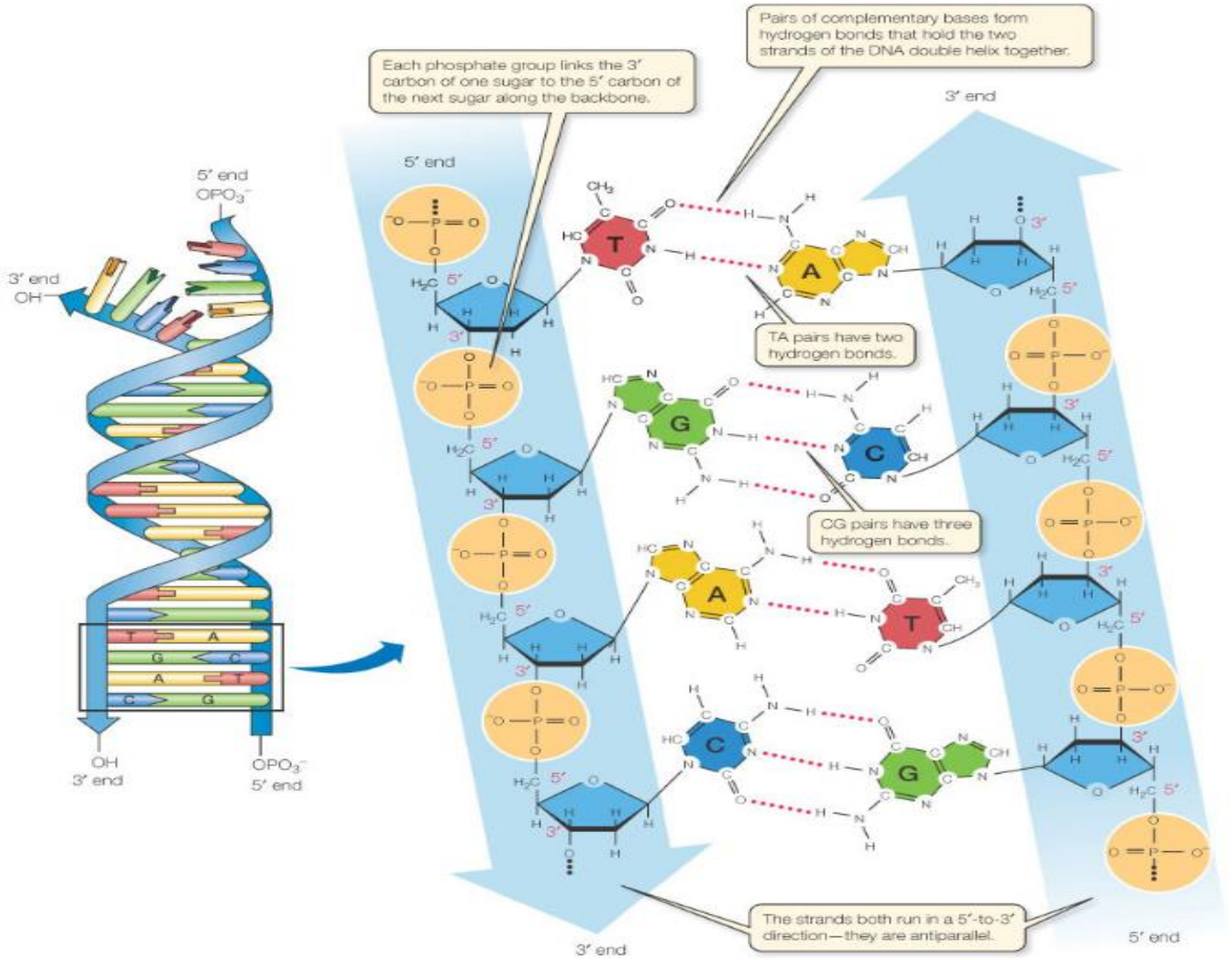


Basi azotate

http://chemistry.gravitywaves.com/CHEMXL153/GeneticsImages/DNATwists.gif

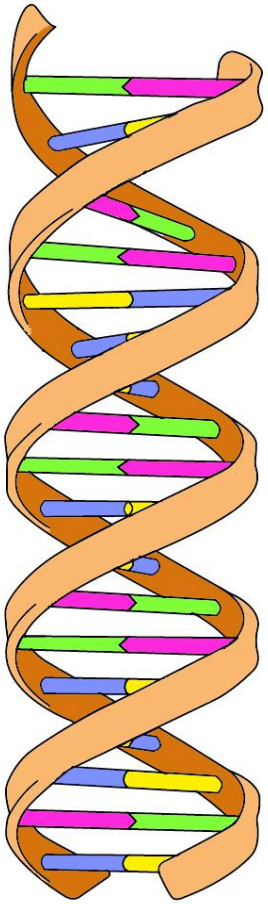
http://superagatoide.altervista.org/images/DNA.jpg

DNA: Double Helix

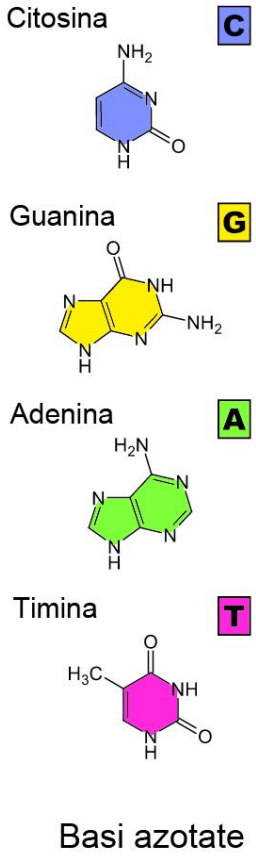


http://www.nature.com/scitable/content/24298/sadava_11_9_large_2.jpg

DNA encodes genes

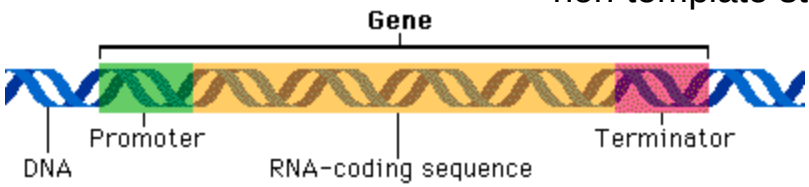


DNA
Acido Desossiribonucleico

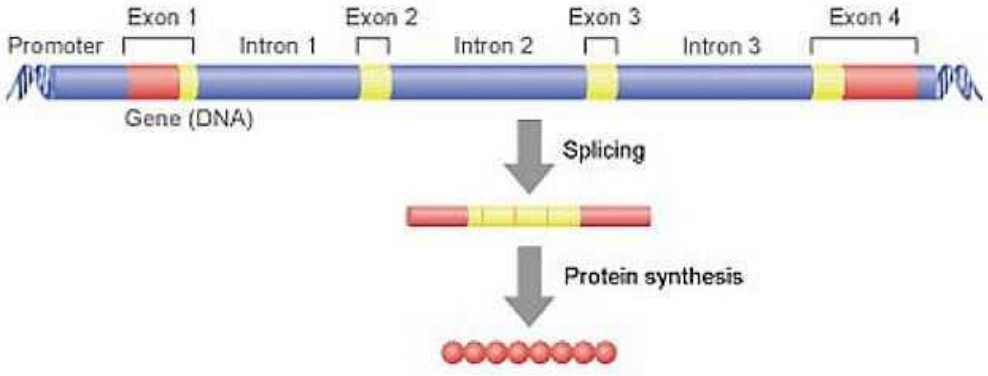


Basi azotate

Template strand and non-template strand



Gene = single and contiguous unit of inheritance, encoding a (coding) protein or (non-coding) RNA product



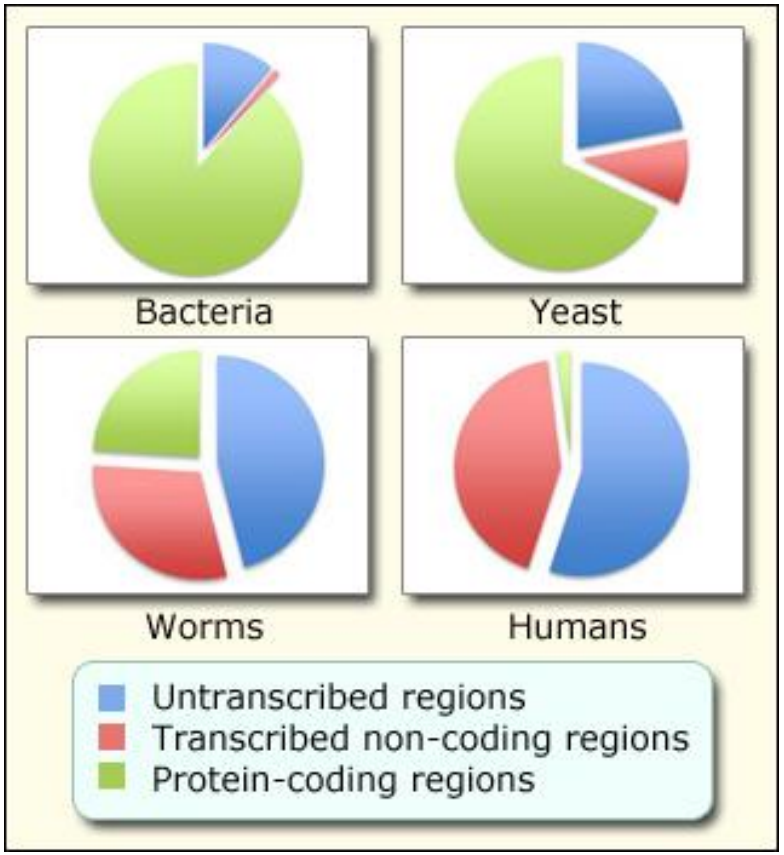
http://superagatoide.altervista.org/images/DNA.jpg

http://www.phschool.com/science/biology_place/biocoach/images/transcription/gene.gif

Coding region is only a fraction of the genome

Number of genes in species:

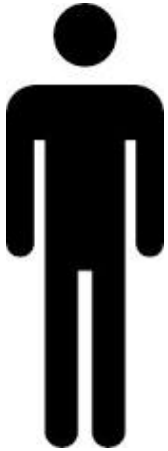
- Human: 20,000 ~ 30, 000
- Rice: 28,000
- Fly: 14,000
- Yeast: 6,000
- *E.coli*: 4,000



The noncoding DNA is not junk region

Human and Fugu fish share roughly the same genes and regulatory sequences, but the Fugu fish has a more condensed form with little junk region.

VS.

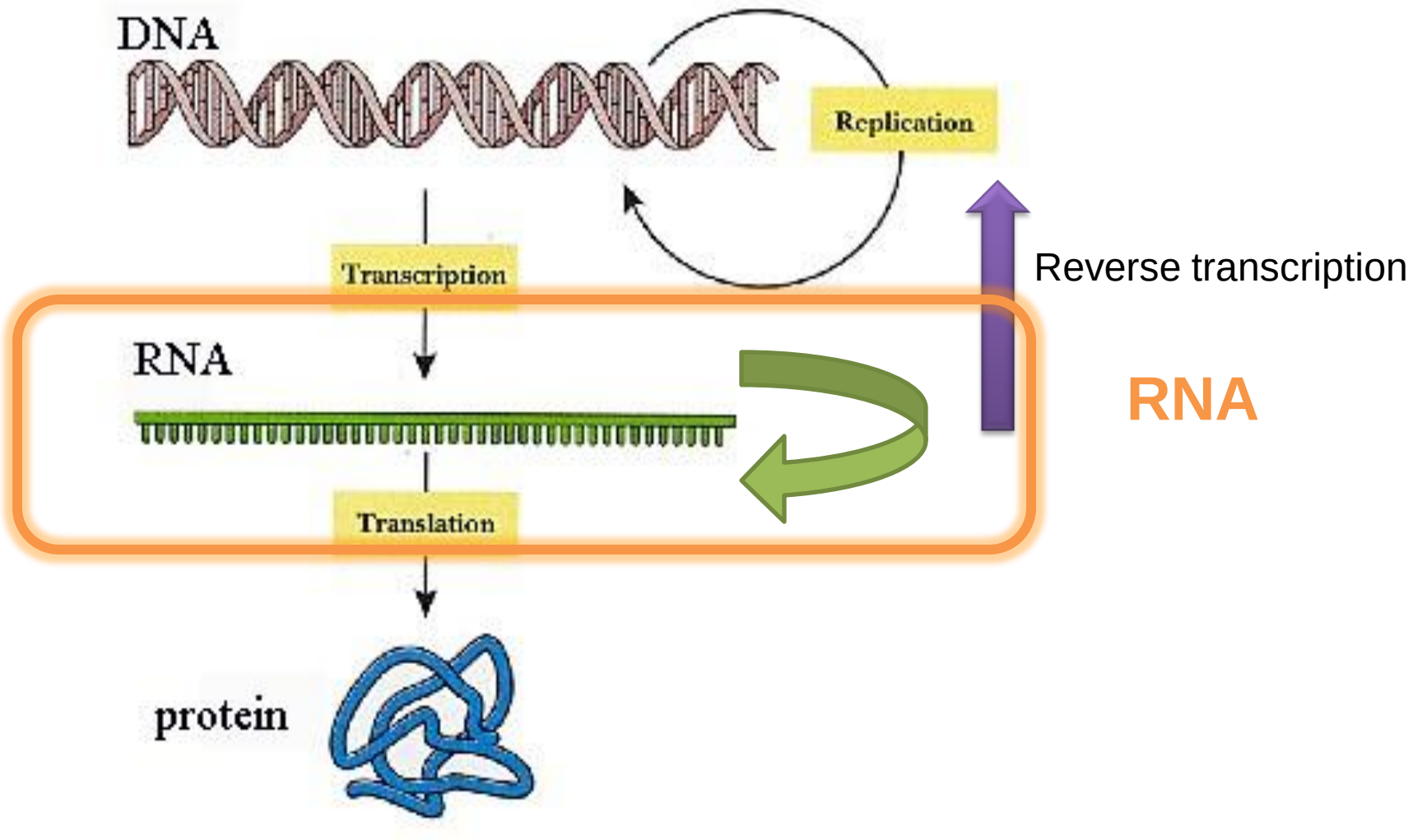


Fugu fish: 400 million base

Human: 3 billion base

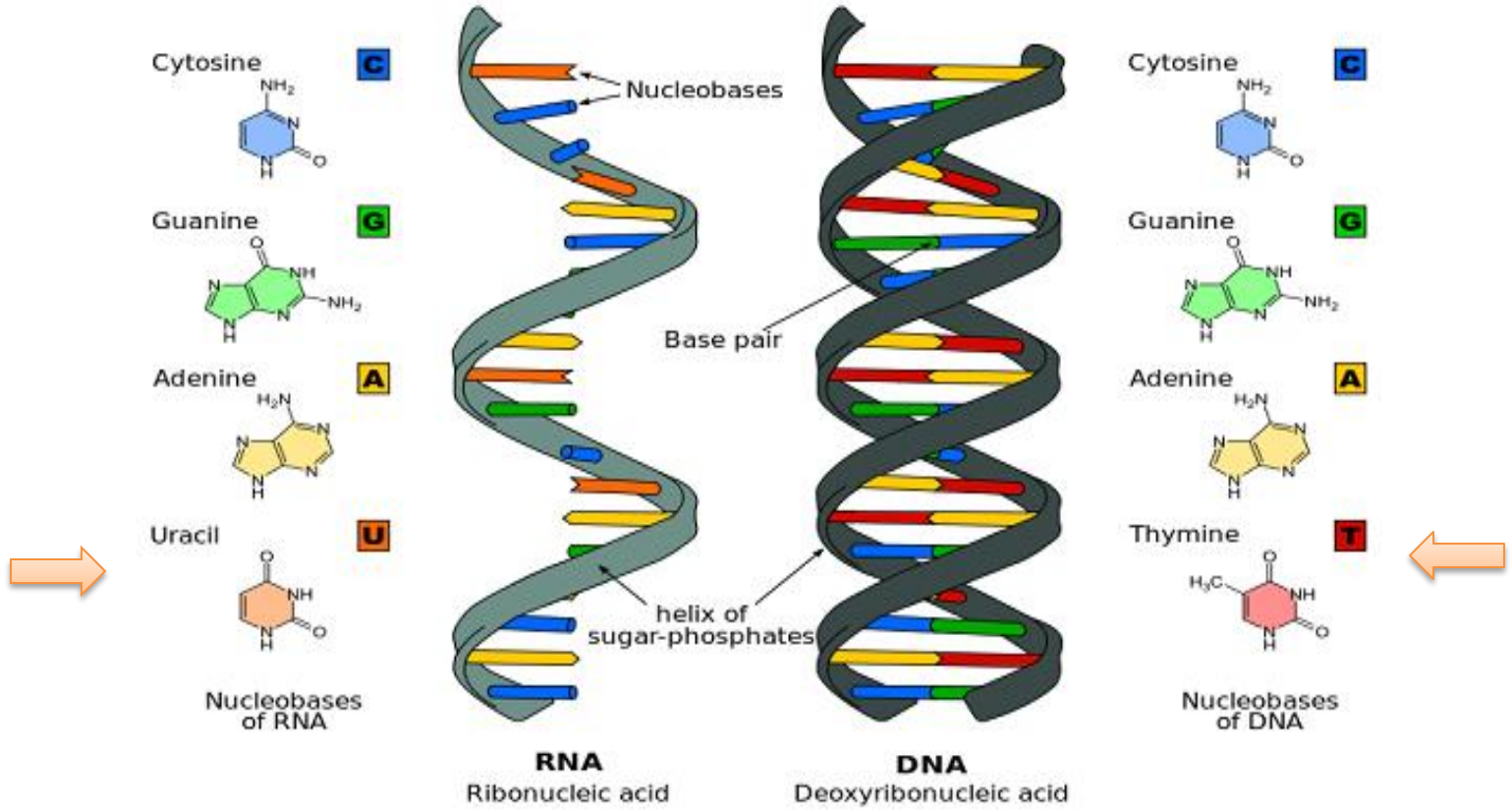
http://www.genomenewsnetwork.org/gnn_images/news_content/08_02/pufferfish/fugu.jpg

The **Central Dogma**: Flow of heritable information



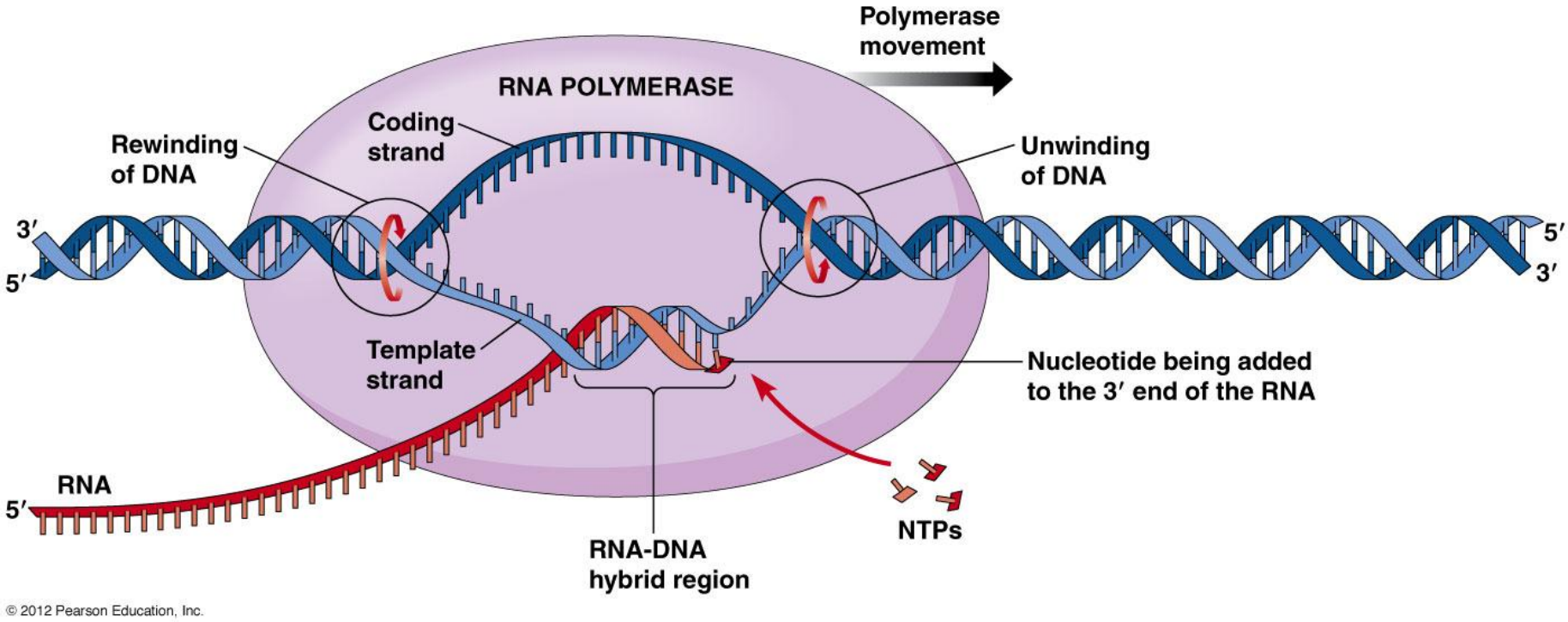
http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

RNA



http://education-portal.com/cimages/multimages/16/Difference_DNA_RNA-EN.svg.png

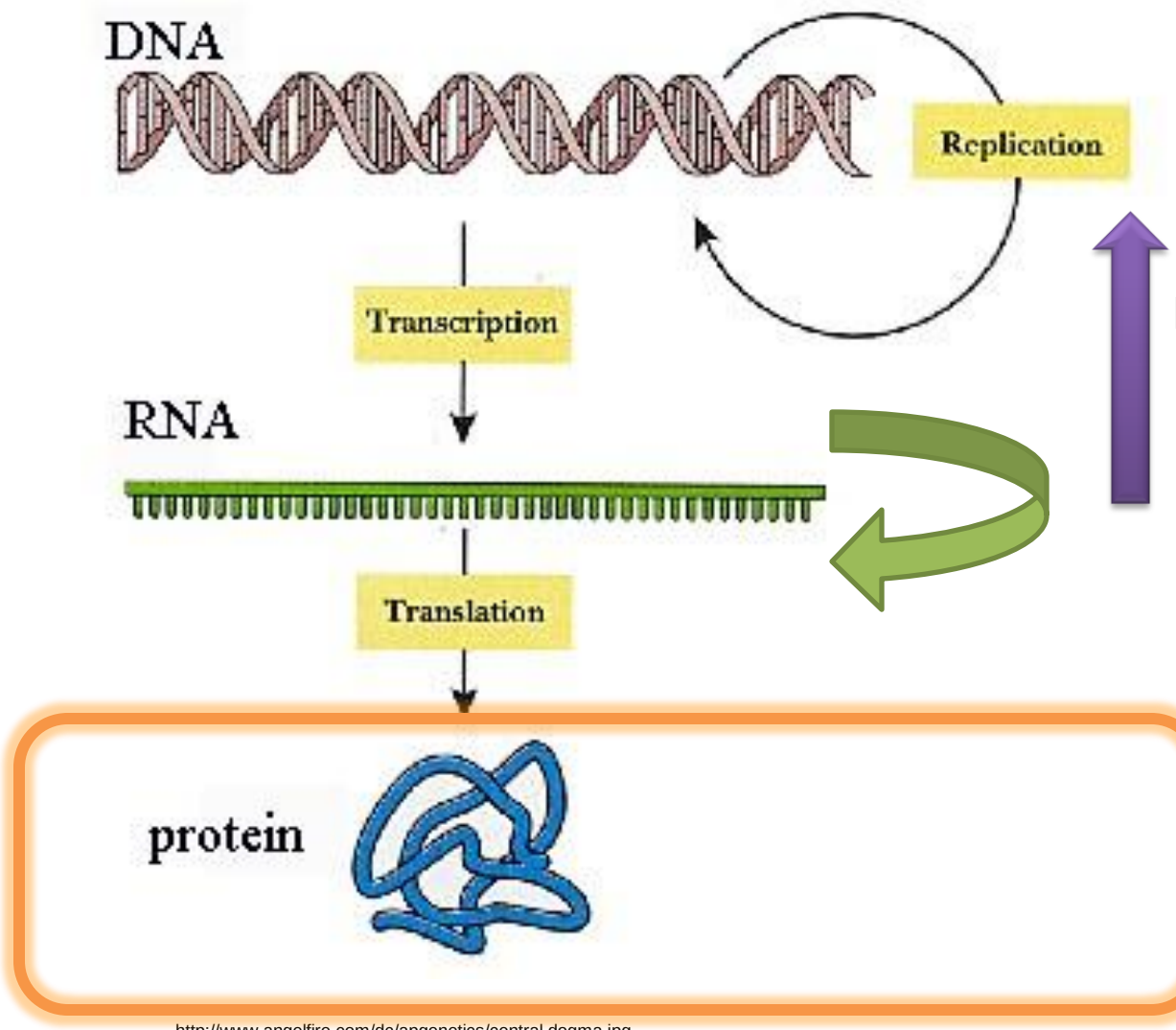
RNA is synthesized by transcription



Transcription proceeds from 5’ to 3’ on the RNA, *i.e.* from 3' to 5' on the **complementary** template DNA

http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-21/21_09.jpg

The **Central Dogma**: Flow of heritable information



Reverse transcription

Protein

http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

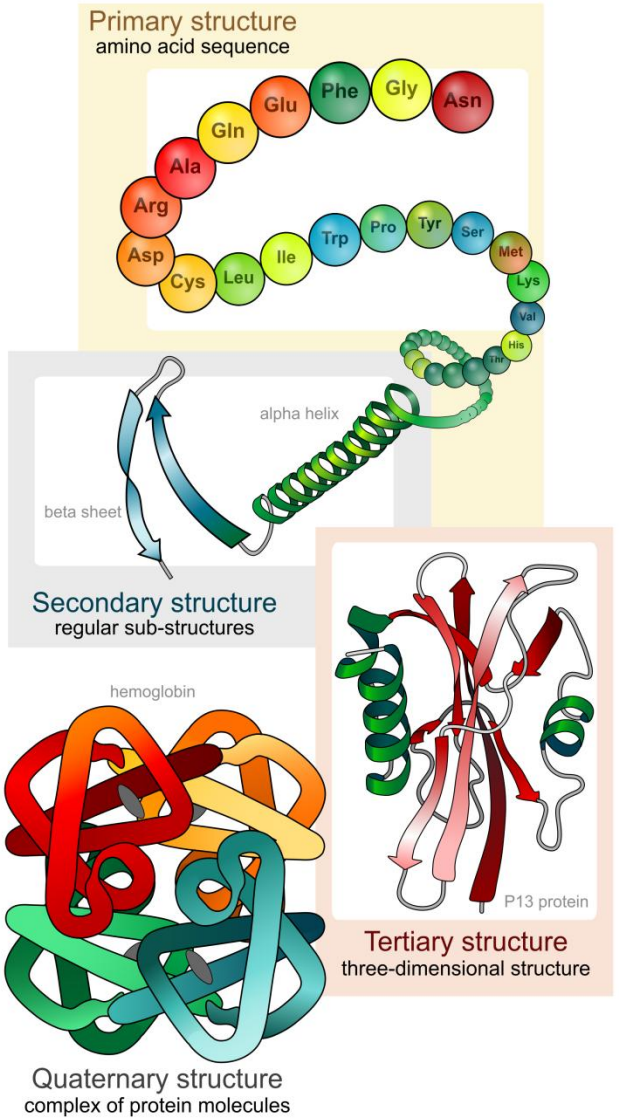
Protein Structure

Primary Protein structure
sequence of a chain of animo acids

Secondary Protein structure
hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern

Tertiary protein structure
three-dimensional folding pattern of a protein due to side chain interactions

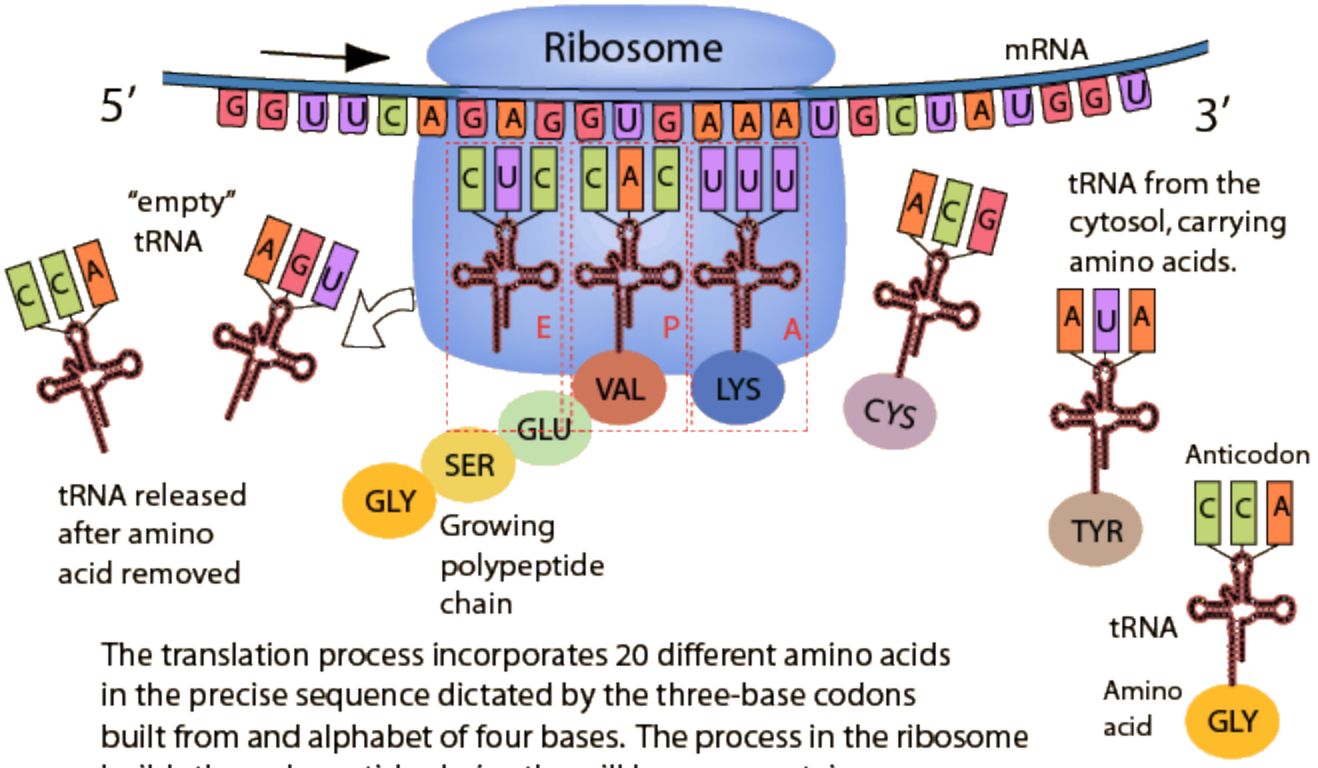
Quaternary protein structure
protein consisting of more than one amino acid chain



http://cnx.org/content/m44402/latest/Figure_03_04_09.jpg

http://upload.wikimedia.org/wikipedia/commons/thumb/c/c9/Main_protein_structure_levels_en.svg/2000px-Main_protein_structure_levels_en.svg.png

Protein is synthesized by translation

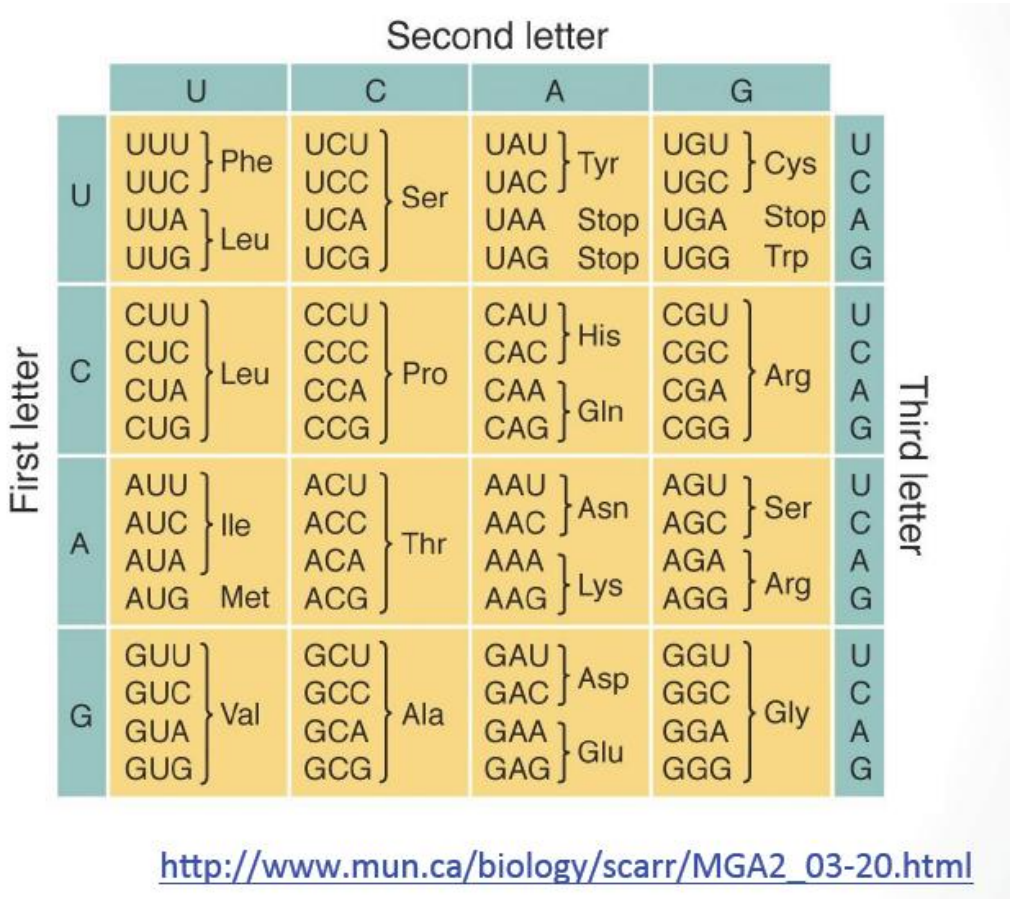


The translation process incorporates 20 different amino acids in the precise sequence dictated by the three-base codons built from and alphabet of four bases. The process in the ribosome builds the polypeptide chains tha will become proteins.

http://hyperphysics.phy-astr.gsu.edu/hbase/organic/imgorg/translation2.gif

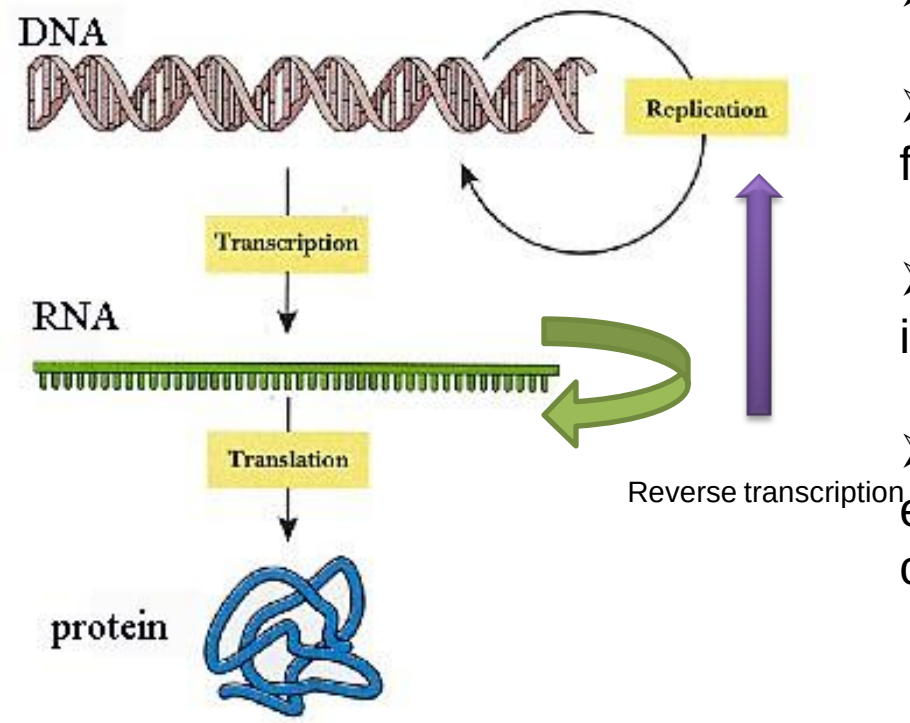
Translation is interpreted via Codon

Codon Degeneracy:
Many distinctive codons can redundantly map onto the same amino acid



The **Central Dogma**: Summary

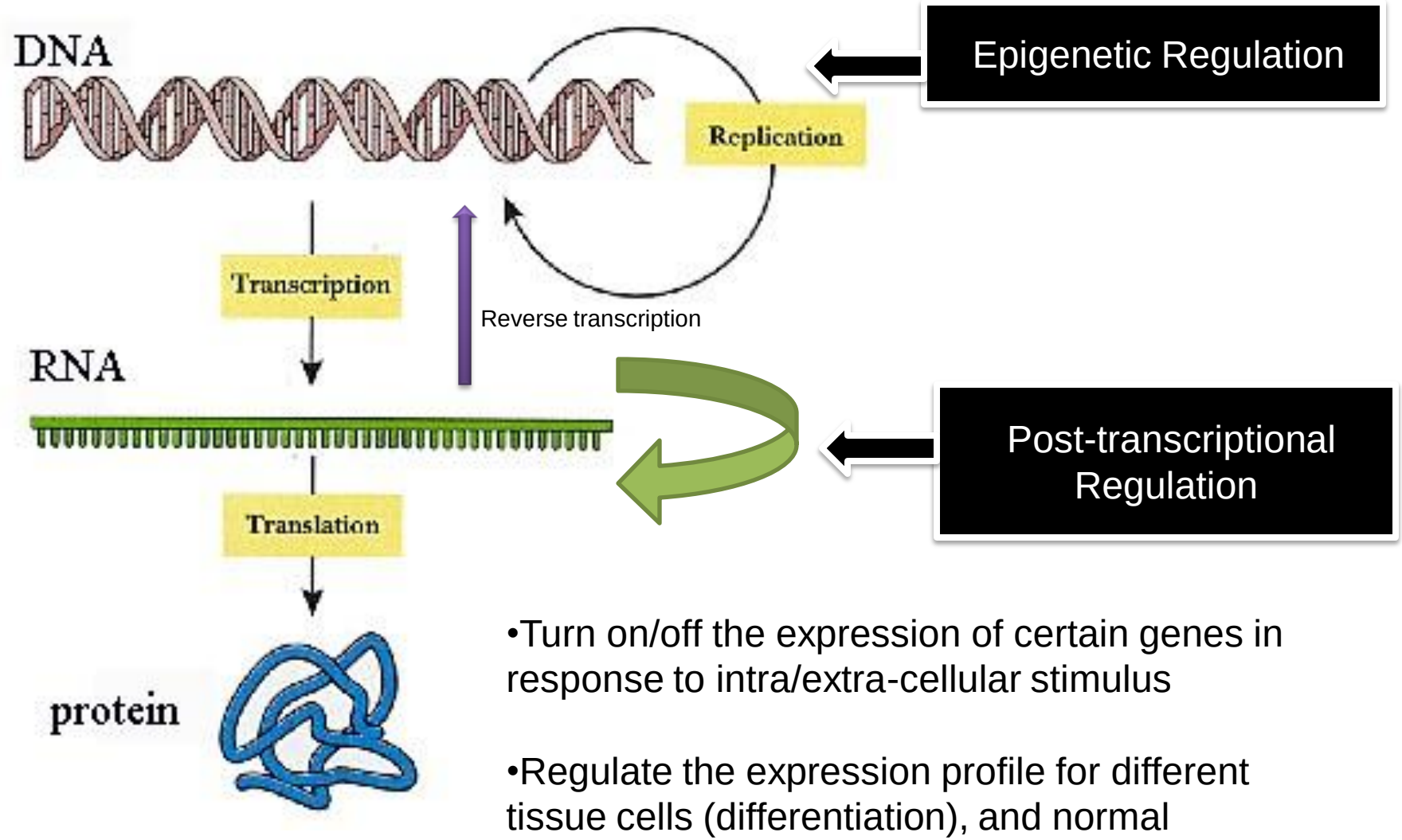
http://www.angelfire.com/dc/apgenetics/central.dogma.jpg



- The inherited information is stored in DNA
- RNA carries the inheritance information from nucleus to cytoplasm
- The inheritance information is expressed in proteins with RNA being the mediator
- The proteins come in versatile structures equipped with diverse functions to meet cellular demand

Example for the flow of heritable information, quantitatively:
in human,
DNA: 3 billion bp (haploid) → RNA transcripts: 62.1-74.7% of DNA transcribed (Djebali, Davis et al, 2012) → Proteins: about 20,000-25,000

Regulation of the gene expression

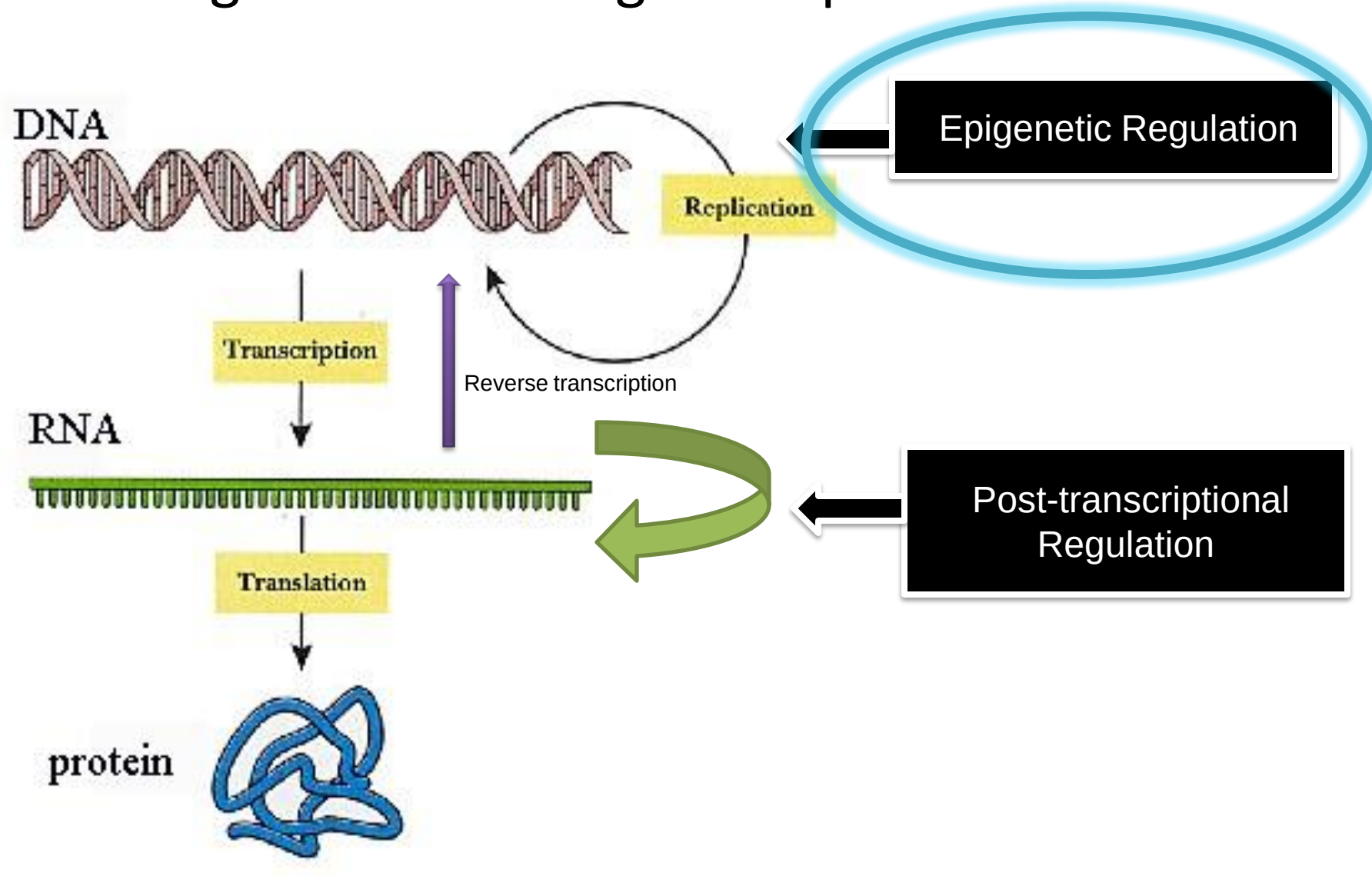


http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

- Turn on/off the expression of certain genes in response to intra/extra-cellular stimulus

- Regulate the expression profile for different tissue cells (differentiation), and normal development , e.g. imprinting

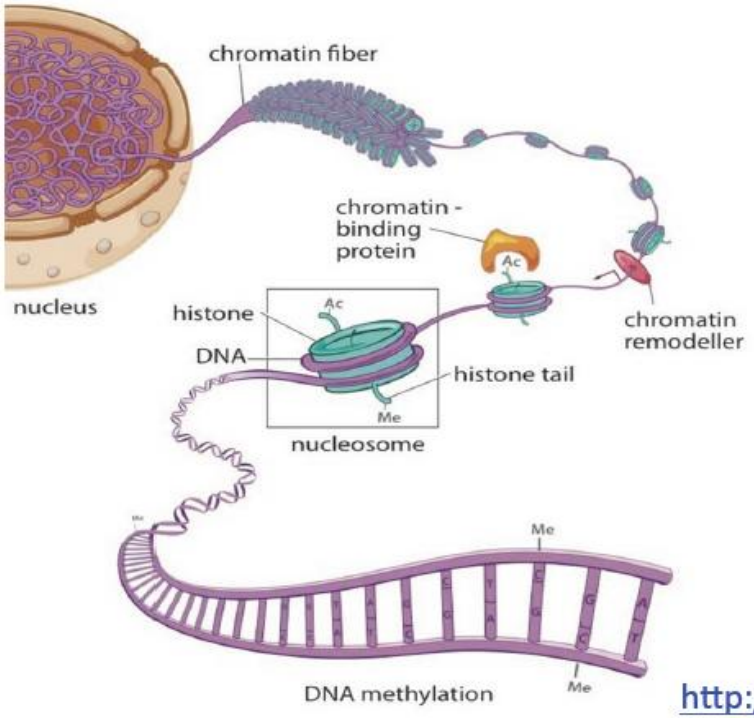
Regulation of the gene expression

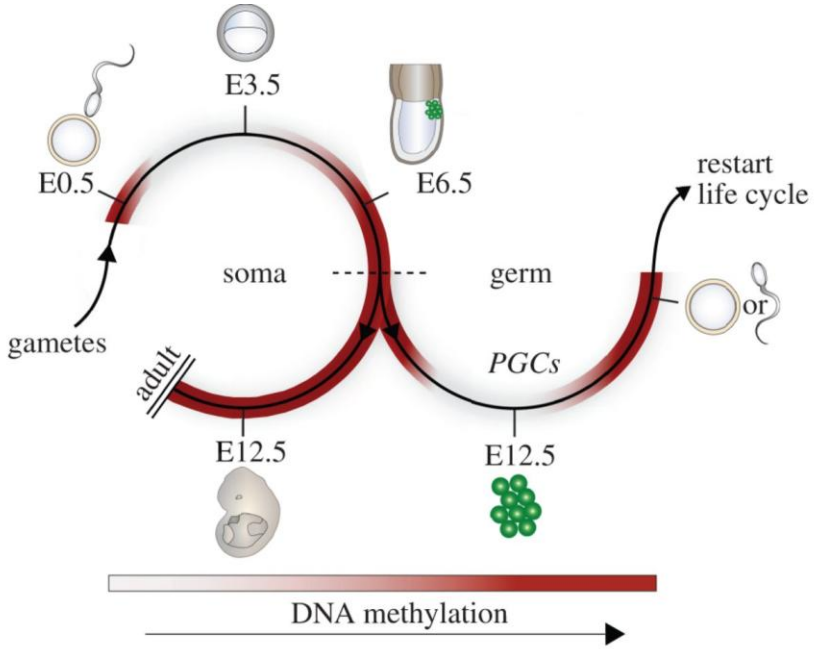


http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

Epigenetic Regulation

• Genes are silenced or activated by epigenetically modifying the access of transcriptional machinery to the DNA
• e.g. signalled by methylation of DNA = off genes.
•The access is moderated by regulating how compact the DNA is,

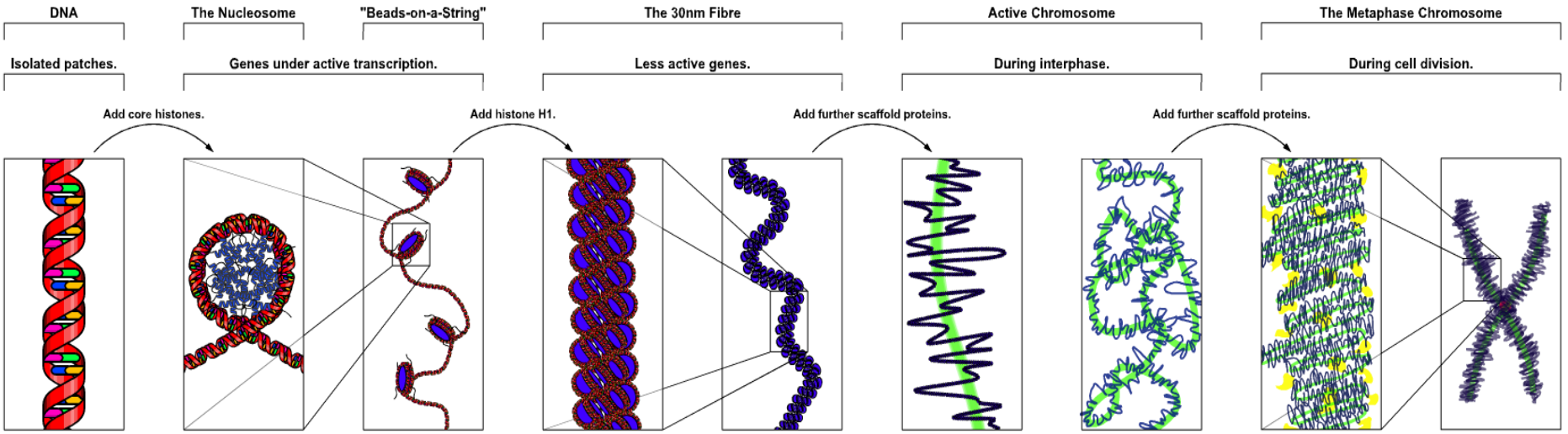




http://www.landesbioscience.com/

http://rstb.royalsocietypublishing.org/content/royptb/368/1609/20110328/F1.large.jpg

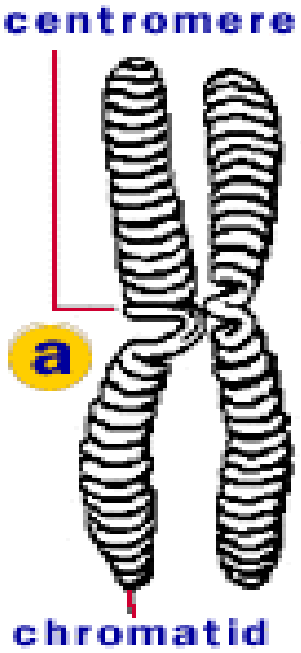
Chromatin and Chromosome compact the genome

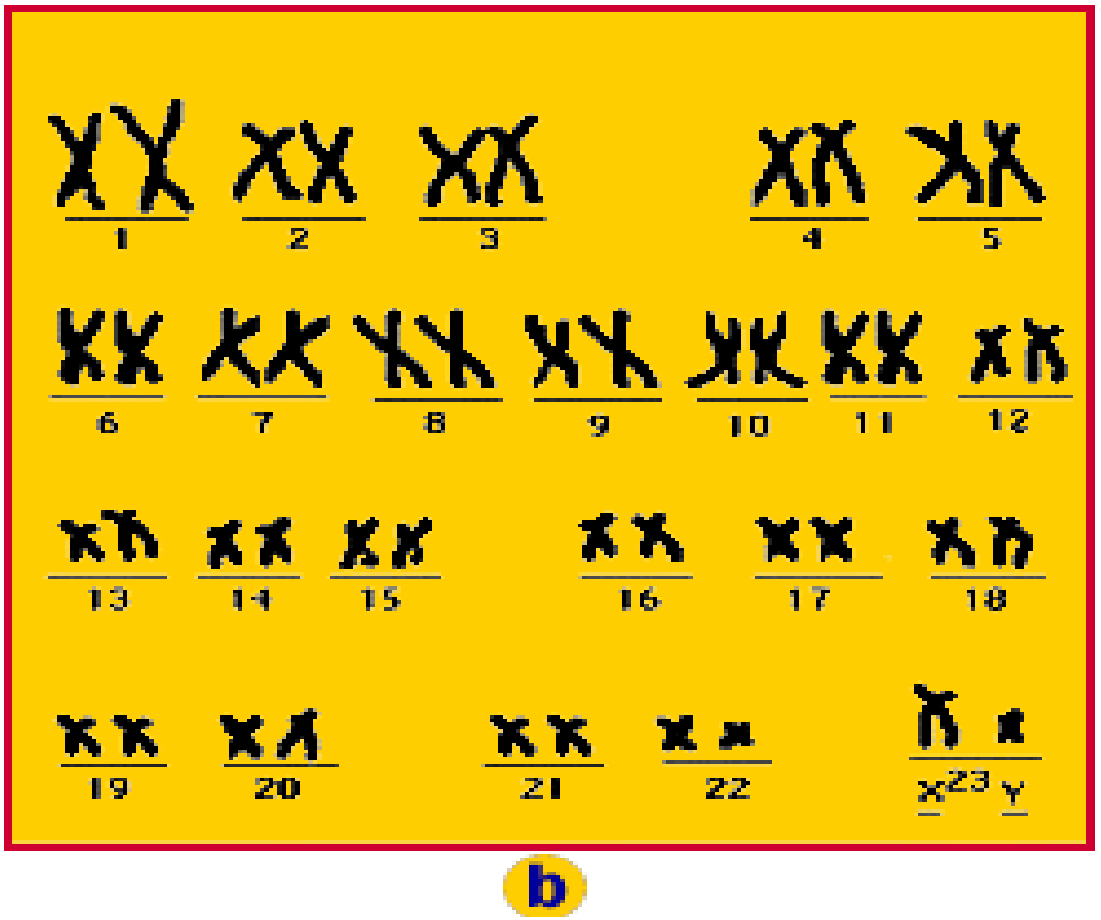


http://upload.wikimedia.org/wikipedia/commons/4/4b/Chromatin_Structures.png

In eukaryotic cells only, the prokaryotes counterpart is genophore

Chromosome





http://whyfiles.org/034clone/images/human_chrom.gif

http://www.iitk.ac.in/infocell/Archive/dirnov3/scienc3.jpg

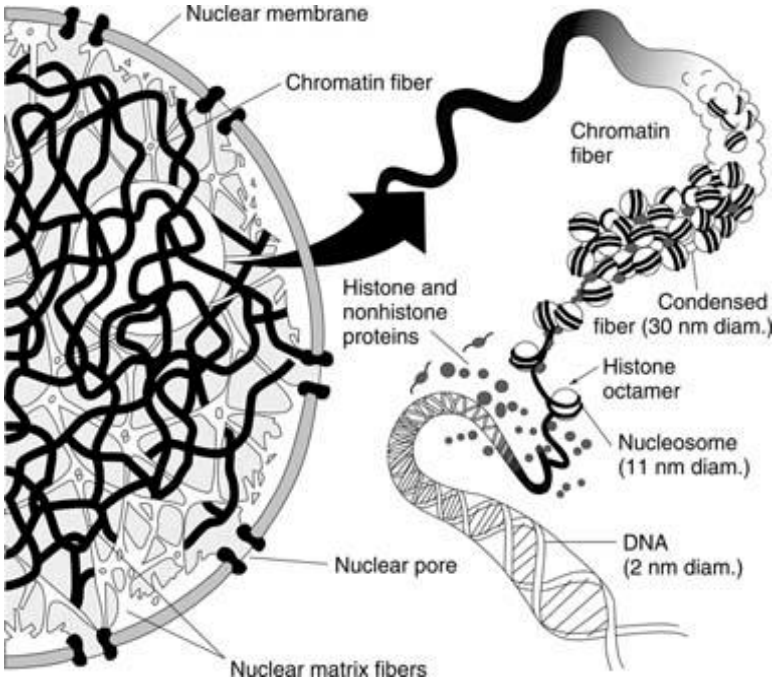
Size of genome vs. size of nucleus

General GENOME Sizes

Yeast 12 million bp
Worm 100 million bp
Fruit Fly 133 million bp
Human 3.3 billion bp
Mouse 3.4 billion bp
Red Viscacha Rat 8.2 billion bp
Mountain Grasshopper 16.5 billion bp

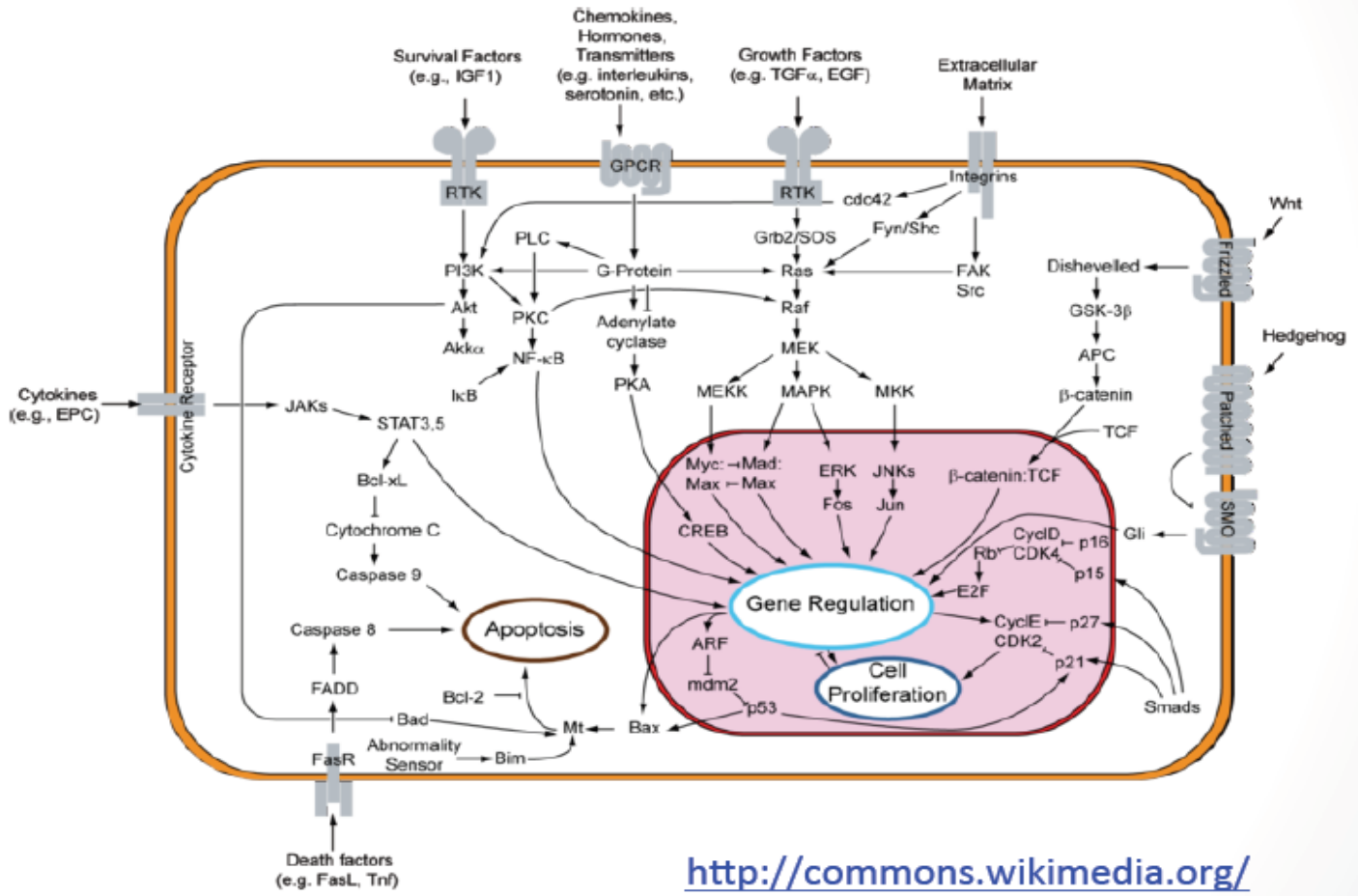
http://info.gersteinlab.org/Genome_Statistics

Avg. diameter of nucleus of mammalian cells
= 6 micrometers



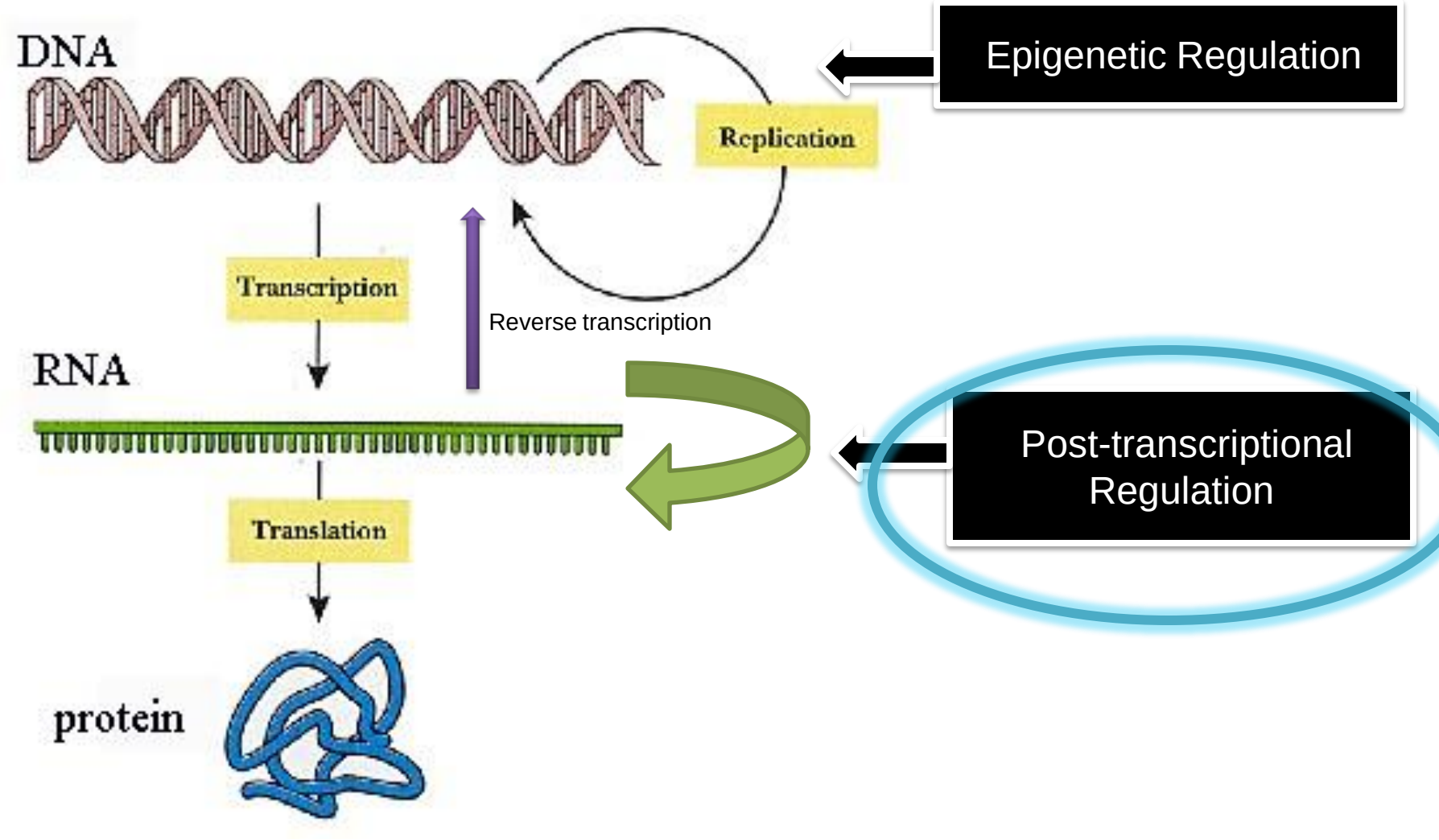
http://www.cliffsnotes.com/assets/24452.jpg

Regulation – Signal Transduction



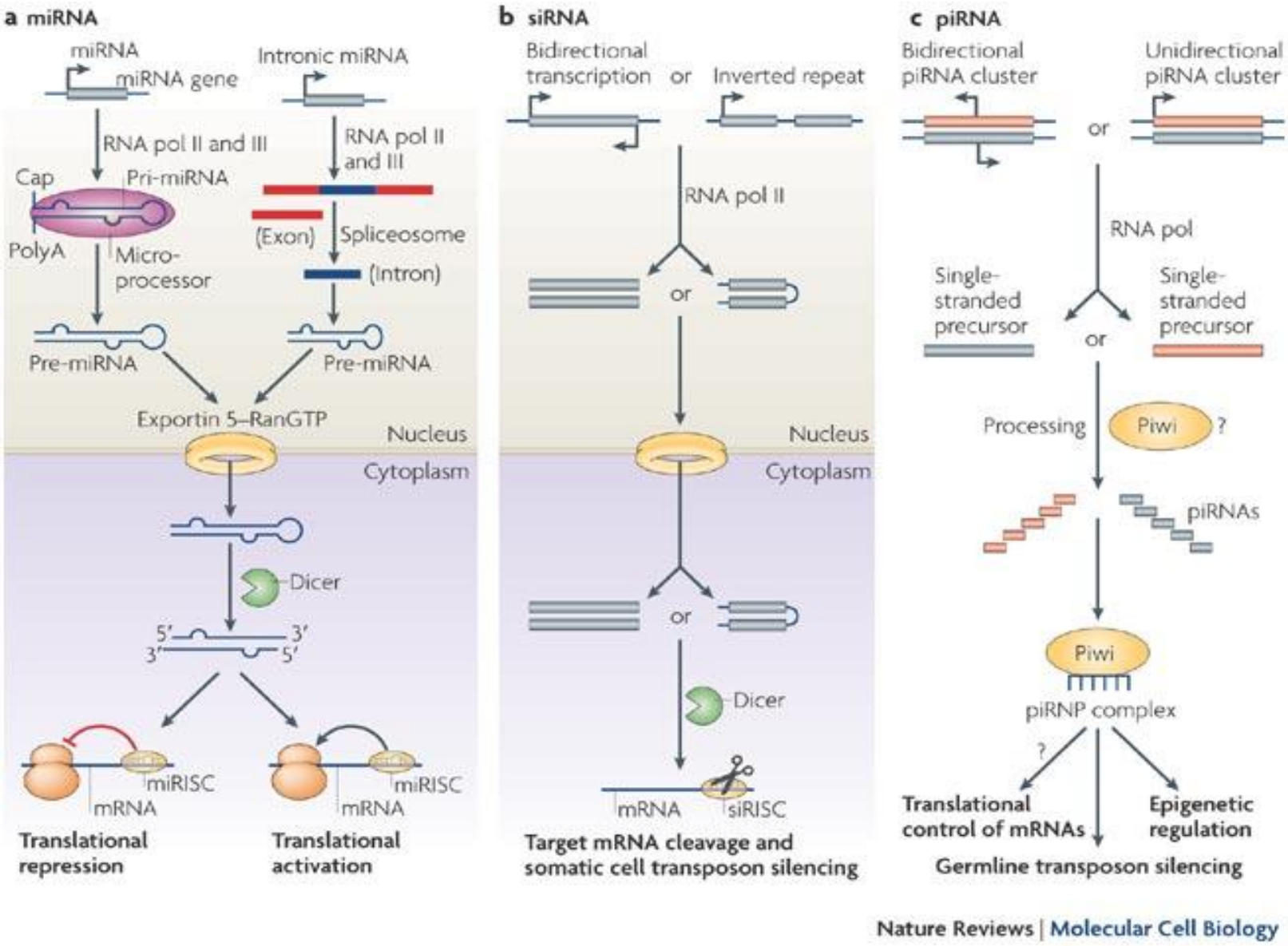
http://commons.wikimedia.org/

Regulation of the gene expression



http://www.angelfire.com/dc/apgenetics/central.dogma.jpg

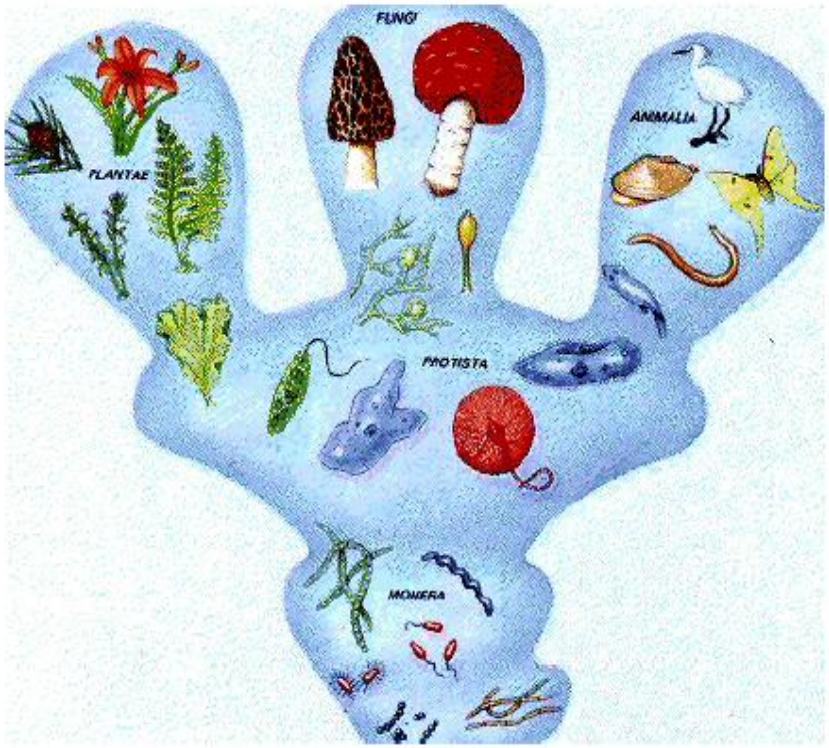
Post-transcriptional regulation: miRNA, siRNA, piRNA



Summary: Regulation of the gene expression

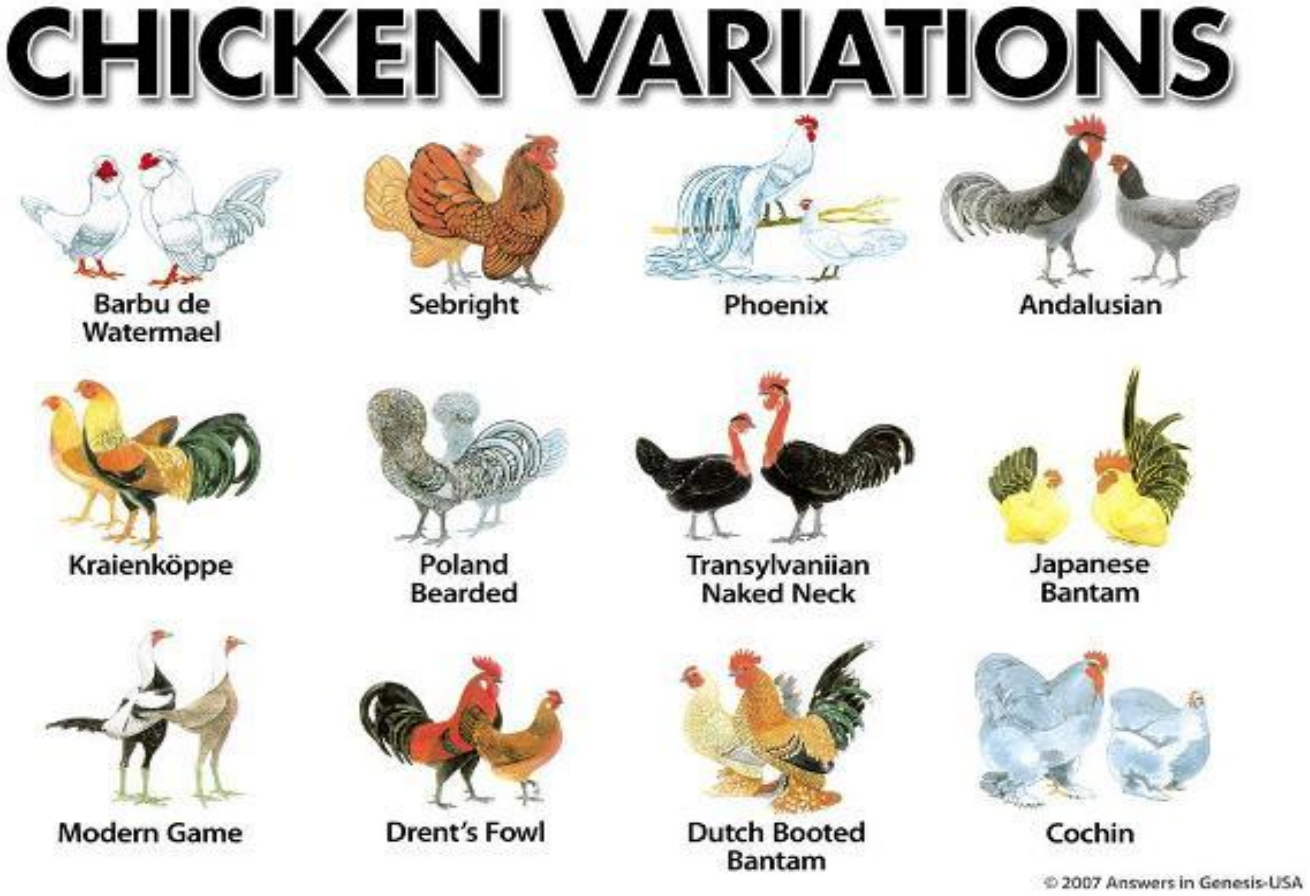
- Activate or silence gene expression
- Control on each level of the cascading flow for gene expression,
- Some control may entail multiple levels
- The regulation involves various type and quantity of molecules, such as RNA, noncoding RNA, protein, DNA, which is still in need of understanding.

Genetic variation: inter-species



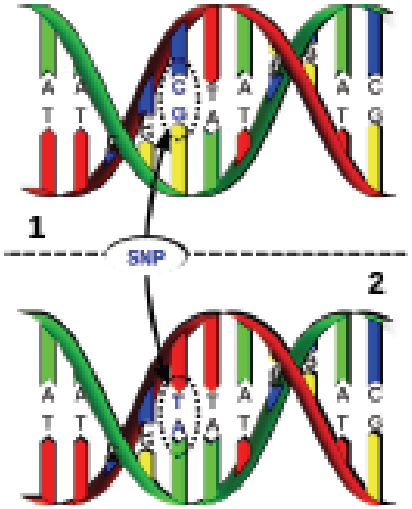
http://bio8.wikispaces.com/

Genetic variation: intra-species



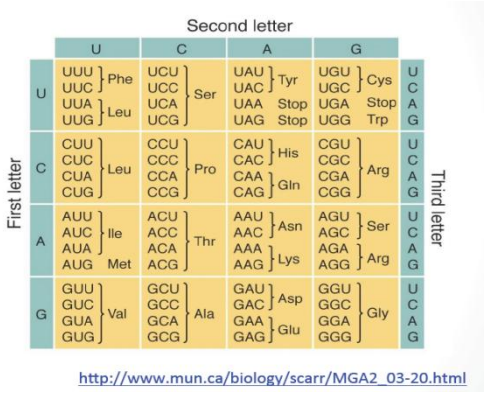
https://cdn-assets.answersingenesis.org/img/articles/2008/08/chicken-variations.jpg

Single Nucleotide Variation (SNV)
Single Nucleotide Polymorphism (SNP)



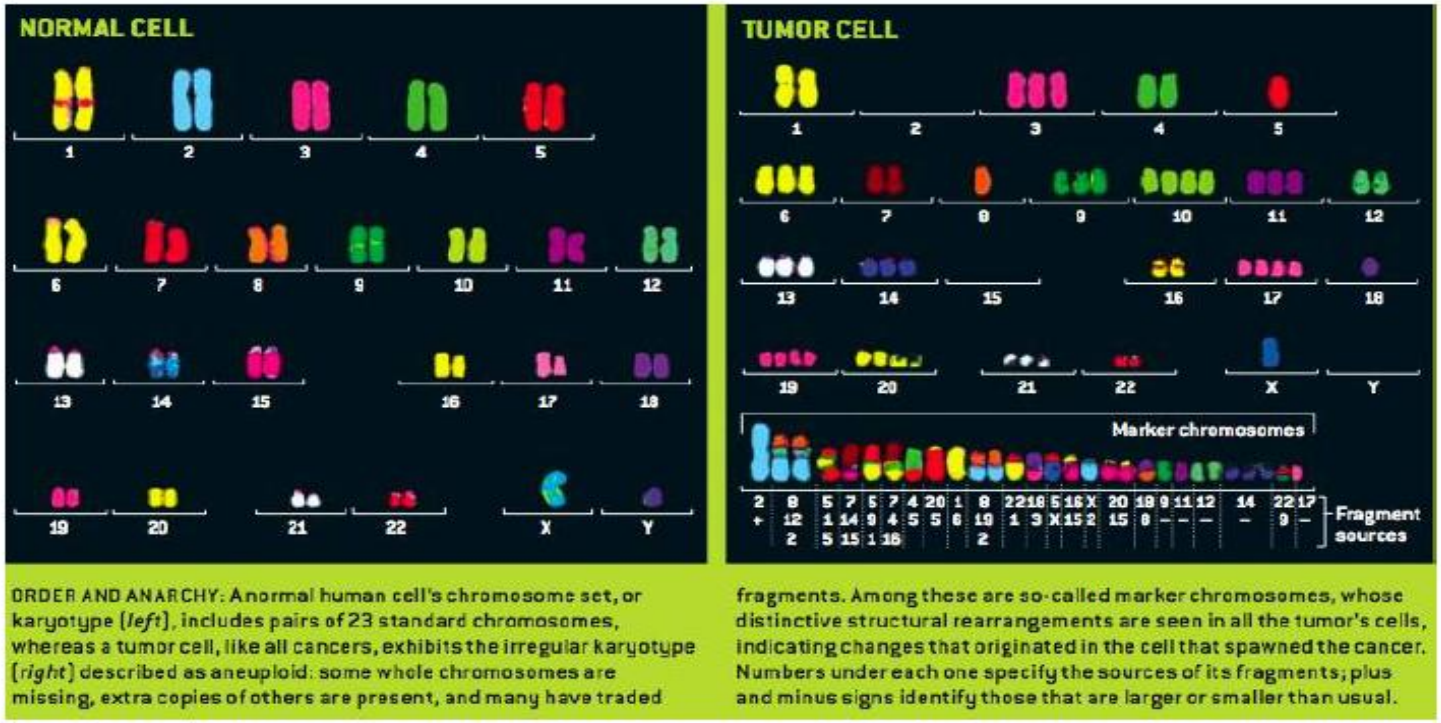
h2p://wikipedia.org

- SNP: the alternative base of the SNV occur > 1% population
- Most are biallelic, synonymous or non-synonymous due to codon degeneracy
- *e.g.* 1 per 1,000-2,000 bp in human genome, giving rise to genetic diversity
- may be disease-related: Osteoporosis(SMAD1) sickle-cell anemia(SNP in β-globin gene)
- Popular in Genome-wide association study(GWAS) which investigates the association (cosegregation) between the SNPs and a trait (disease or response to drug, *e.g.* Warfarin)



http://www.mun.ca/biology/scarr/MGA2_03-20.html

Disease-related: Genetic variation within the same species or within one individual



ORDER AND ANARCHY: A normal human cell's chromosome set, or karyotype (*left*), includes pairs of 23 standard chromosomes, whereas a tumor cell, like all cancers, exhibits the irregular karyotype (*right*) described as aneuploid: some whole chromosomes are missing, extra copies of others are present, and many have traded fragments. Among these are so-called marker chromosomes, whose distinctive structural rearrangements are seen in all the tumor's cells, indicating changes that originated in the cell that spawned the cancer. Numbers under each one specify the sources of its fragments; plus and minus signs identify those that are larger or smaller than usual.

Reference: Chromosomal Chaos and Cancer, Peter Duesberg

Examples:

Copy Number Variation(CNV): Addition or deletion of copy number , e.g. Huntington's disease

Loss of Heterozygosity(LOH): initially heterozygous for a mutant allele, but mutation (point or deletion, chromosomal deletion/breaks, recombination) converts it to homozygous of mutant allele.

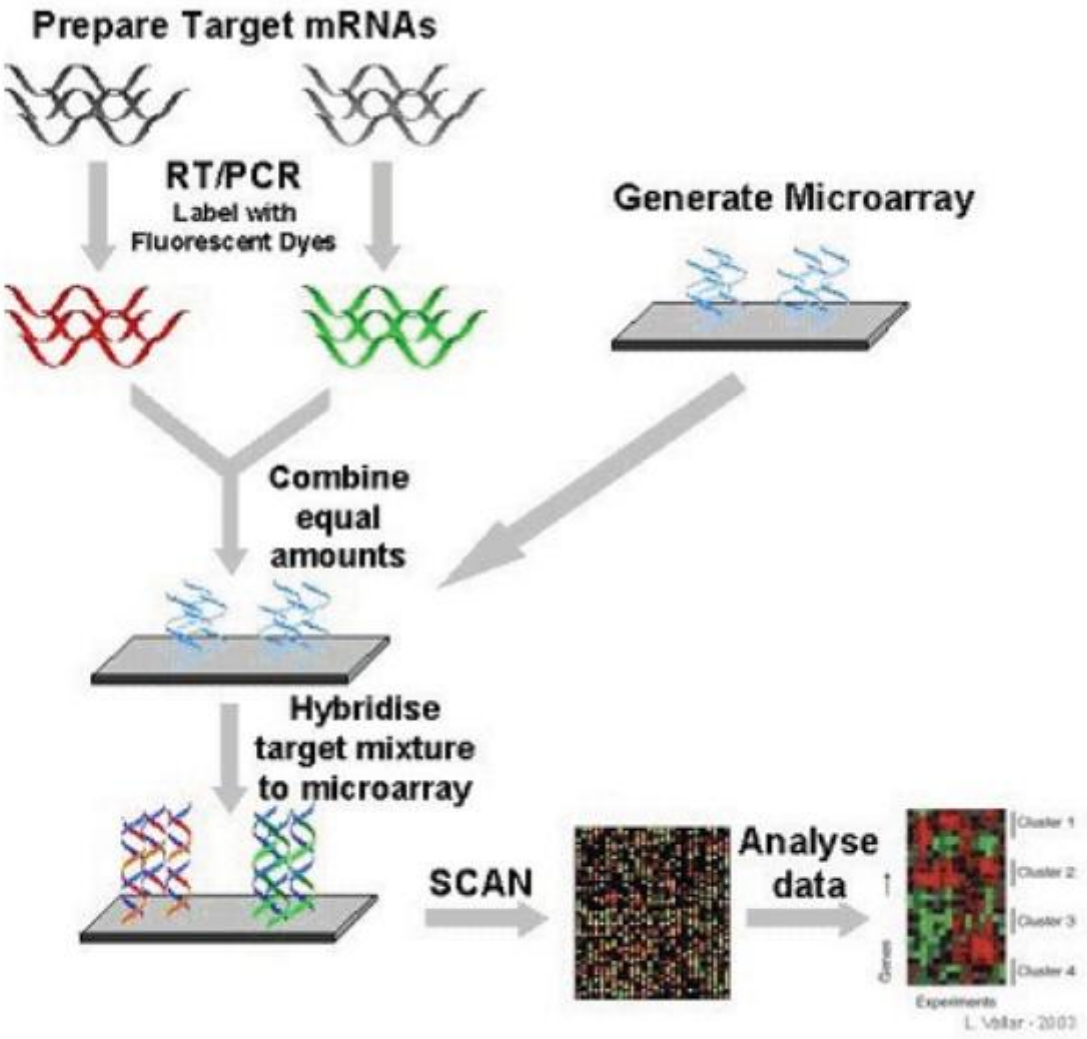
Genetic Variation: Summary

- Inter- or intra- species, or among the cells within the same individual

- Could be disease-causing, or the coordination between a number of variation together contribute to the disease manifestation

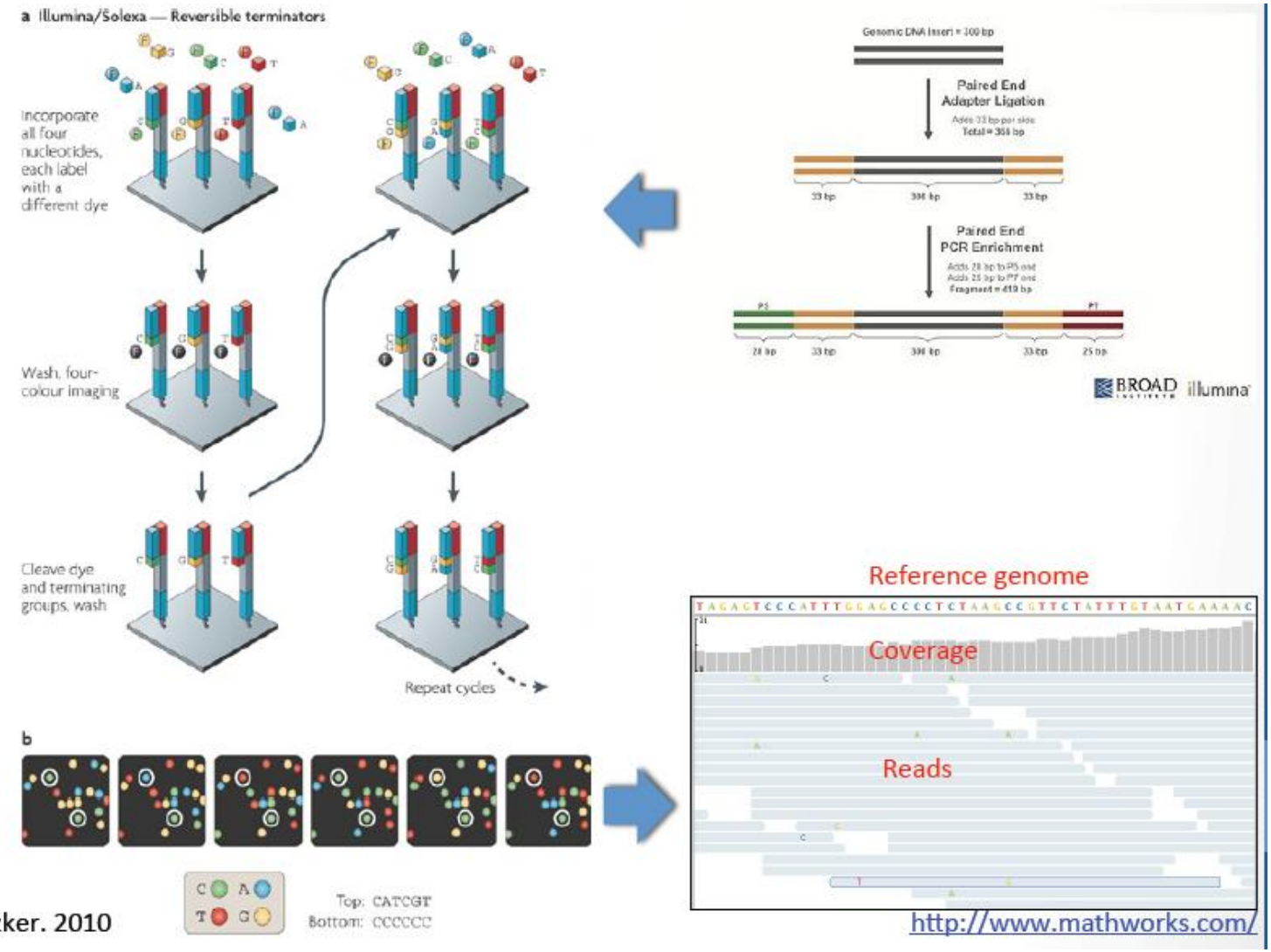
- SNP is widely utilized in GWAS study

How is the genome data generated?



http://www.microarray.lu/en/MICROARRAY_Overview.shtml

Next Generation Sequencing



Metzker. 2010

NGS

High-end sequencing- Platform†	Sequencing chemistry	Read lengths/ through put	Run time	Template prep	Application
Roche 454 -Titanium FLX	Pyrosequencing	400 bp 400 Mb/run	10 hours	Emulsion PCR	Denovo WGS of microbes, pathogen discovery, Exome seq
Illumina/Solexa -HiSeq 2000	Reversible terminator chemistry	2×100bp 600 GB/ run (dual cell)	11.5 days	Solid-phase	Human WGS, exome seq, RNA-seq, Methylation
ABI/LifeTechnology-SOLiD 5550XL	Sequencing by ligation	2×60bp 15 GB/day	8 days	Emulsion PCR	Human WGS, exome seq, RNA-seq, Methylation
HelicosBiotechnologies	Reversible Terminator chemistry	25-55 bp 28 GB/run (avg)	>1 GB/hour	Single molecule	Human WGS, exome seq, RNA-seq, Methylation
Roche 454- GS Junior	Pyrosequencing	400 bp 50 Mb/run	10 hours	Emulsion PCR	Denovo WGS of microbes, pathogen discovery, Exome seq
Illumina/Solexa- MiSeq	Reversible terminator chemistry	2×150bp 1.0-1.4 Gb	26 hours	Solid-phase	Microbial discovery, Exome seq, Targeted capture
ABI/ Lifetechnology- Iontorrent	H+ Ion sensitive transistor	320 Mb/run	8 hours*	Emulsion PCR	Microbial discovery, Exome seq, Targeted capture

*Sample preparation – 6 hours, sequencing time – 2 hours, †Data shown here represent the highest figures currently available on the company website and is highly likely to change by the time this article is published

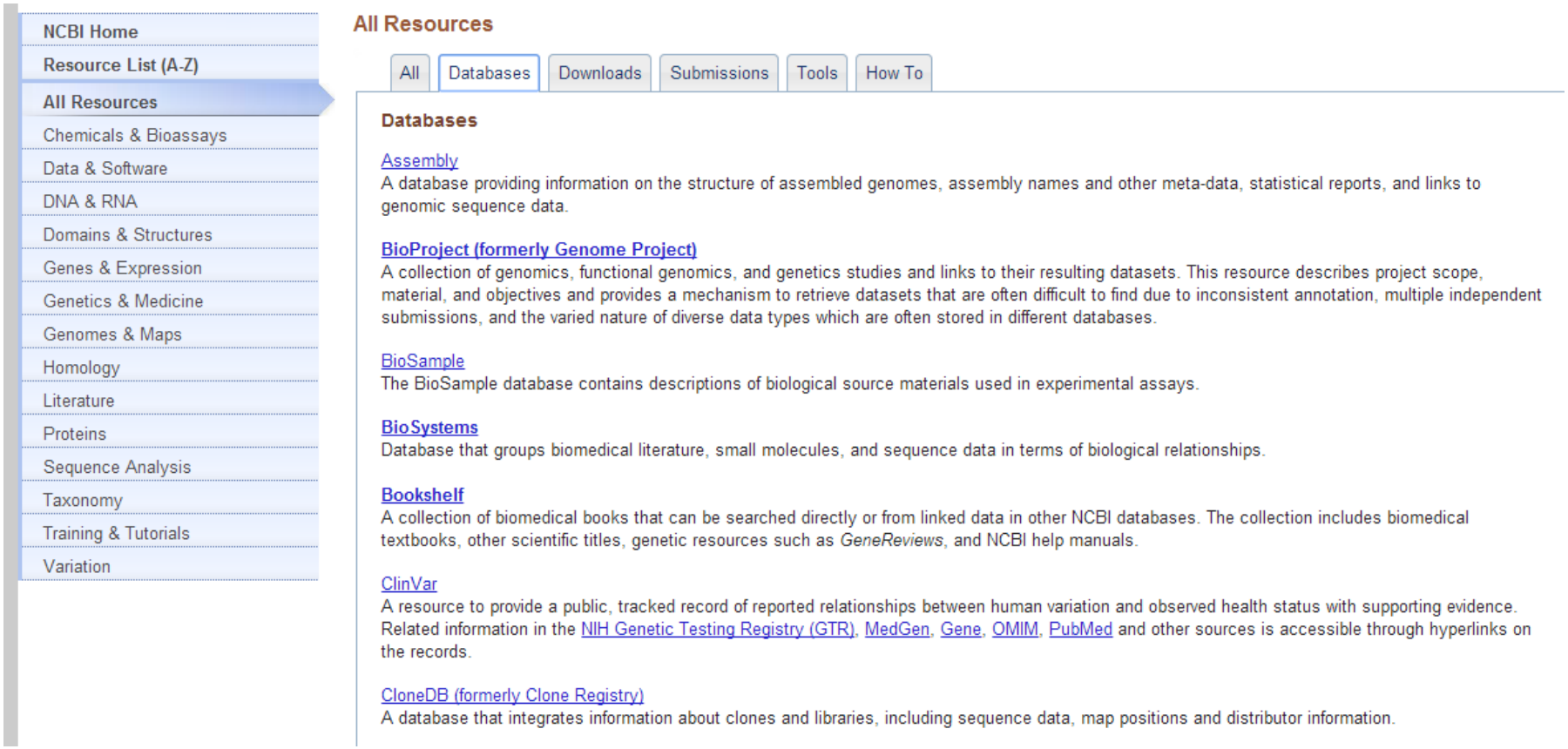
http://www.jpathinformatics.org/articles/2012/3/1/images/JPatholInform_2012_3_1_40_103013_b2.jpg

Technology – Applications

- DNA => genetic variations
  - SNP array
  - Whole genome / exome / targeted sequencing
- RNA => differential expression
  - Expression array
  - RNA-seq / miRNA-seq
- Epigenome => differential methylation / chromatin state
  - DNA methylation array
  - Bisulfite sequencing
  - Methylated DNA immunoprecipitation (MeDIP-seq)
  - Chromatin immunoprecipitation sequencing (ChIP-seq) for histone modifications

From Gloria's slide

Database


NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
All Resources
All
Databases
Downloads
Submissions
Tools
How To
Databases
Assembly
A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.
BioProject (formerly Genome Project)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.
BioSample
The BioSample database contains descriptions of biological source materials used in experimental assays.
BioSystems
Database that groups biomedical literature, small molecules, and sequence data in terms of biological relationships.
Bookshelf
A collection of biomedical books that can be searched directly or from linked data in other NCBI databases. The collection includes biomedical textbooks, other scientific titles, genetic resources such as GeneReviews, and NCBI help manuals.
ClinVar
A resource to provide a public, tracked record of reported relationships between human variation and observed health status with supporting evidence. Related information in the NIH Genetic Testing Registry (GTR), MedGen, Gene, OMIM, PubMed and other sources is accessible through hyperlinks on the records.
CloneDB (formerly Clone Registry)
A database that integrates information about clones and libraries, including sequence data, map positions and distributor information.

Summary

- DNA → RNA → Protein
  → non-coding RNA

- Genes from the genome are expressed under intricate regulation.

- Variation in the genome may result in (i) disease manifestation, (ii) response to treatment, and need to be extracted from the background noise.

- Technology enables high-throughput production of data pertaining to DNA, RNA and Protein, and could be used in search for underlying genomic cause of certain conditions.