Introduction to Bioinformatics

Introduction to Molecular Biology

Lecturer: Jan Baumbach
Teaching assistant(s): Diogo Marinho

Course details

- When: Weeks 36 - 44
- Where: IMADA seminar room

- Wednesdays: Exercises (12-14)
- Thursdays: Classes (10-12)

- Afterwards… Group projects

Course details

Exercises

- Exercise sheets published online on Thursdays after the class at www.baumbachlab.net
- Hand-in due Monday, the following week
- Prepare in small groups
- **Part of the examination!** → Each week one student will be randomly chosen to present the results of one exercise task.
- TAs will help to correct these results

Course details

Week 46 – Nov 12 at 17-19: Get together at OUH for breath analysis live demo and **project data sampling**

Course details

Week 46 – Nov 12 at 17-19: Get together at OUH for breath analysis live demo and **project data sampling**

Afterwards: Breath analysis **group projects** with weekly office hours for help requests (TAs)

Week 48: Short intermediate reports by email to TAs

Week 50: Short intermediate reports by email to TAs

Week 3: Final report and project hand-in by email to TAs

Course details

Somewhen (exact day to be determined):

FINAL EXAM:

- Individual oral examination with external examiner
- 15 mins slots – between 8.00 and 18.00 (looong day)
  - 5 mins presentation about your project
  - 2x5 = 5 mins for two questions/answers
  - Questions will be **picked randomly** → Better be prepared for everything covered in the class, project and examination sheets. ;-)

What is bioinformatics?

Define: Bioinformatics

(Molecular) **bio – informatics:** bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical chemistry) and applying informatics techniques (derived from disciplines such as applied maths, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications

As per Oxford English dictionary

Biological Information

- Bio + Informatics: "information science of the biological world"
- DNA, RNA, Protein: information bundles of the cell

- Levels of information content in these bundles
    i. Sequence and Structure
    ii. Gene / protein expression , DNA mutations
    iii. Network of interactions among the macromolecules

Databases

Protein sequence

SWISS-PROT

PIR-International

Macromolecular structures

Protein Data Bank (PDB)

CATH

SCOP

Protein-Protein Interactions

HPRD

BioGrid

Nucleotide sequences

GenBank

EMBL

DDBJ

Genome sequences

Entrez genomes

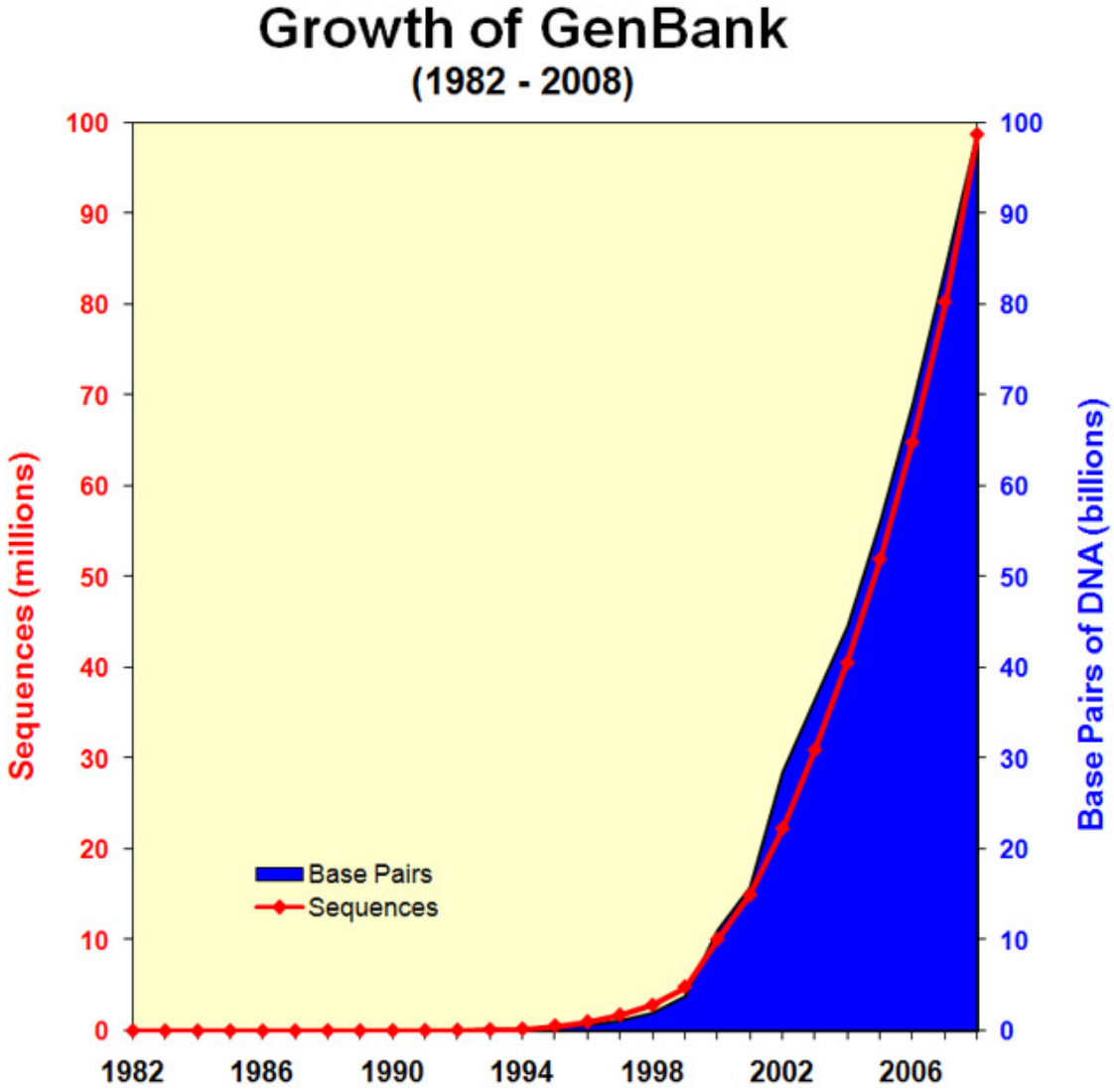
Ensembl

Transcription Factor Binding sites

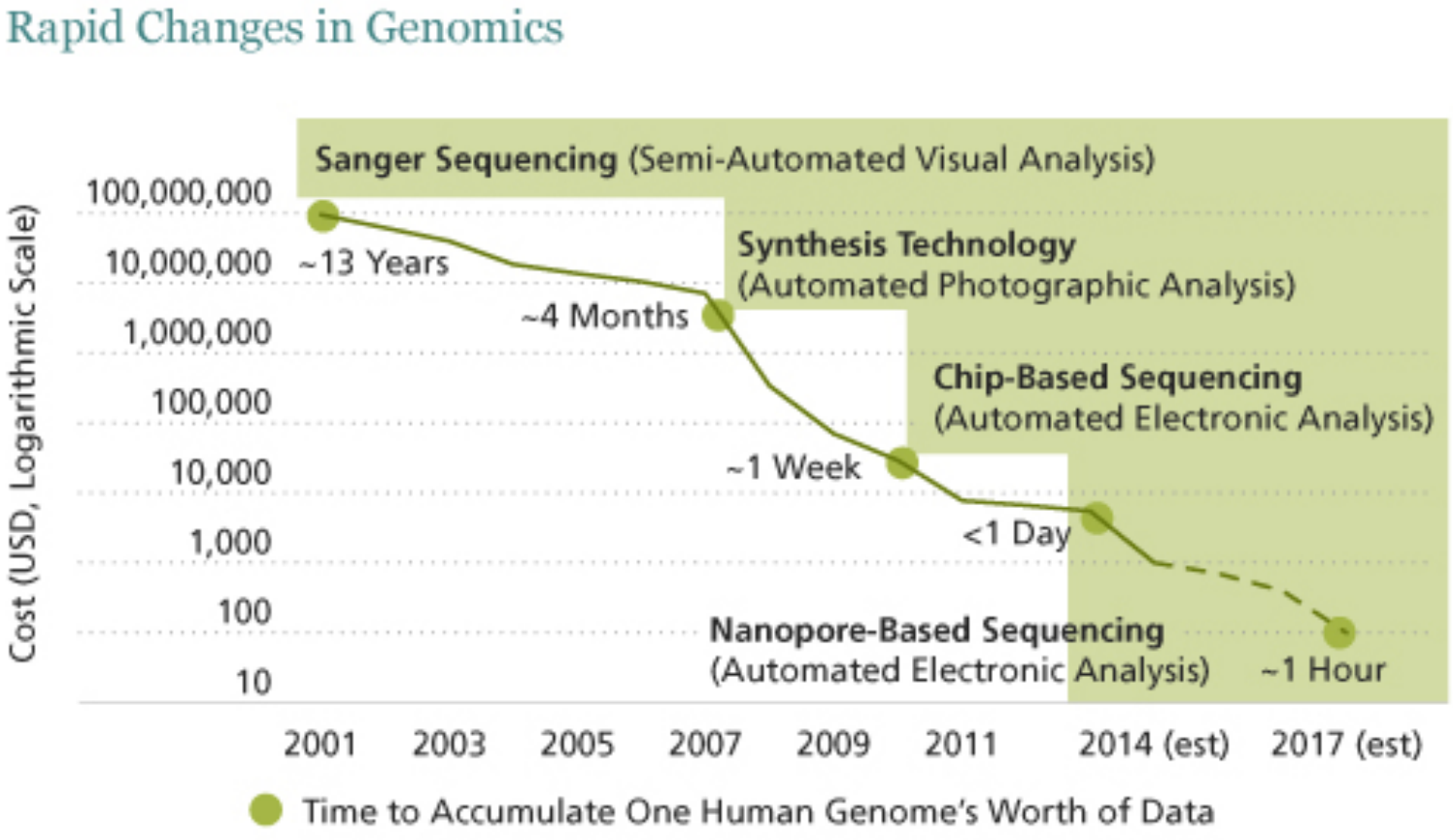
Transfac

Jaspar

Genomic data explosion



And it's getting cheaper and faster



As of November 30, 2013
Source: National Human Genome Research Institute, Genia, Illumina, Life Technologies, Oxford Nanopore Technologies, Washington University and AllianceBernstein

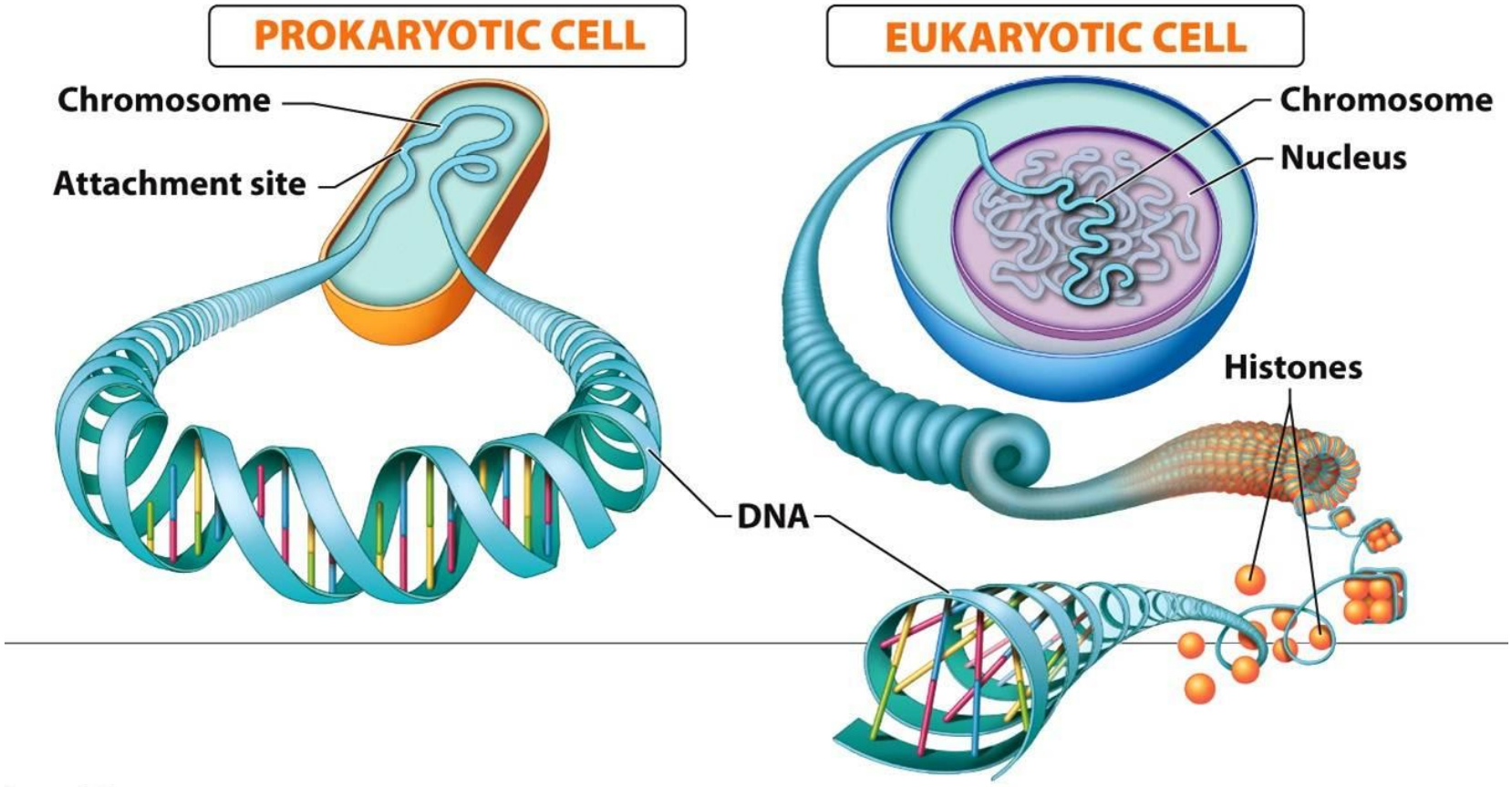
Why do we need bioinformatics?

To understand the molecular basis of life, by developing methods to analyze the massive amount of data at our disposal.

John Nesbit: “We are drowning in information but starving for Knowledge”

Basics of Molecular Biology

Cells



e.g. Bacteria like *E.coli*

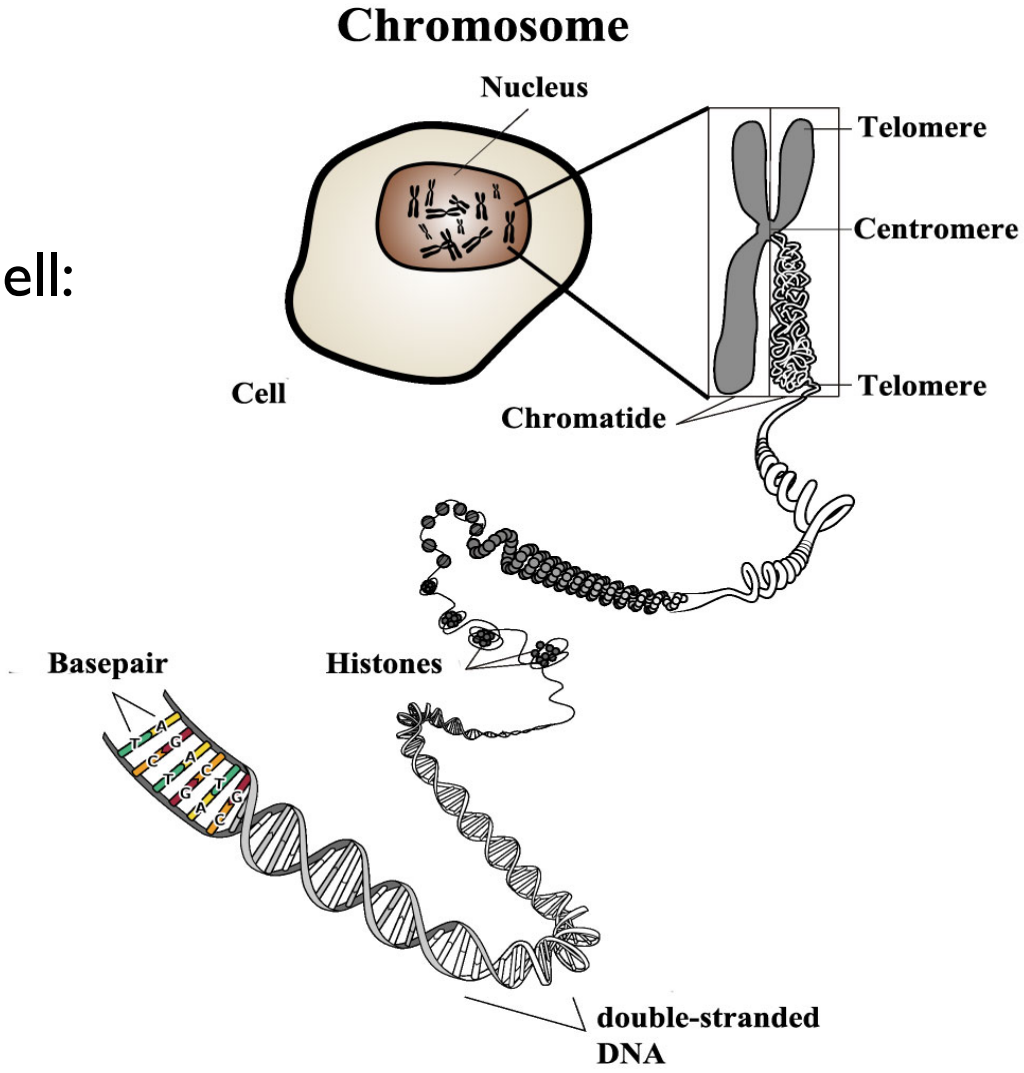
e.g. yeast, potato, fruit fly , humans

Organization of the genome

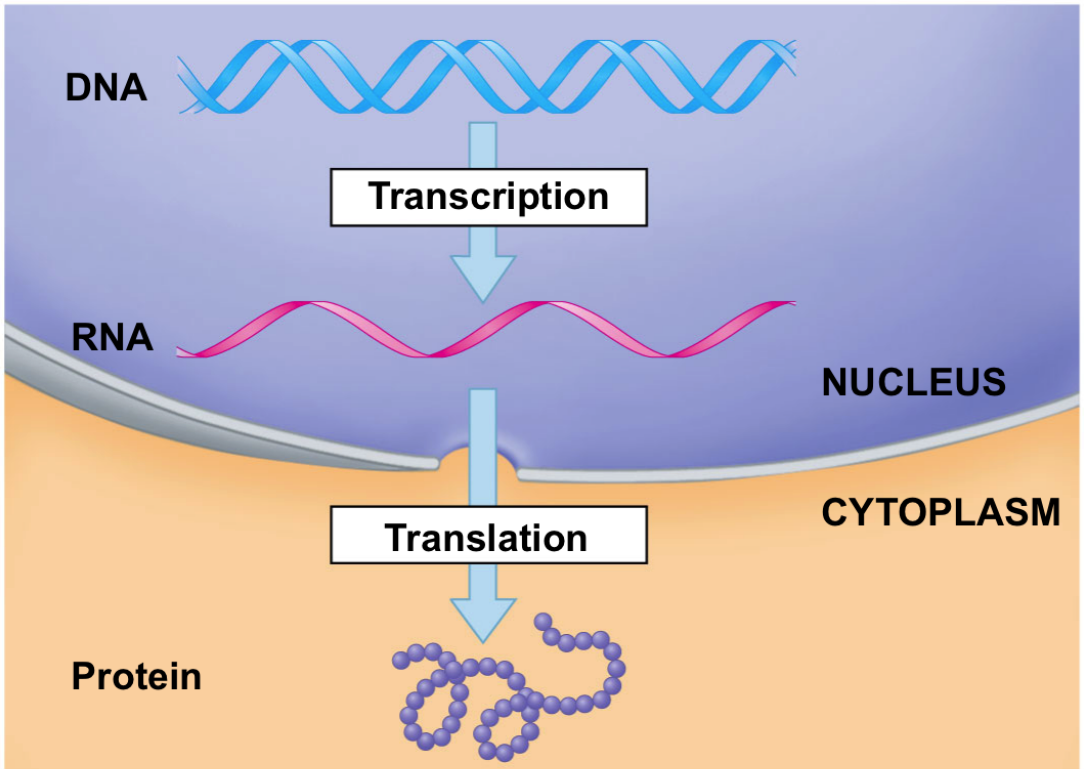
Size of eukaryotic animal cell: 10-30 micro meter

Size of nucleus: 6 micro meter

Length of DNA: 3 meters



Central Dogma of Molecular Biology

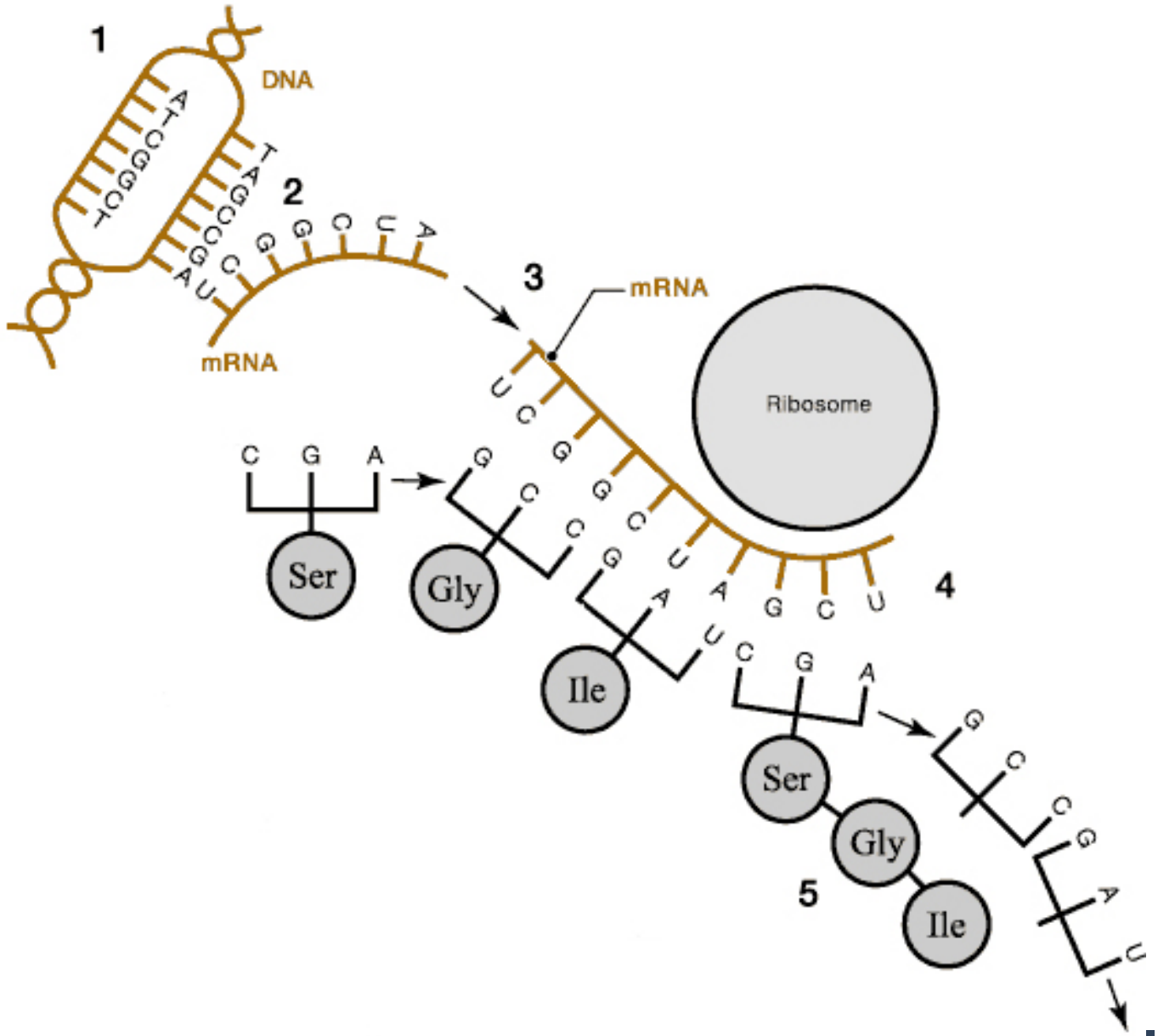


DNA: "the information content of the cell"

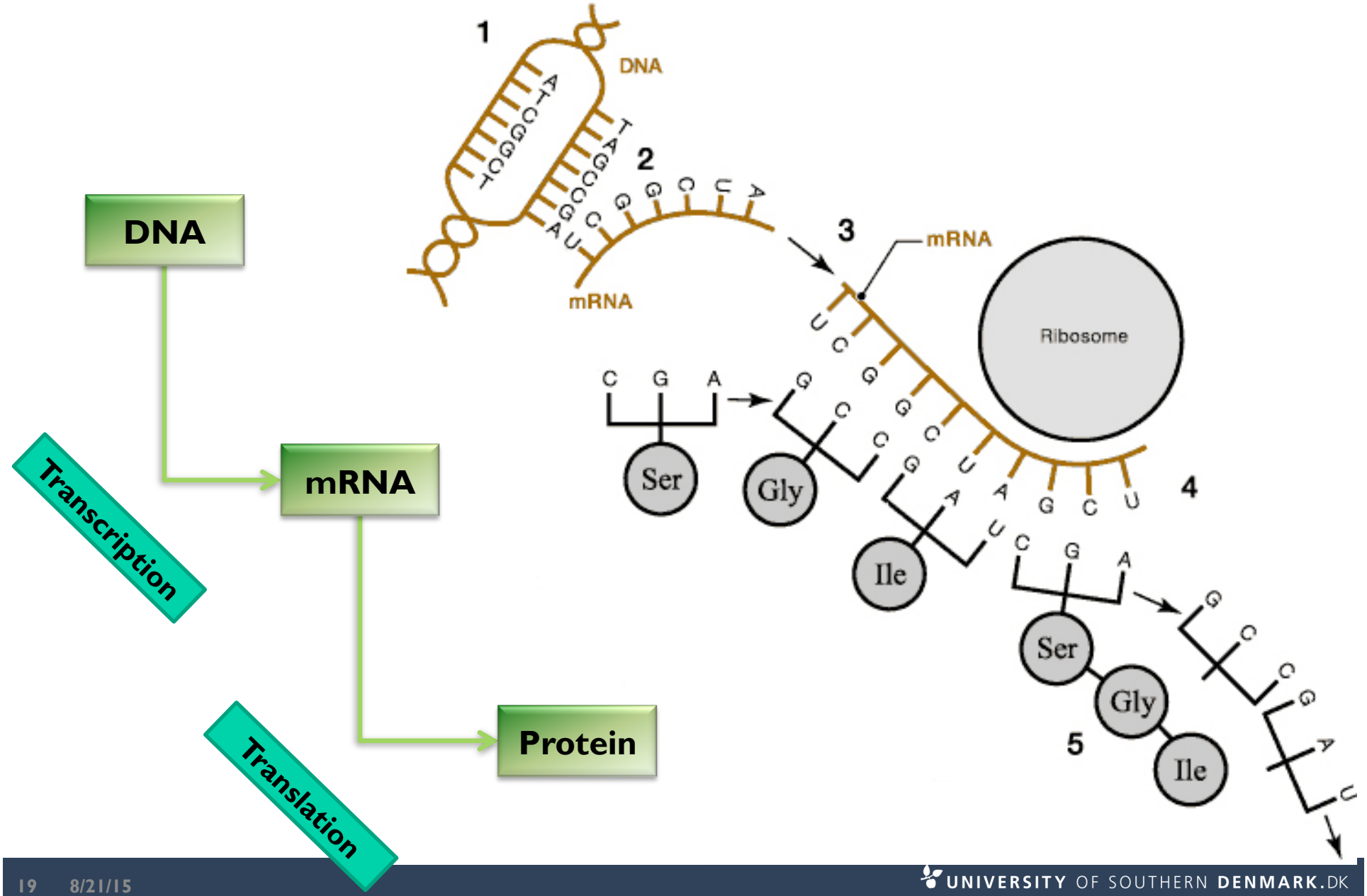
Central dogma: "flow of information in the cell"

Protein: "functional unit of the cell"

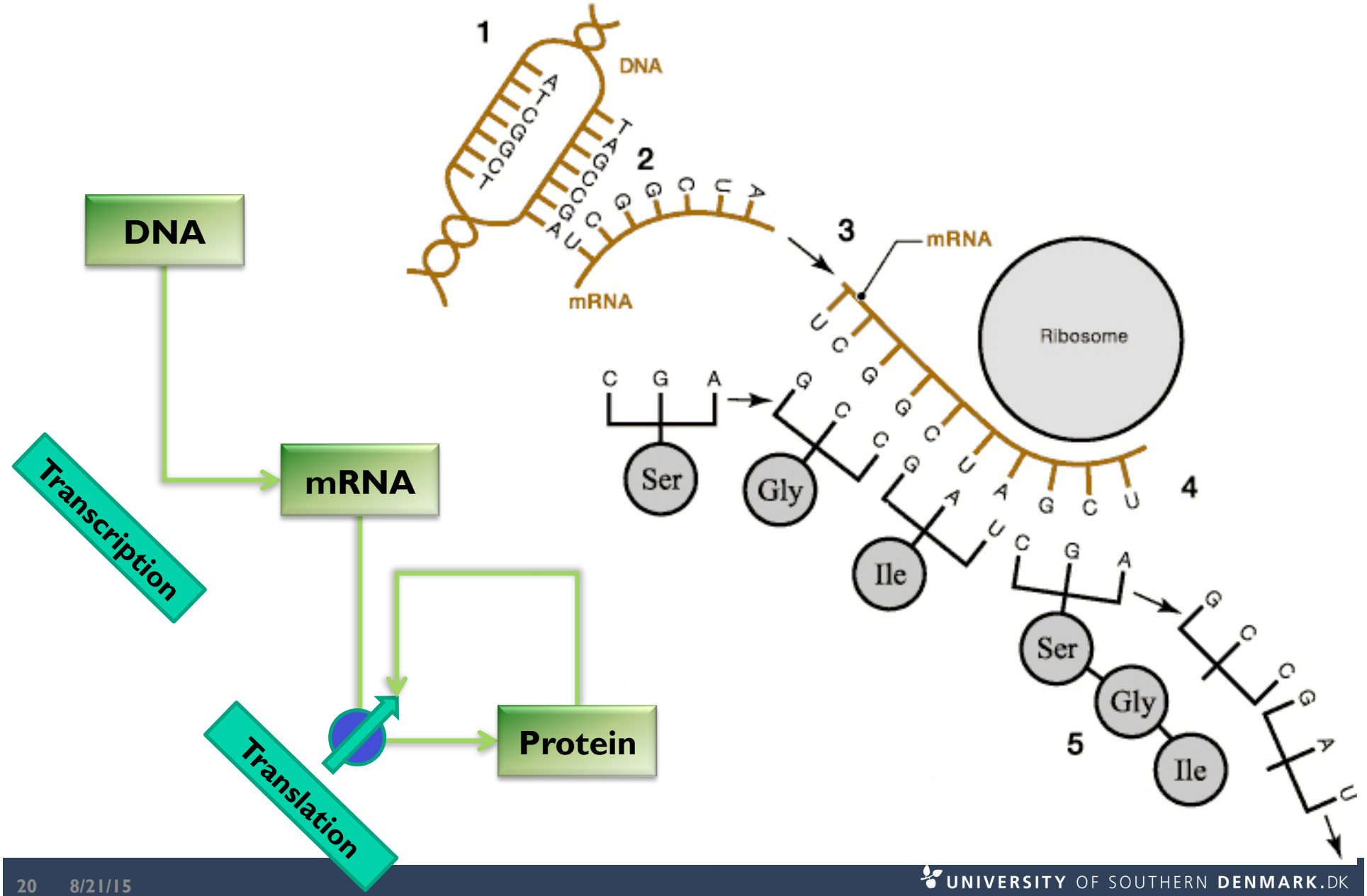
Central dogma



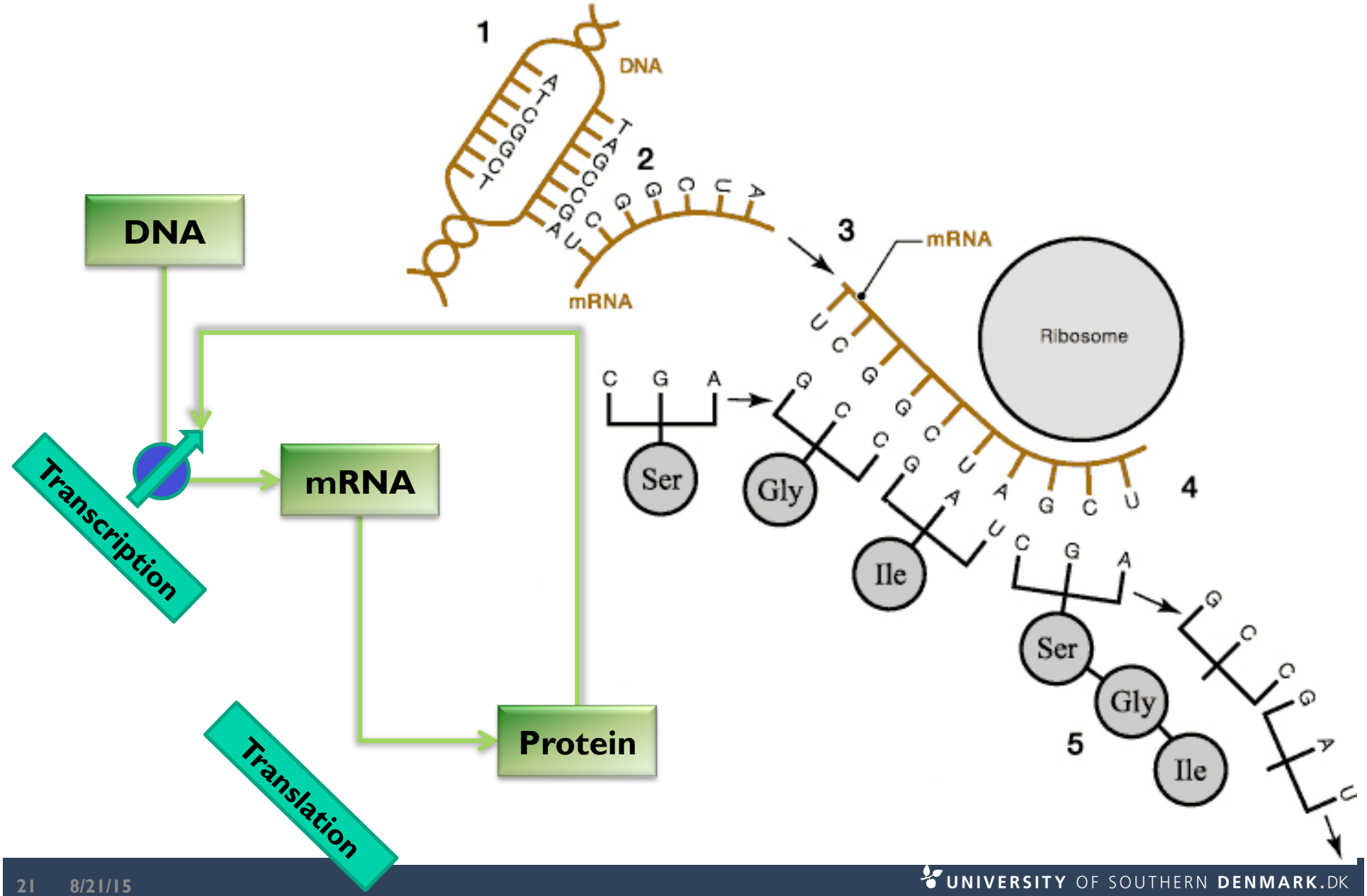
Central dogma



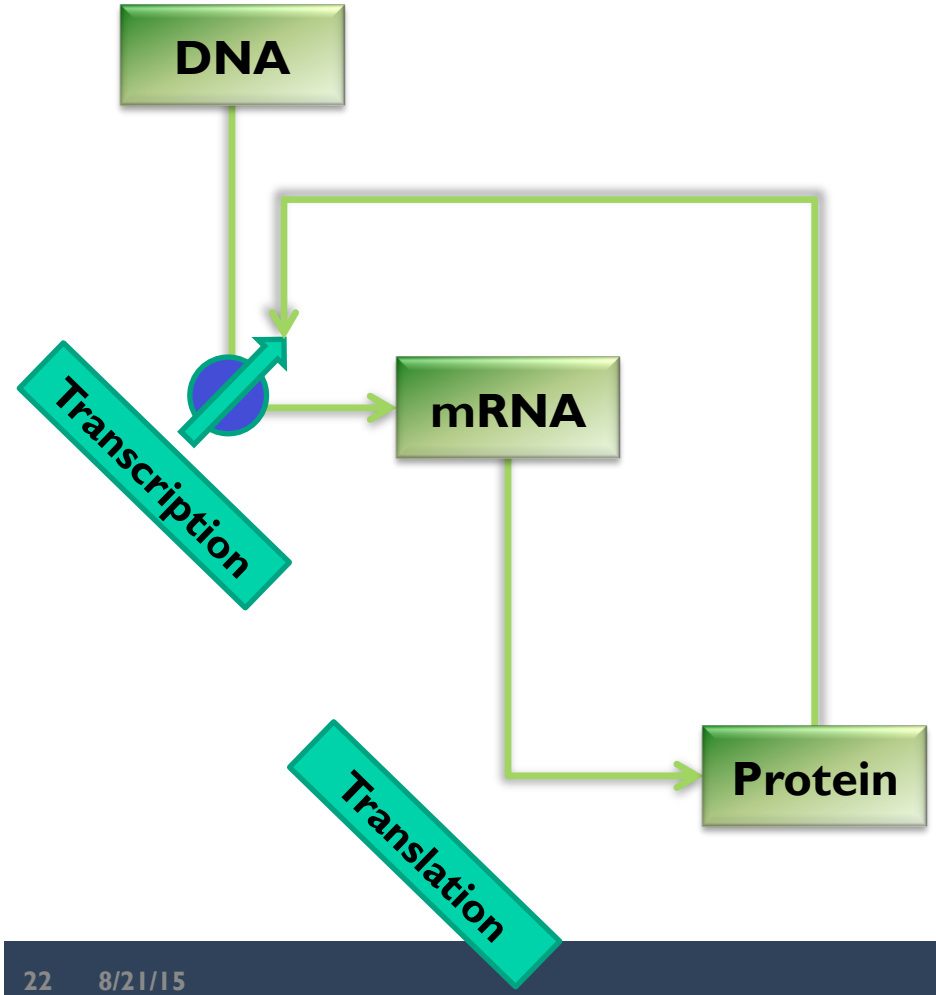
Central dogma



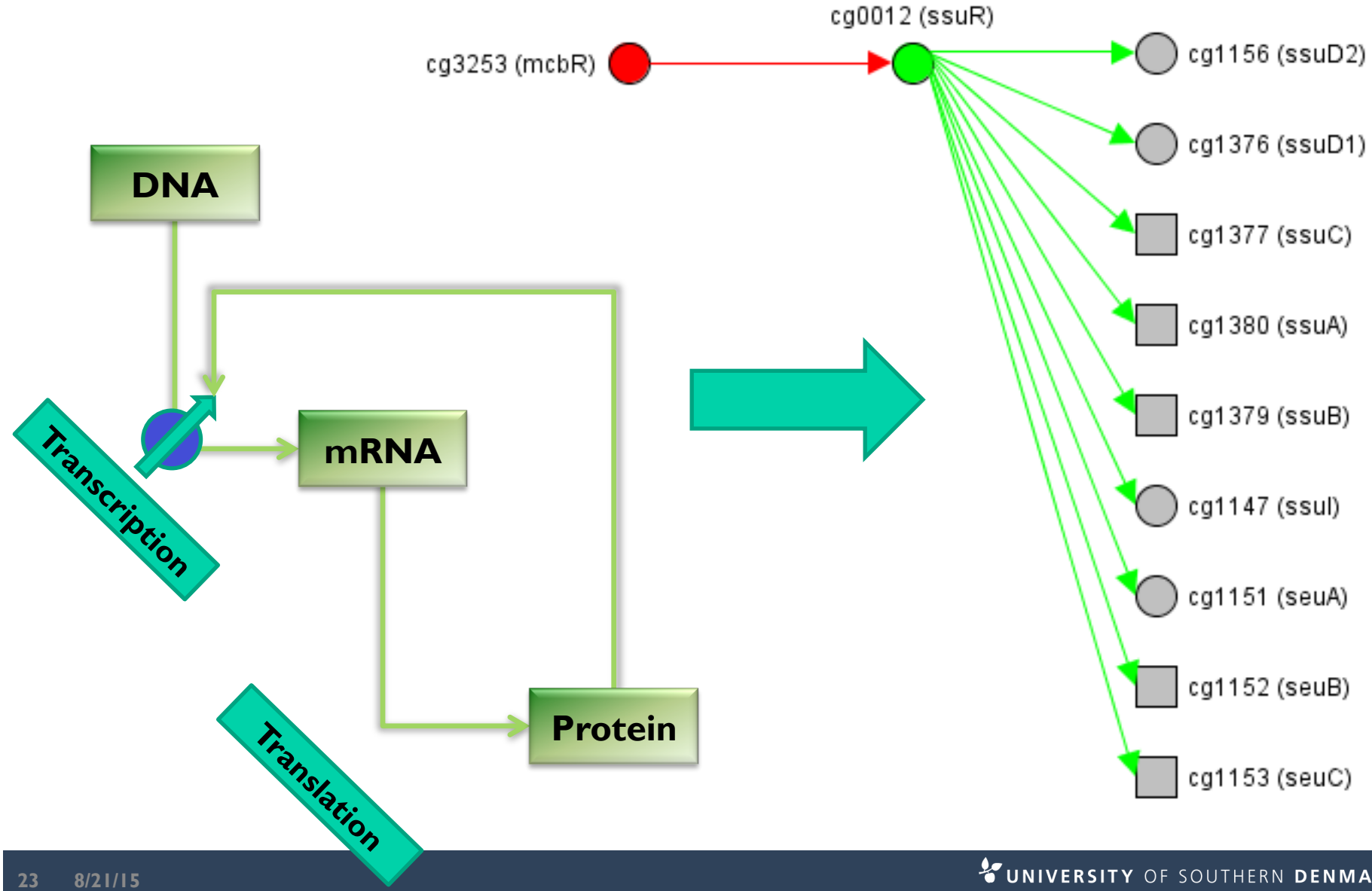
Central dogma



Central dogma

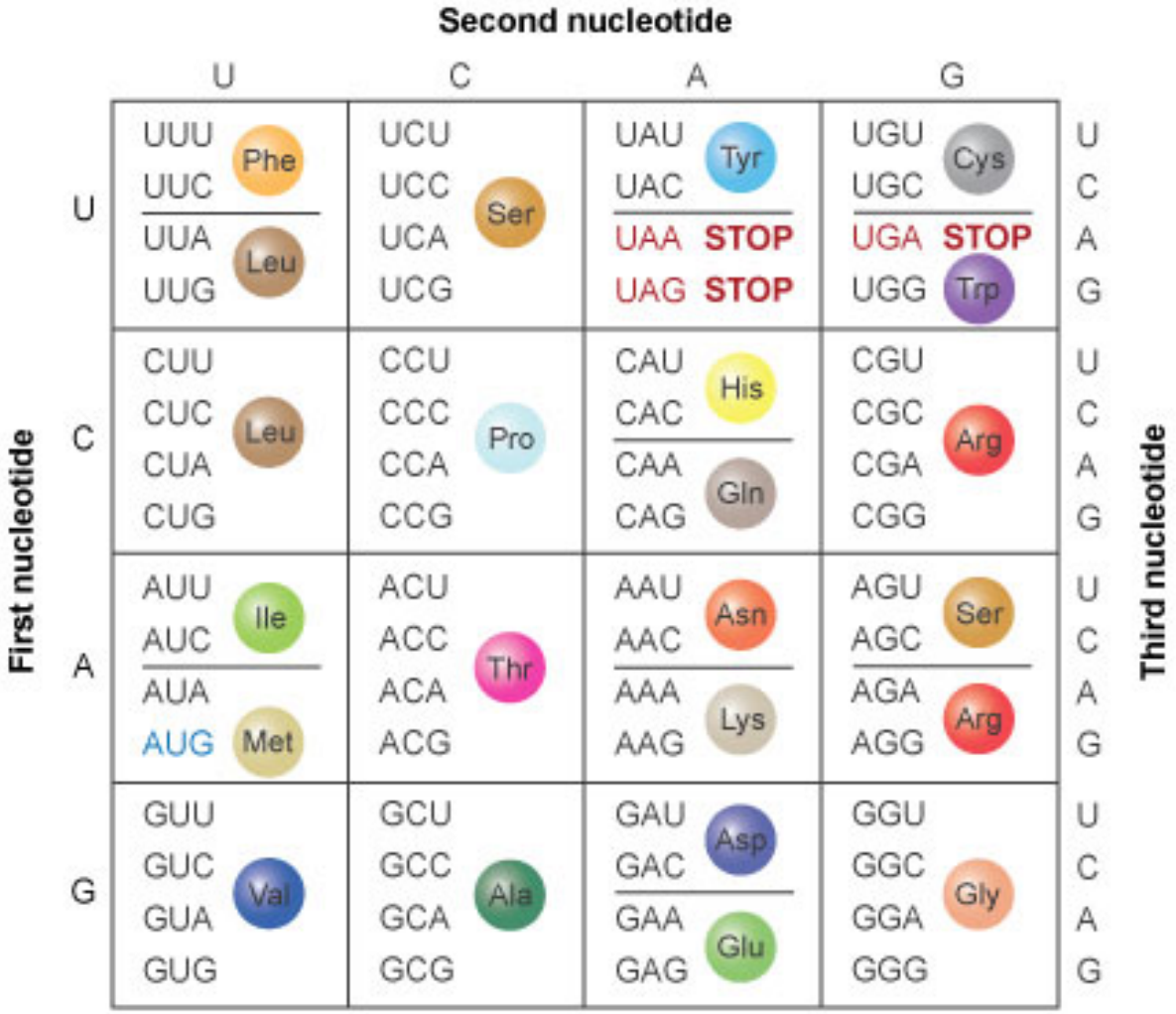


Central dogma

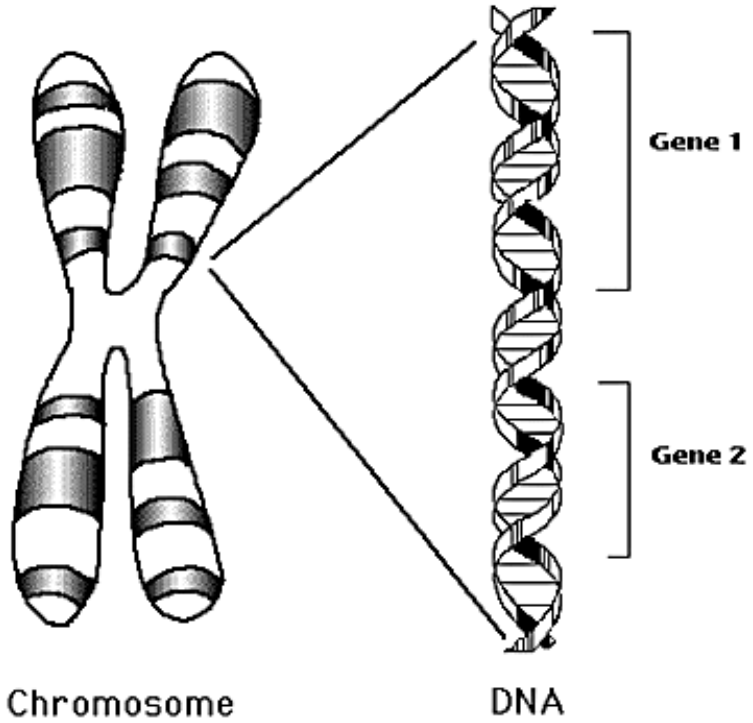


Codon

Codon: "triplet of nucleotide that code for a specific amino acid"



Genome



Human:

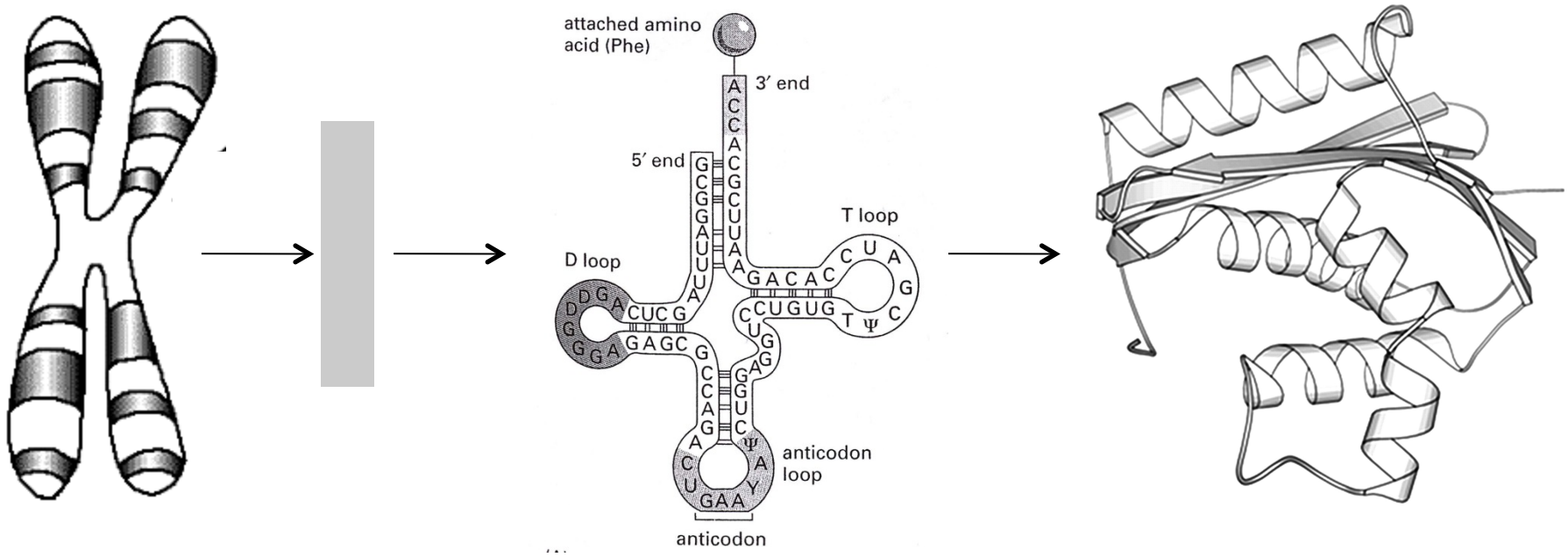
- 46 chromosomes
- With ca. 20-25K protein coding genes

Astonishing:

- 98% of the human genome is noncoding DNA → Only about 2% is coding. Expected: ca. 8-9% functional "somehow".
- But only 2% of typical bacterial genomes are noncoding

Gene expression

Gene expression

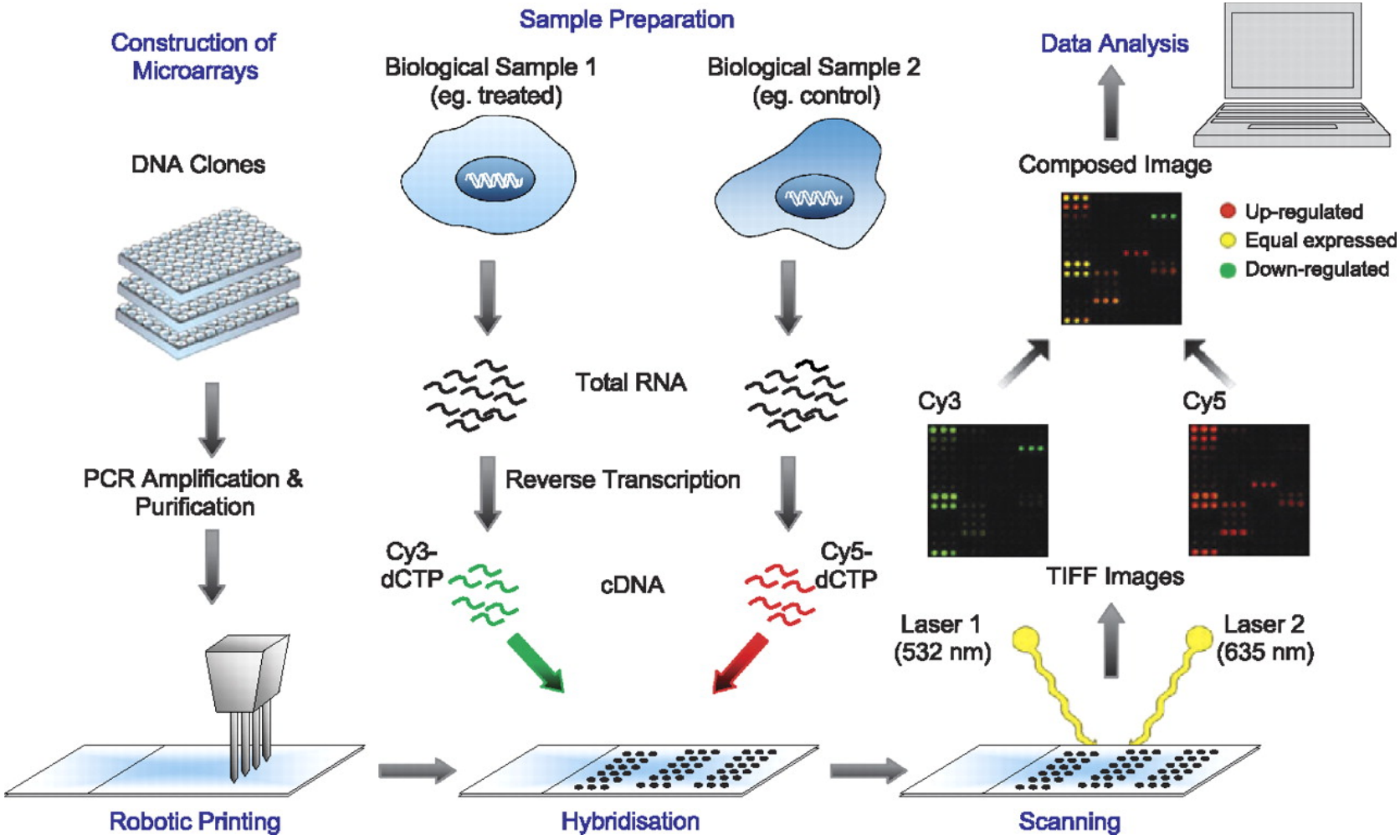


Chromosome **Gene** **mRNA** **Protein**

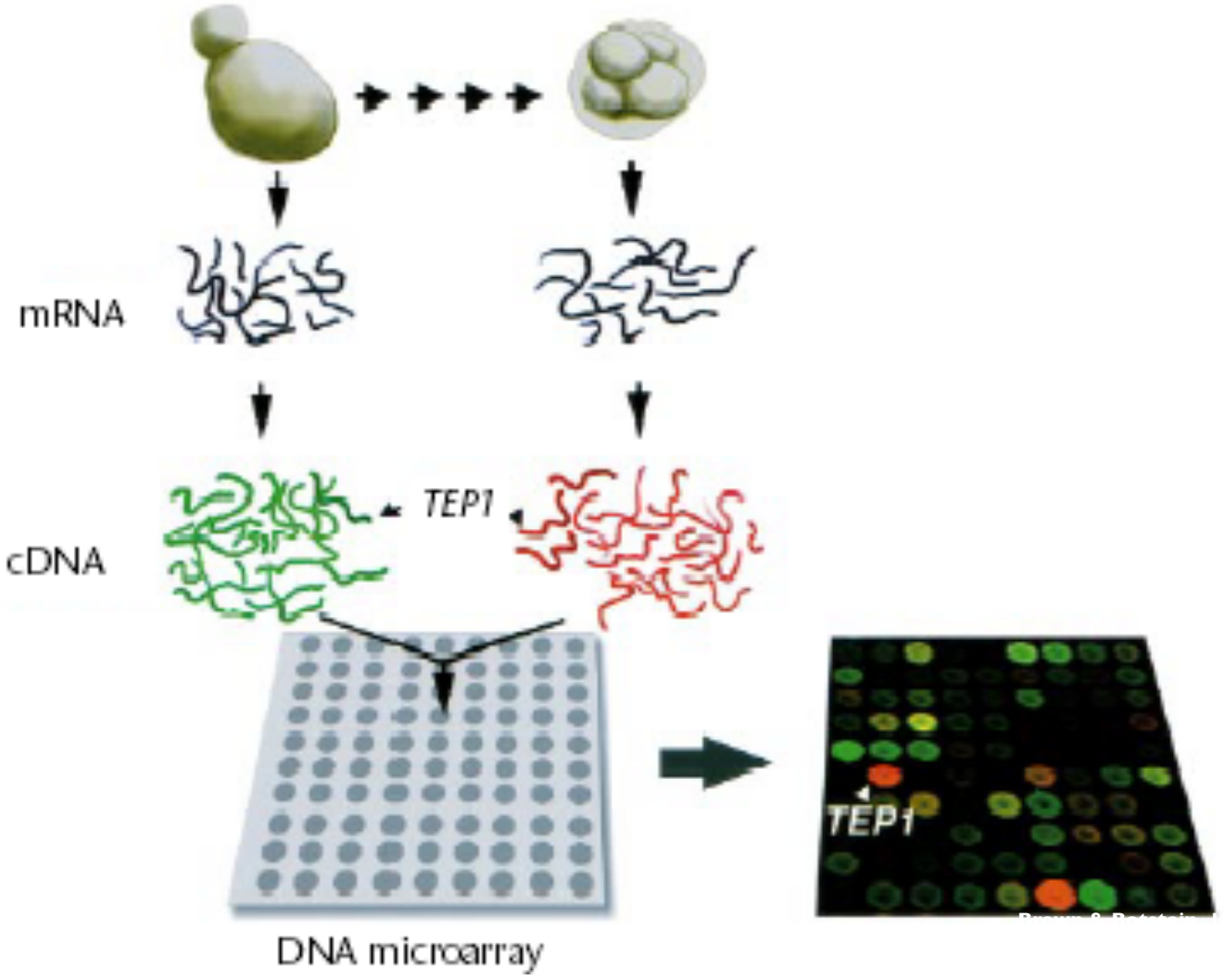
Gene: "segment of DNA that code for a specific protein"

Gene expression: "activation of a gene to synthesize a protein"

Gene expression array

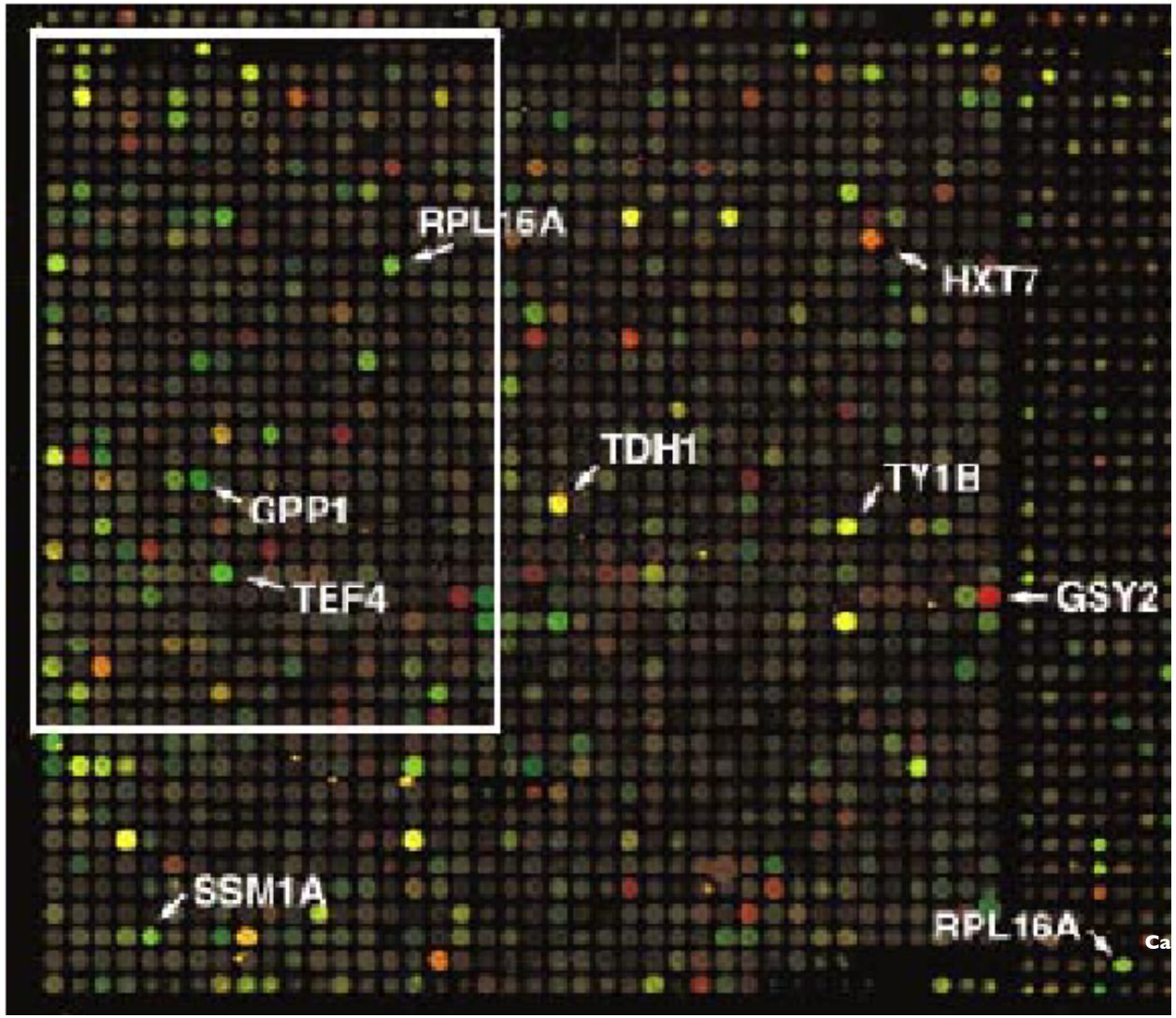


Overview



Borrowed from:
Patrick Schmid
CSE 497
Spring 2004

Real DNA Microarray



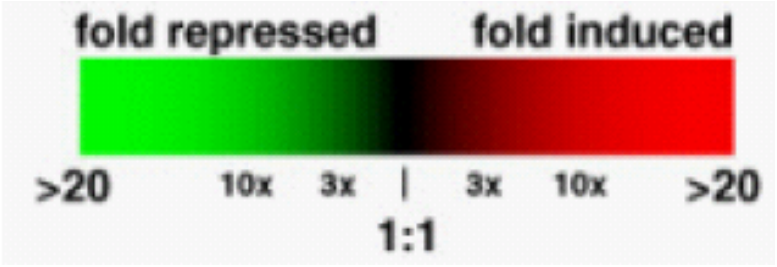
Borrowed from:
Patrick Schmid
CSE 497
Spring 2004

Color Coding

Data is presented with a color scale

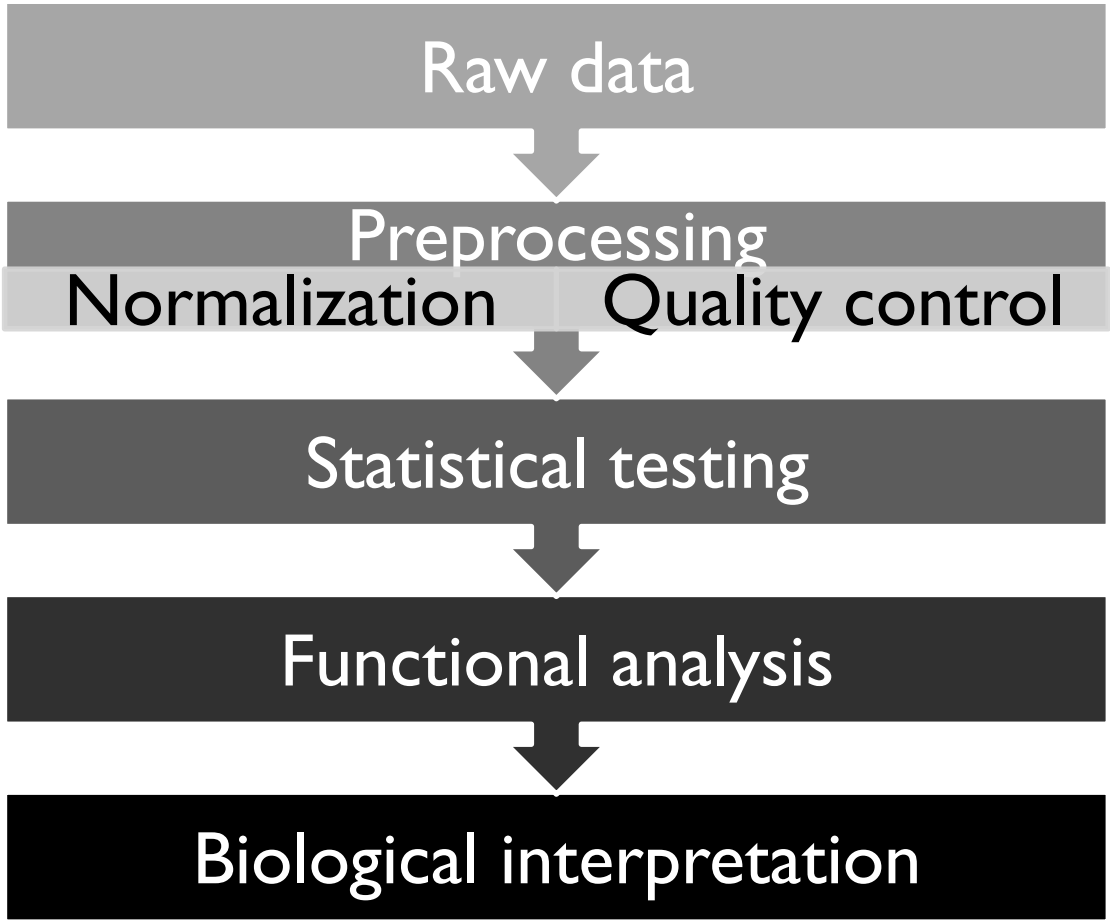
Coding scheme:

- Green = repressed (less mRNA) gene in experiment
- Red = induced (more mRNA) gene in experiment
- Black = no change (1:1 ratio)

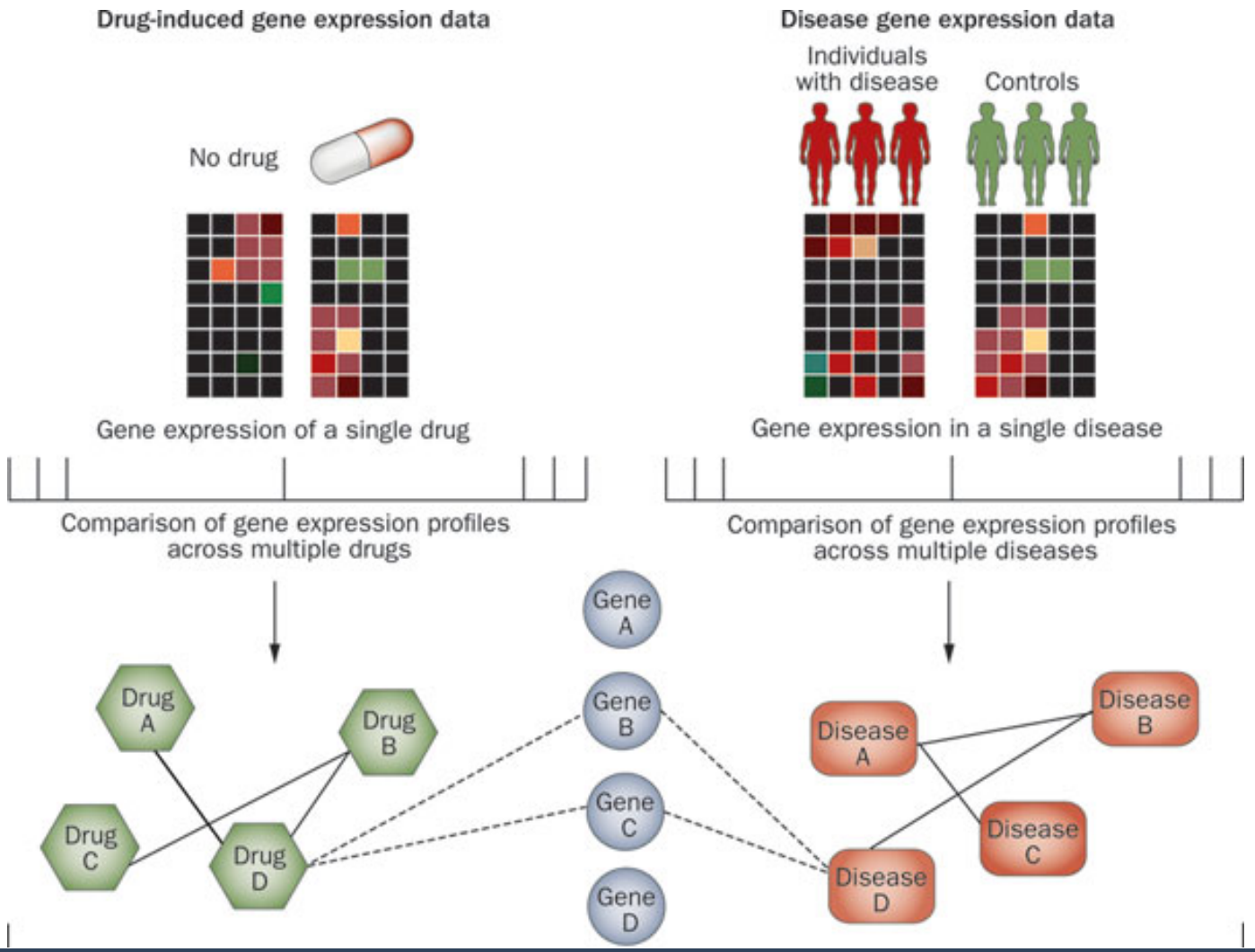


Borrowed from:
Patrick Schmid
CSE 497
Spring 2004

Gene expression analysis

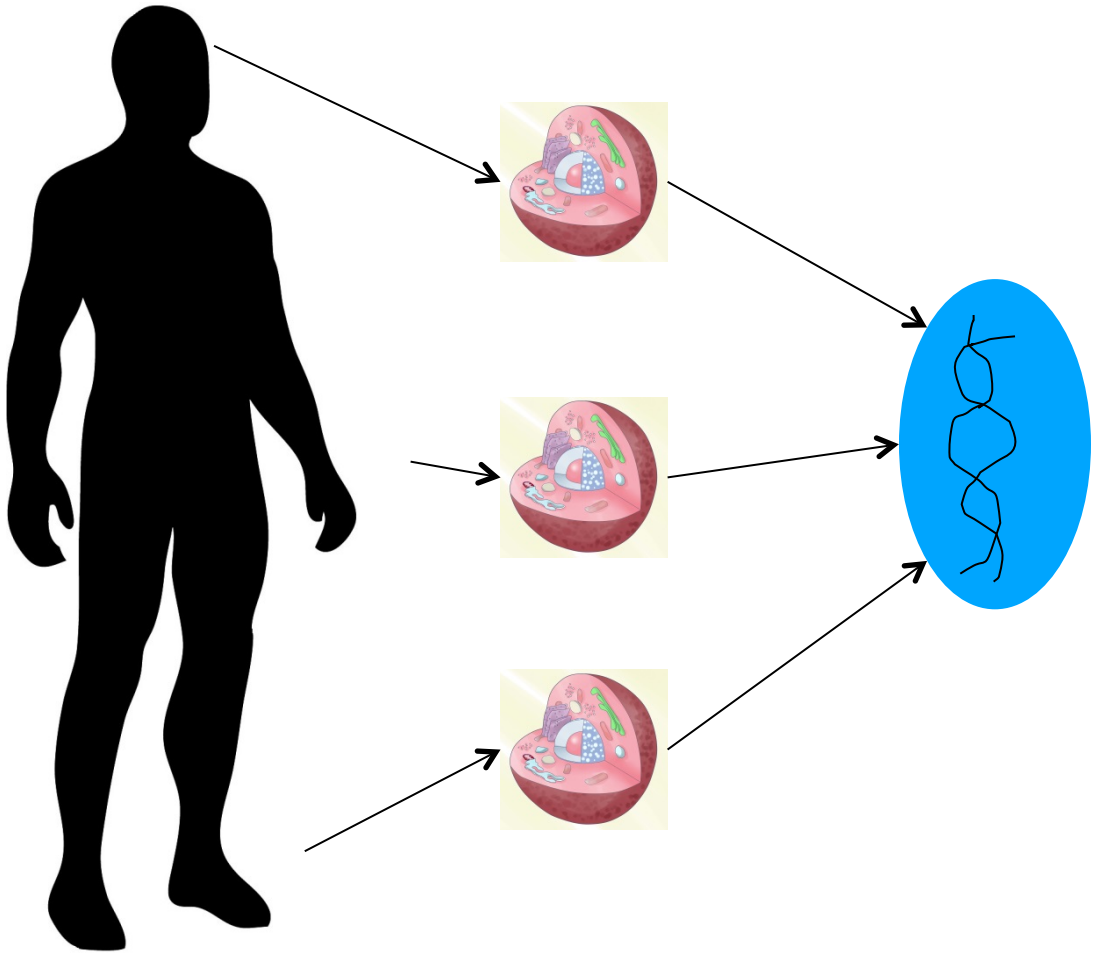


Comparative analysis



Gene regulation

Gene regulation makes a difference



Same DNA
Yet
Different functions

↓

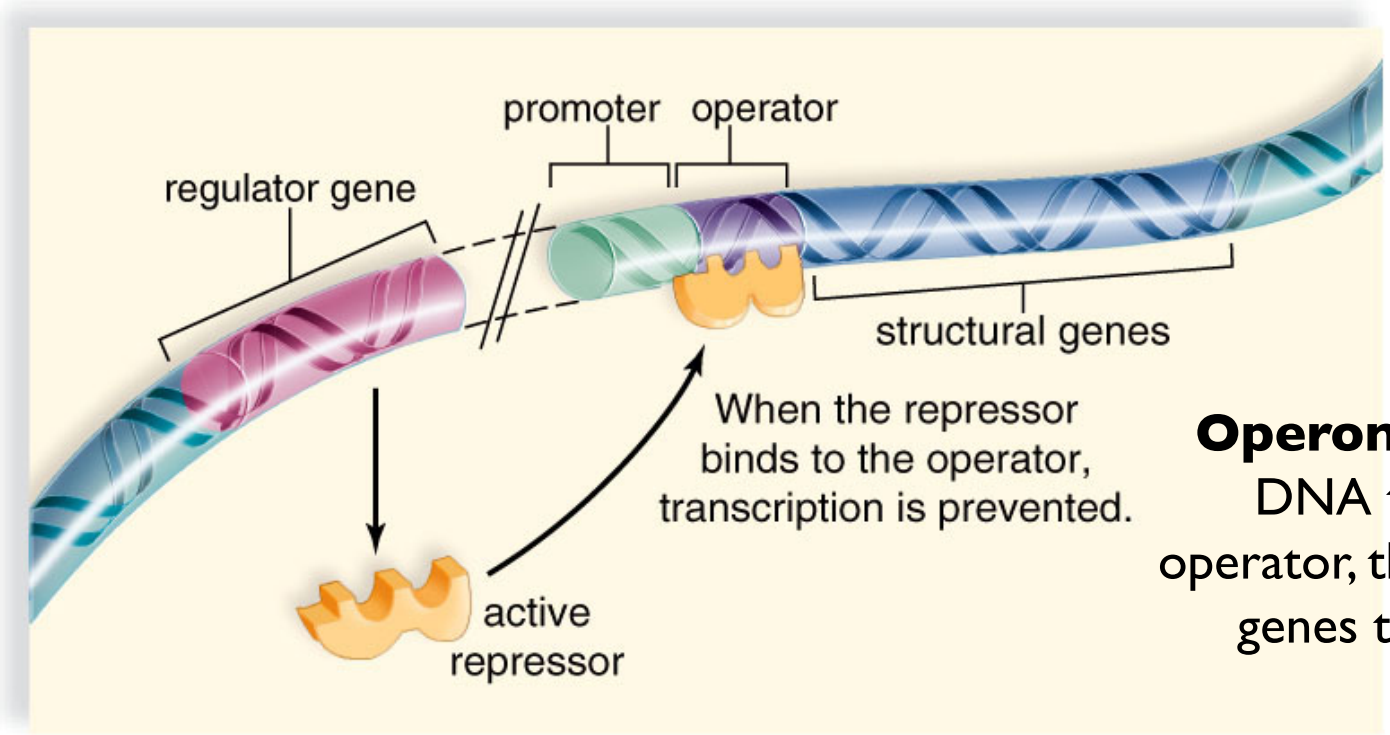
Differences in gene expression

↓

Gene regulation

Prokaryotes

Elements of gene in bacteria



Operon: "entire stretch of DNA that includes the operator, the promoter, and the genes that they control"

promoter: "site of transcriptional activation"

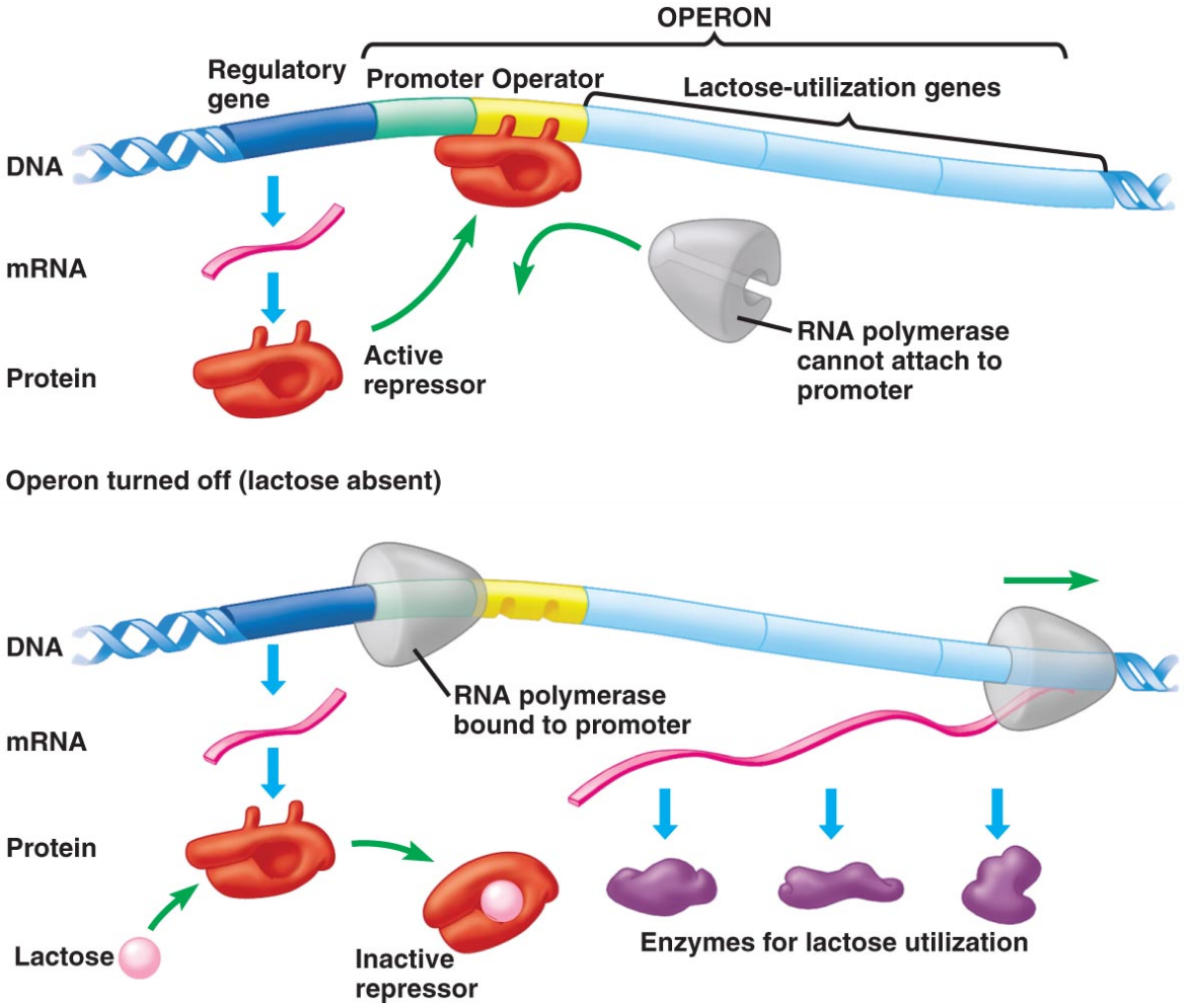
operator: "repressor binding site"

Operon model of regulation

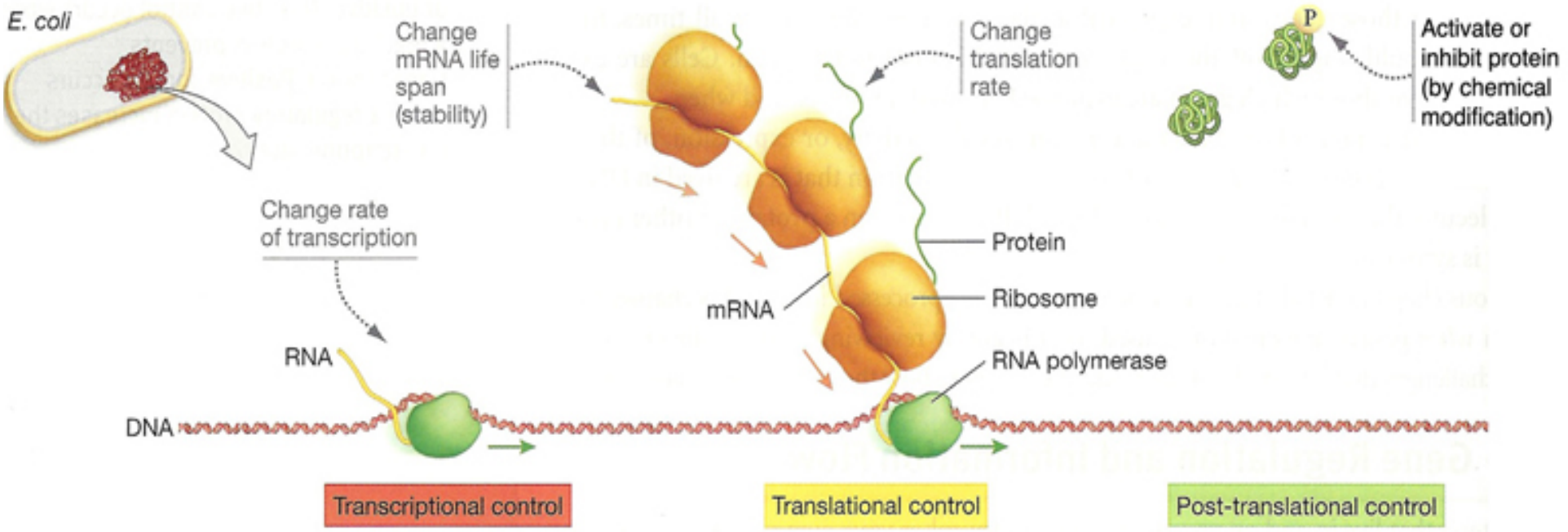
Inducible operon: "operon is activated by small molecule called inducers" e.g. *lac* operon is induced by allolactose

Repressible operon: "operon is shut-off by small-molecule called co-repressors" e.g. *trp* operon is repressed by tryptophan

Lac operon

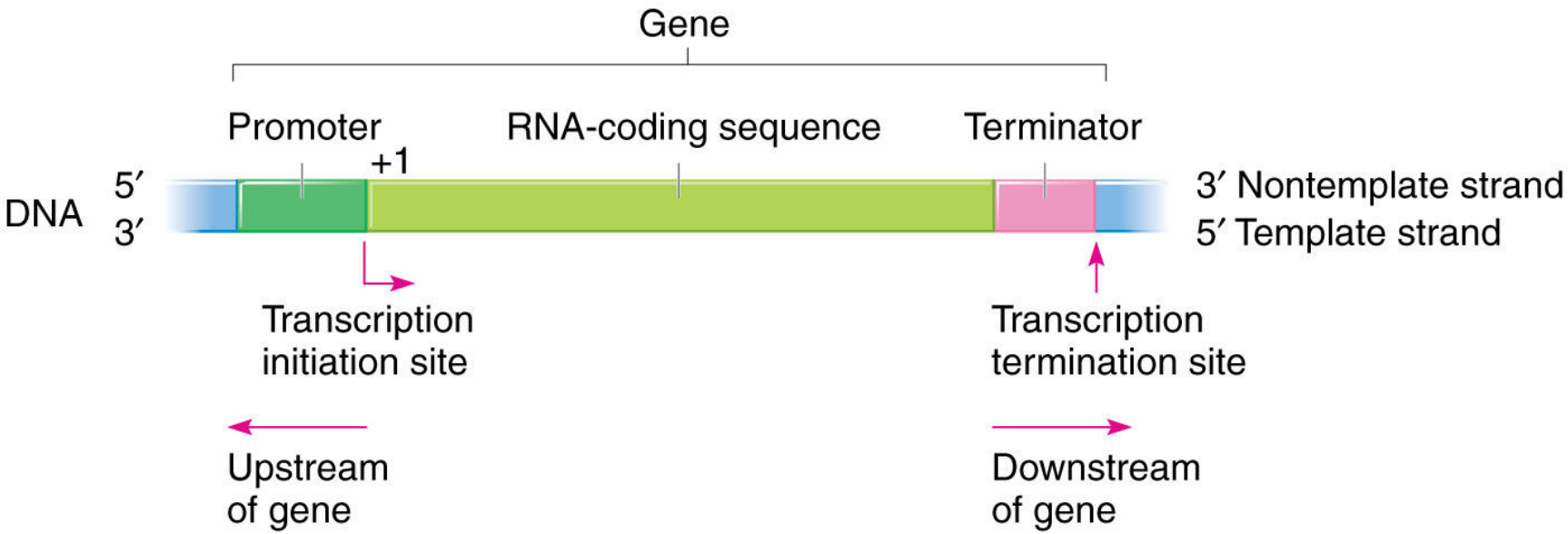


Levels of gene regulation in bacteria

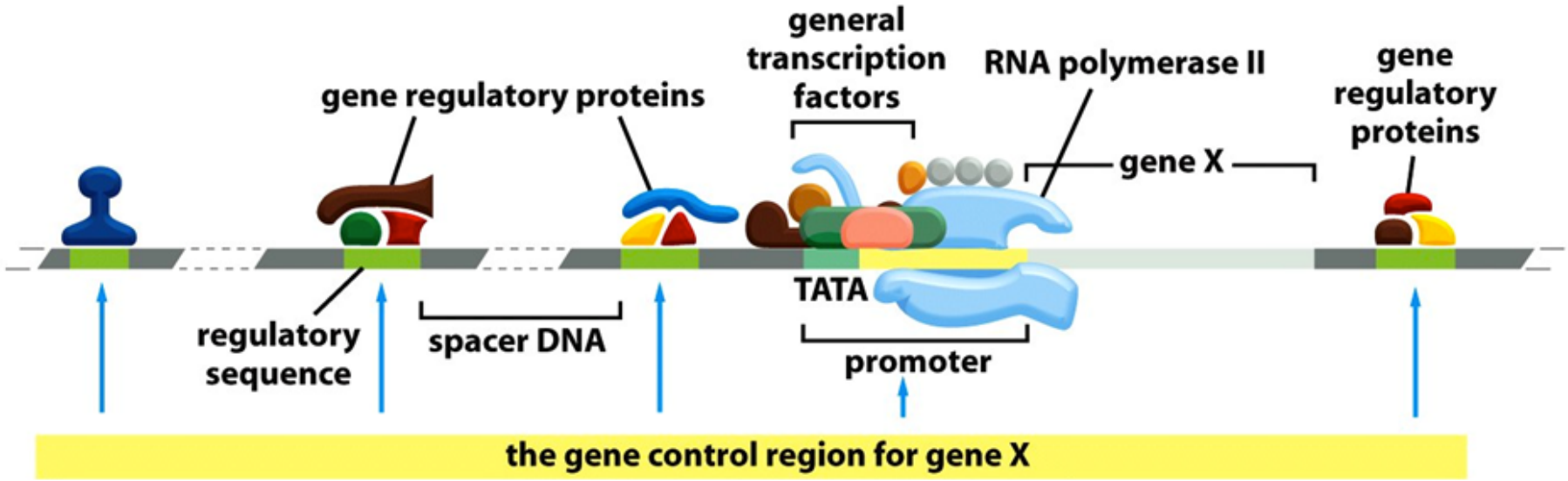


Eukaryotes

Elements of gene

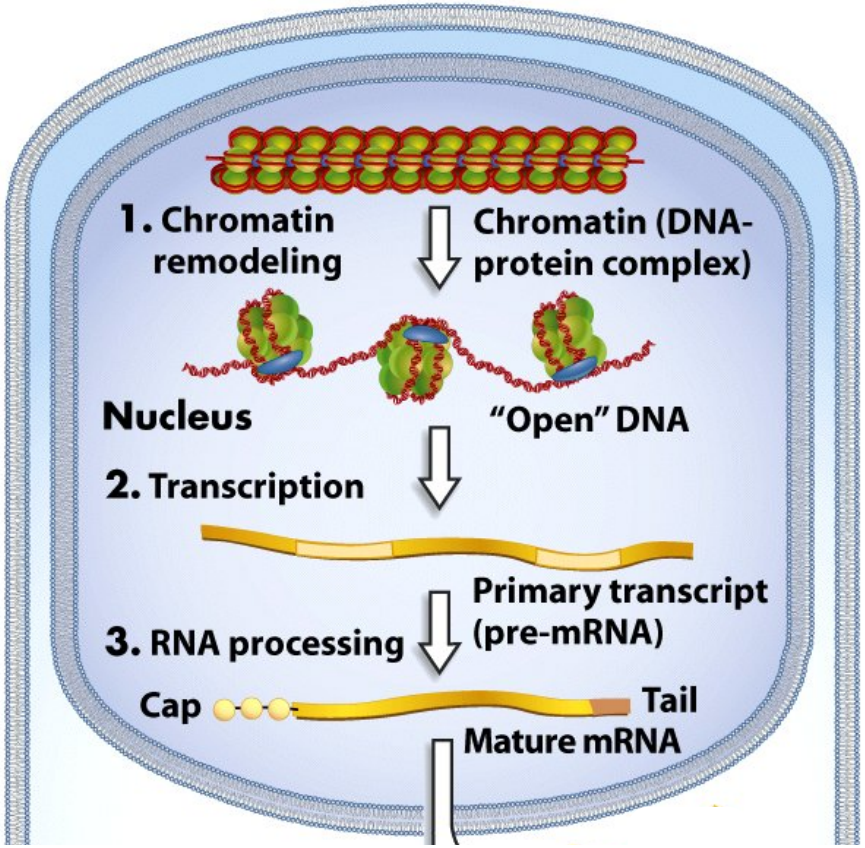


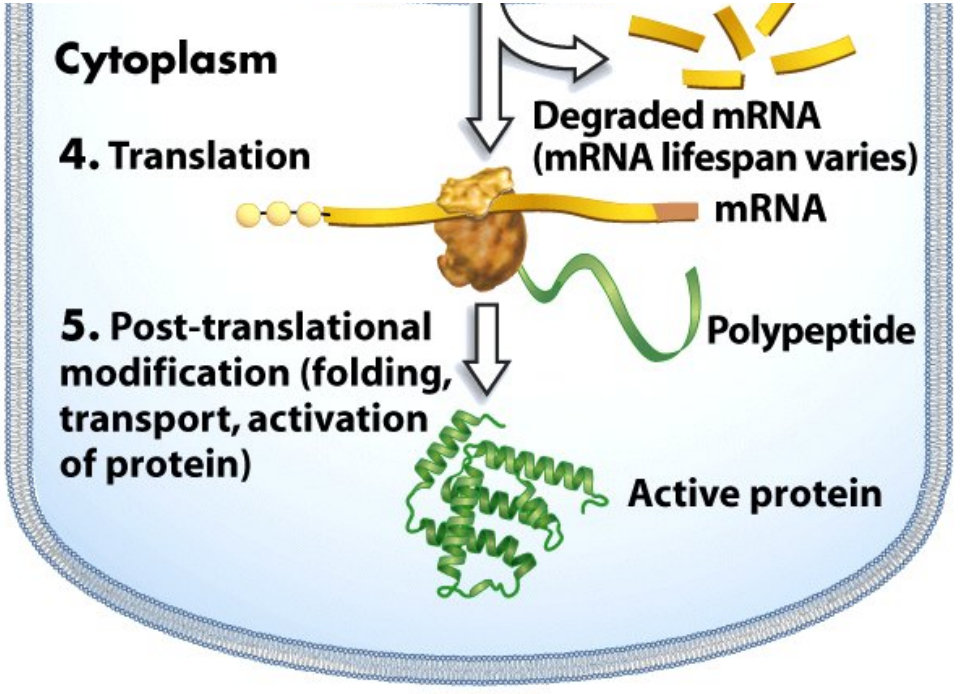
Gene regulation



Gene regulation: "the process of turning on/ off of the genes"

Levels of Gene regulation





Transcriptional control

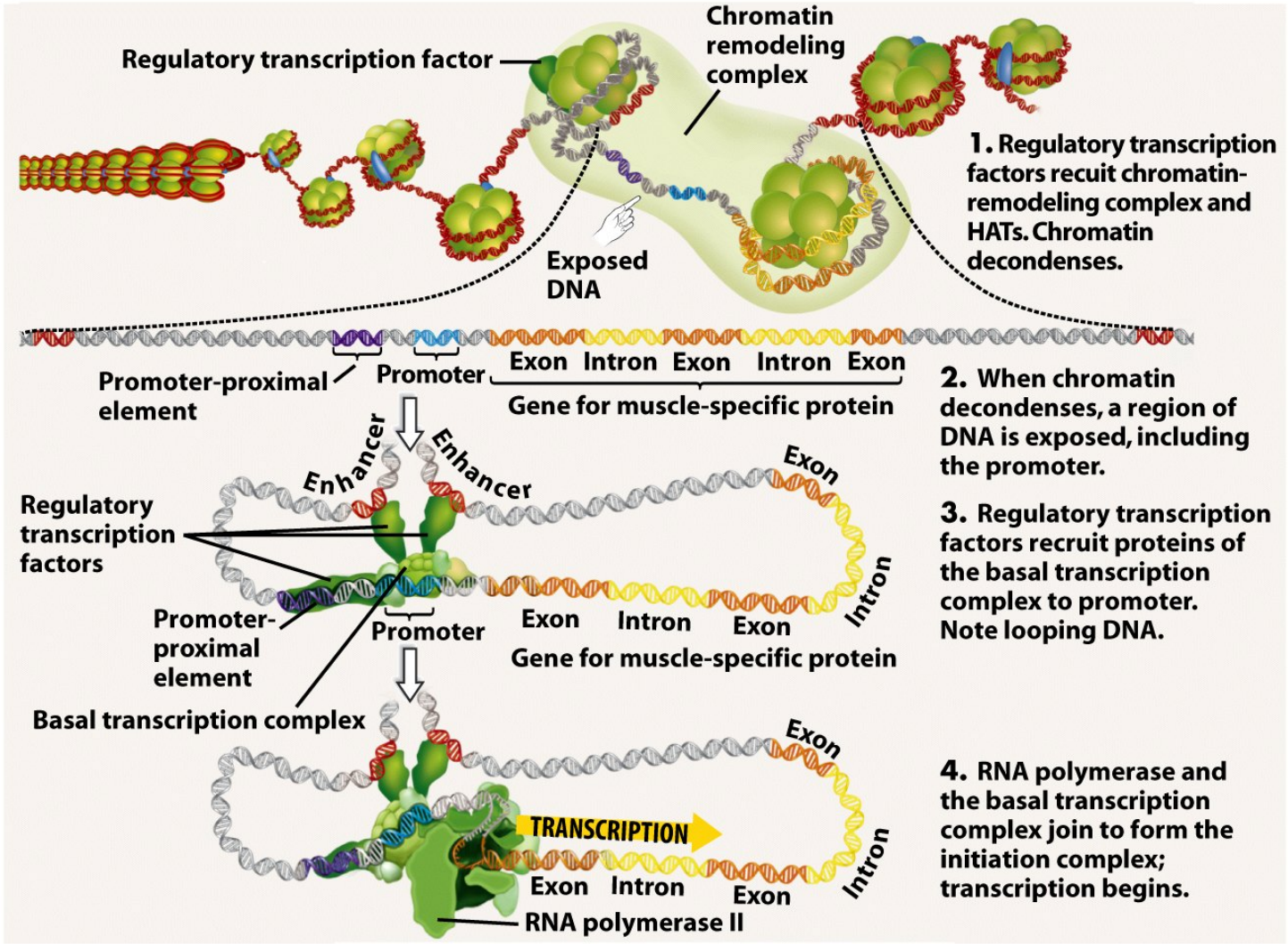
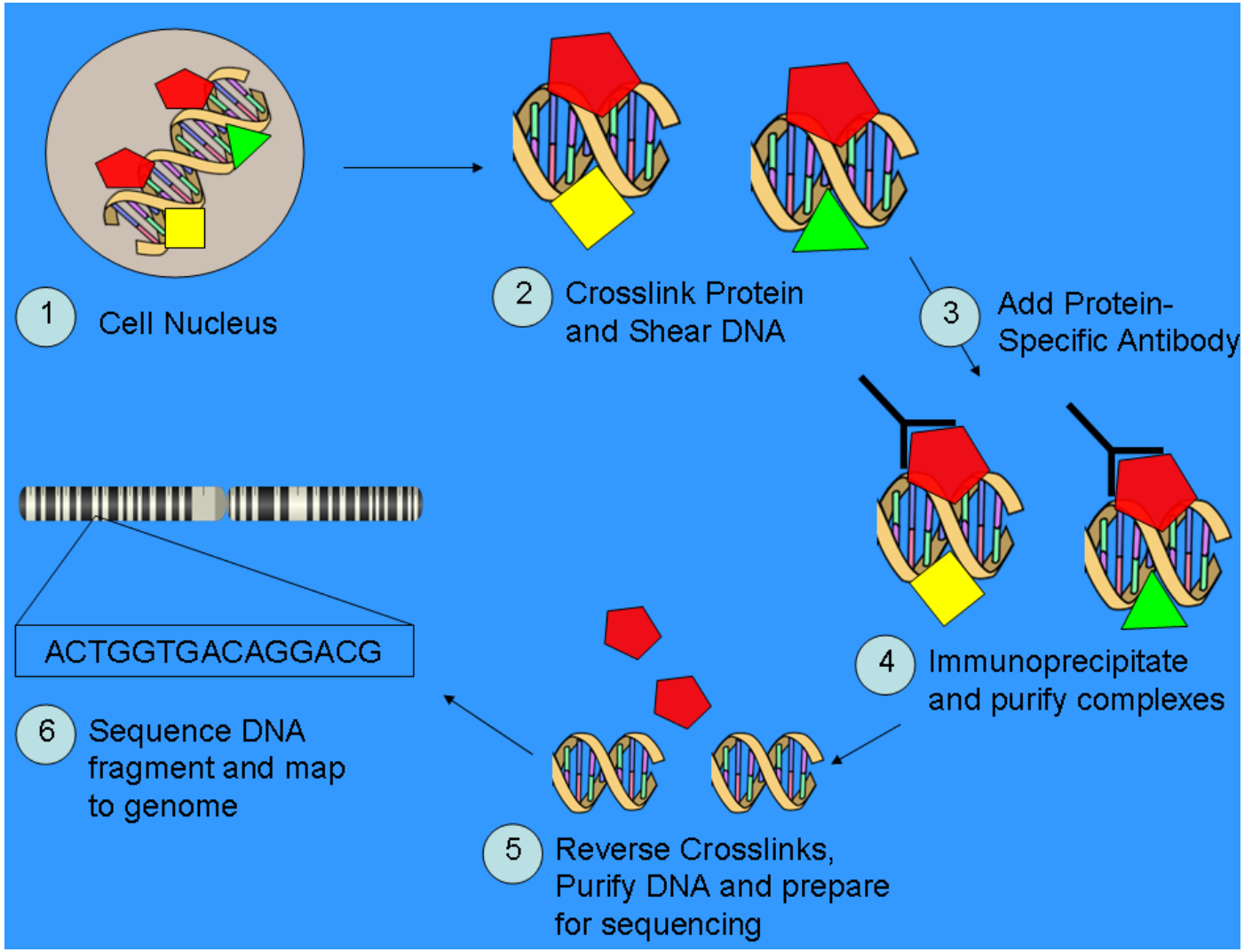


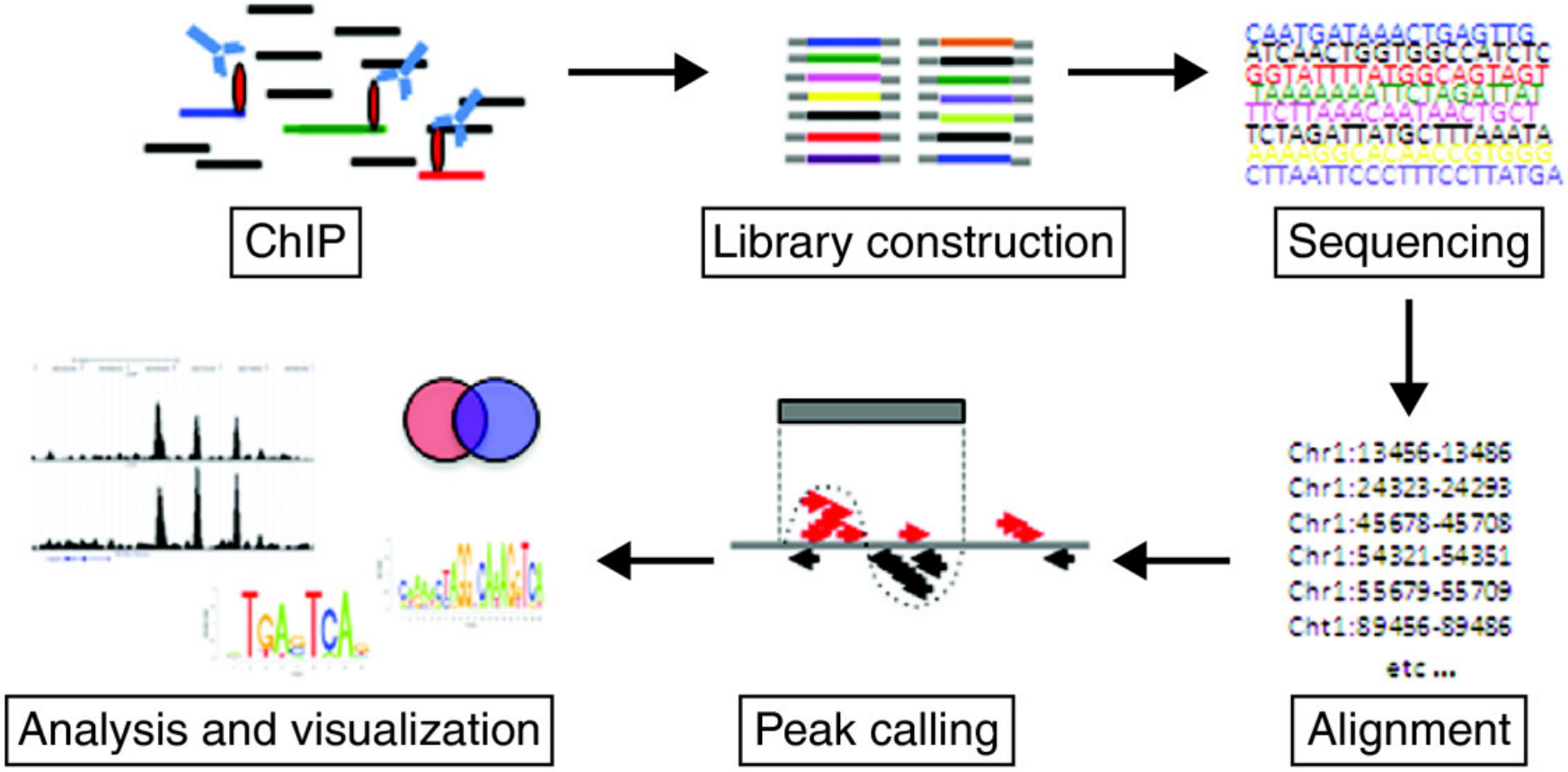
Figure 18-10 Biological Science, 2/e

CHIP-Seq

ChIP-seq:
"sequencing of the genomic DNA fragments that co-precipitate with a DNA-binding protein that is under study"



Flow scheme of ChIP-seq procedure.



Prediction of TFBS

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

1	2	3	4	5	6	7	8	9	10	11	12	13	14
B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

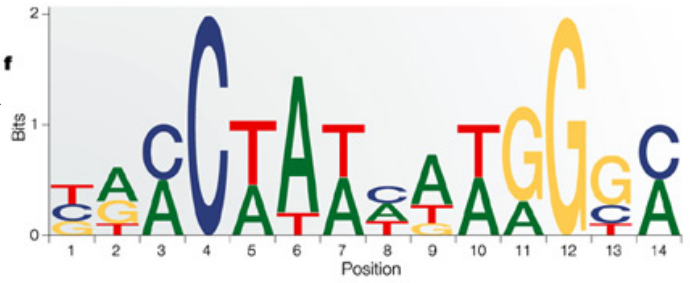
d Position weight matrix (PWM)

A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

e Site scoring

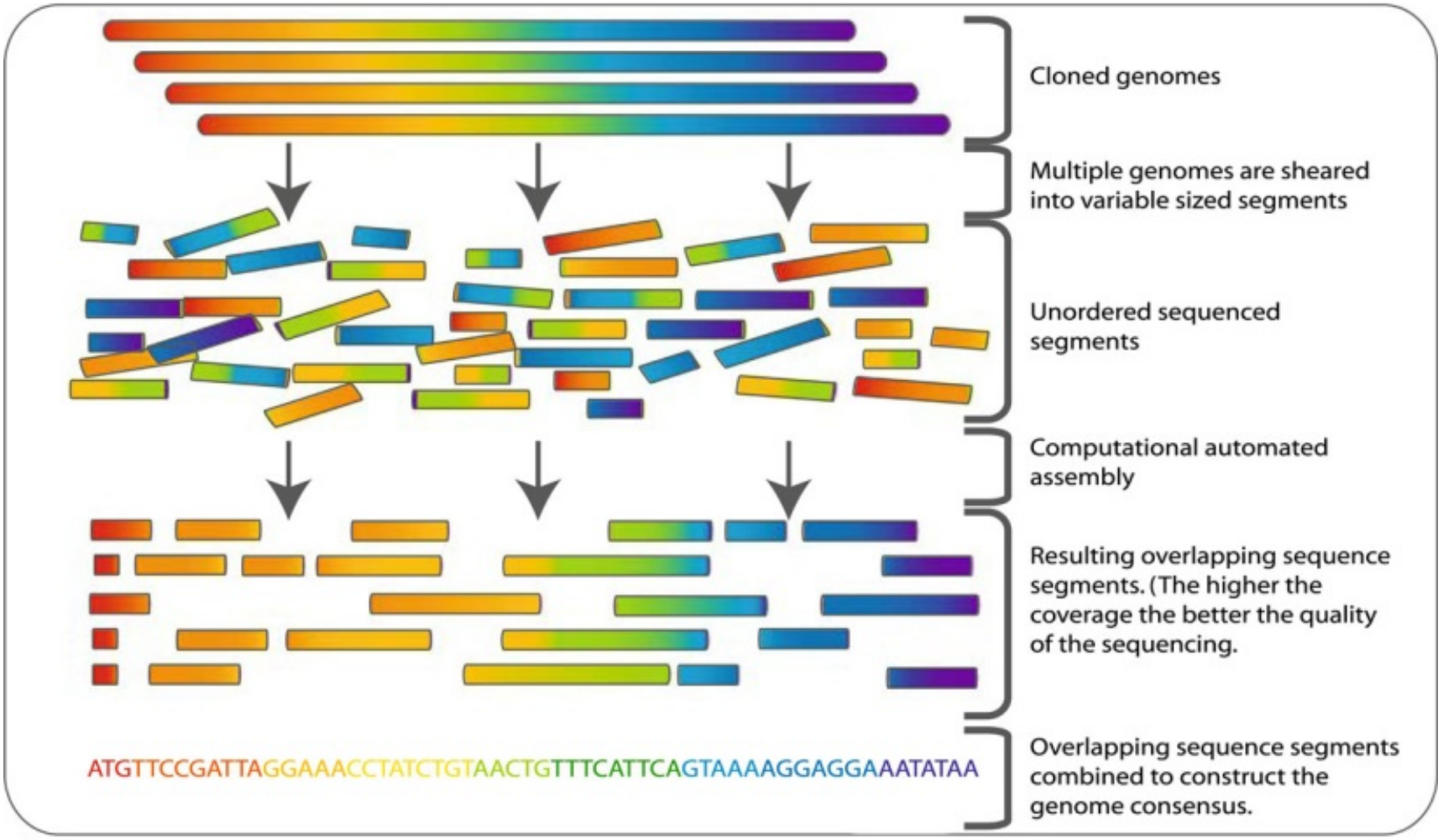
0.45	−0.66	0.79	1.68	0.45	−0.66	0.79	0.45	−0.66	0.79	0.00	1.68	−0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 78% of maximum

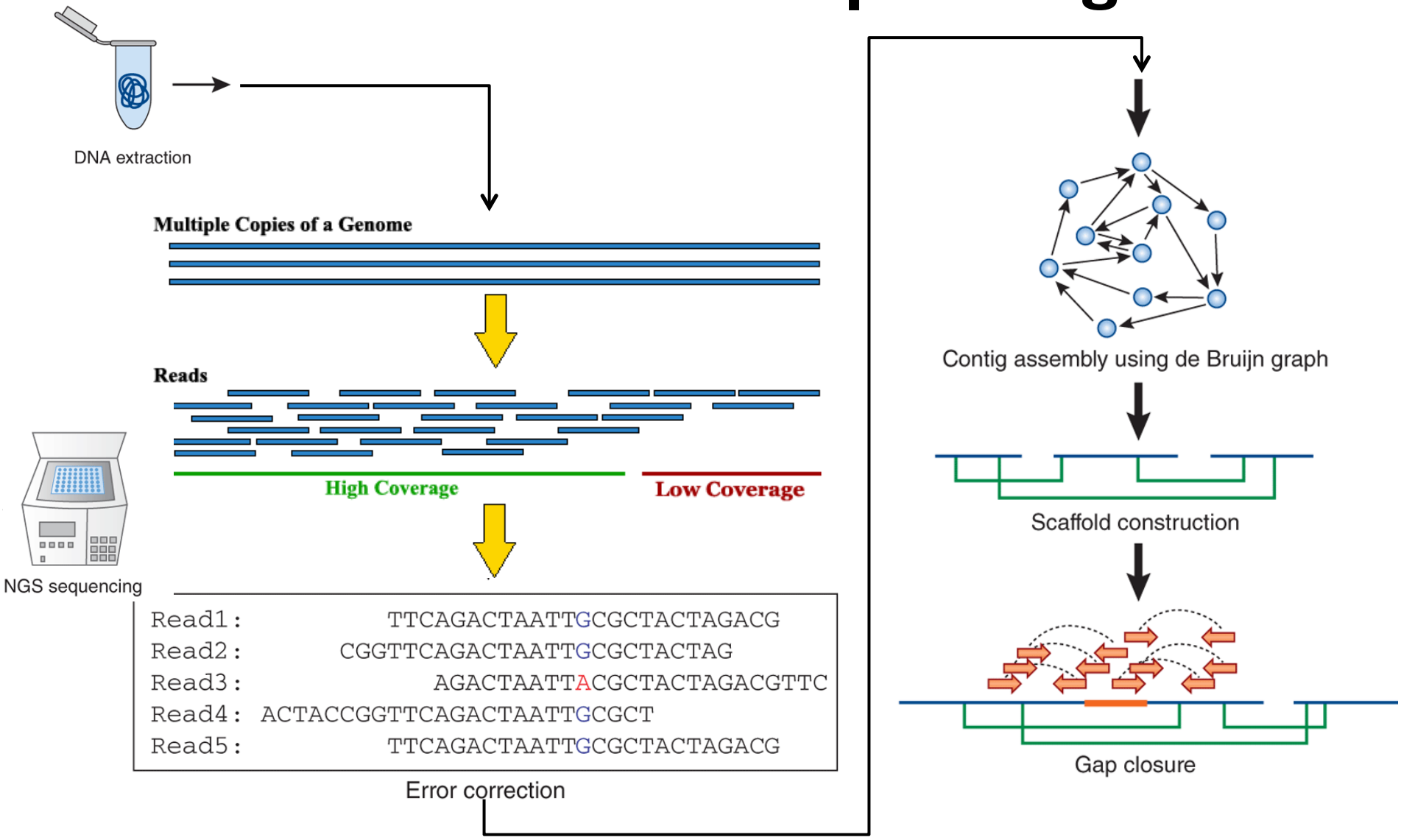


DNA sequencing

Principle of sequencing



Flow chart of sequencing



Genome length

Organism	Estimated size (base pairs)	Chromosome number	Estimated gene number
Human	3 billion	46	~ 25,000
Yeast	12 million	32	6,000
E.coli	4.6 million	1	3,200

Human genome	
Avg. size of chromosomes	~ 130 Mega bps
Genes per chromosome	~ 3000
Avg. size of gene (base pairs)	1148 to 37.7 Kilo bp

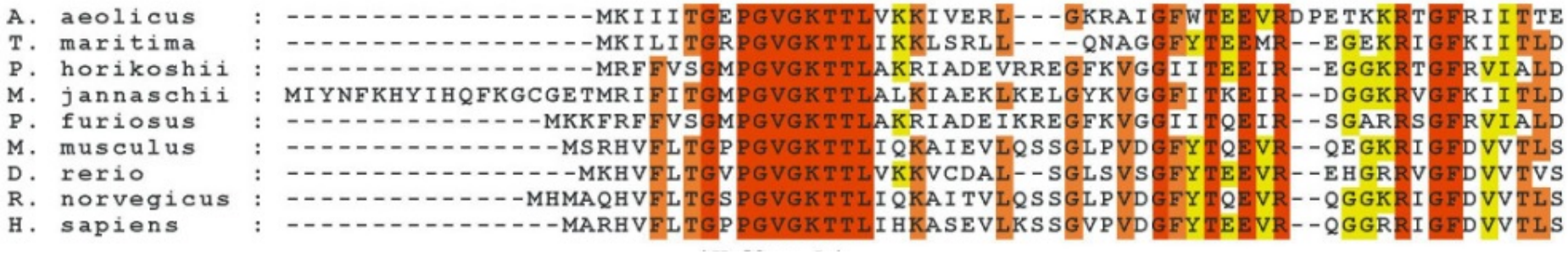
Comparison of next-generation sequencing platforms

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

Sequence alignment

Sequence alignment



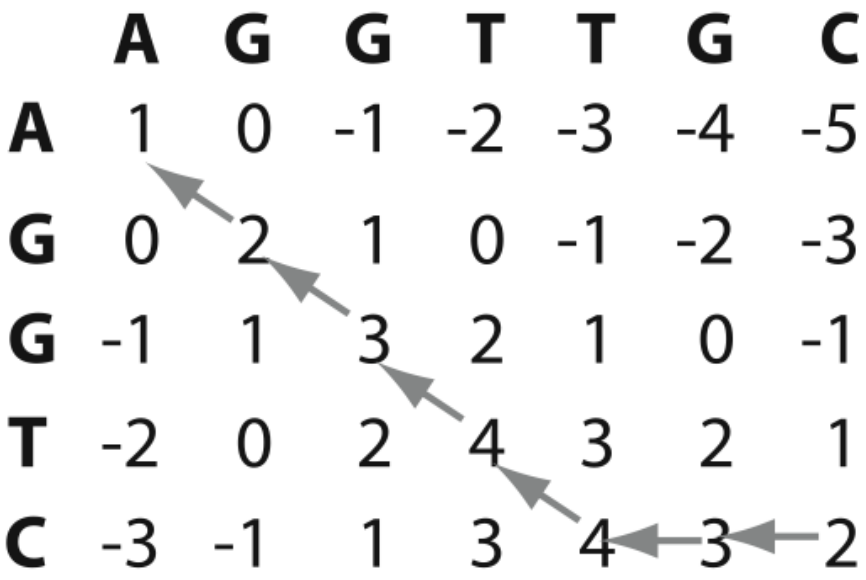
Pairwise sequence alignment: "exactly two nucleotide or protein sequences are aligned to each other to determine the similarity between the two sequences"

Multiple sequence alignment: "process of aligning three or more nucleotide or protein sequences to identify similarities between the sequences"

Local alignment: "will align only similar regions between sequences, and leave regions with too many differences unaligned"

Global alignment: "attempt to align every base (or amino acid) in each aligned sequence"

Principle of sequence alignment



		A	G	G	T	T	G	C
		1	0	-1	-2	-3	-4	-5
A	0	2	1	0	-1	-2	-3
G	-1	1	3	2	1	0	-1
G	-2	0	2	4	3	2	1
T	-3	-1	1	3	4	3	2
C							

Gap: "space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another"

Gap penalty: "when a sequence alignment is scored, gaps in the alignment are given a negative score or gap penalty."

A	G	G	T	T	G	C
A	G	G	T	-	-	C

BLAST

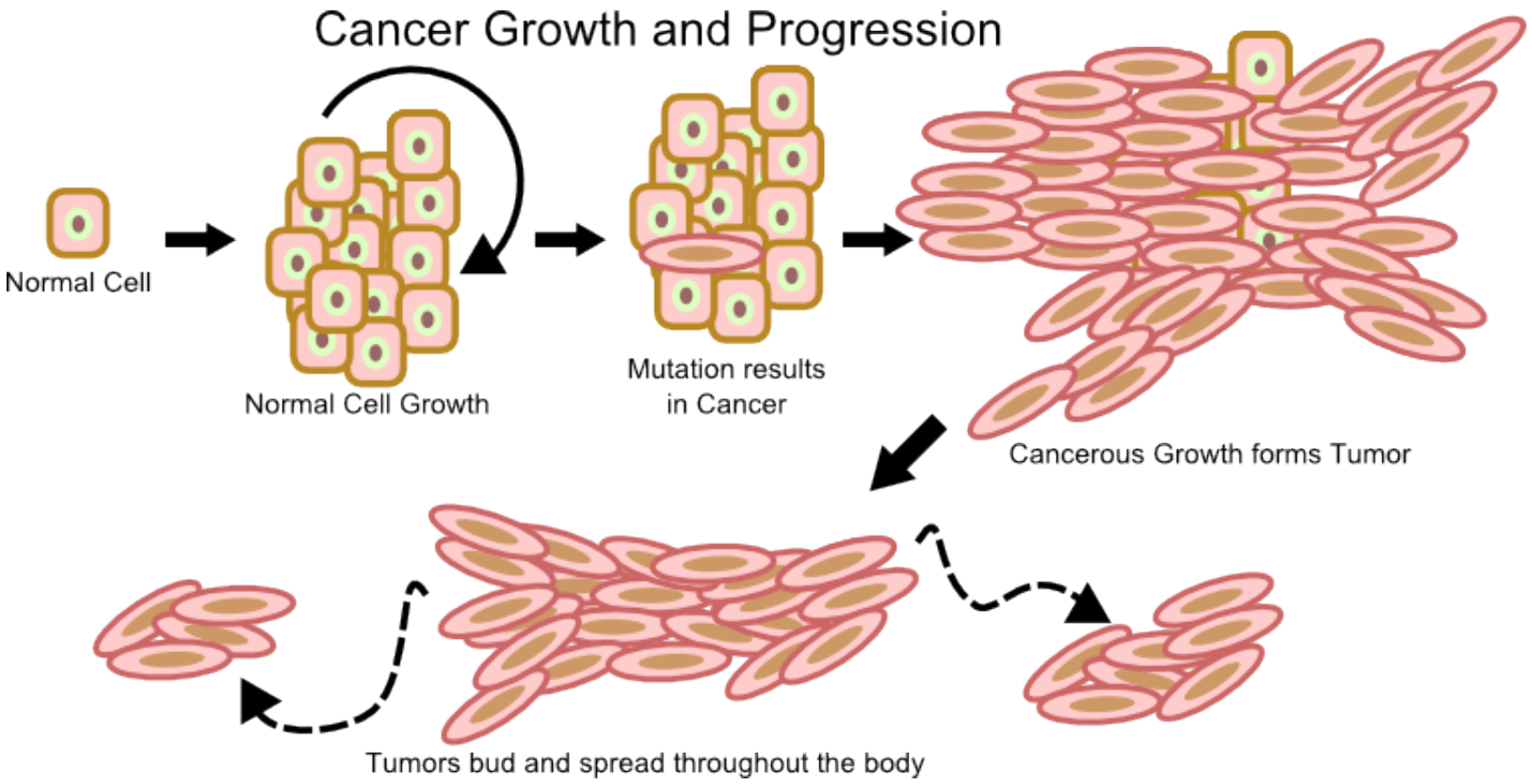
Basic Local Alignment Search Tool

- It is a widely used sequence comparison tool.
- The algorithm is optimized for speed used to search sequence databases for optimal local alignments to a query.
- The alignments are assigned an significance score i.e. E-value

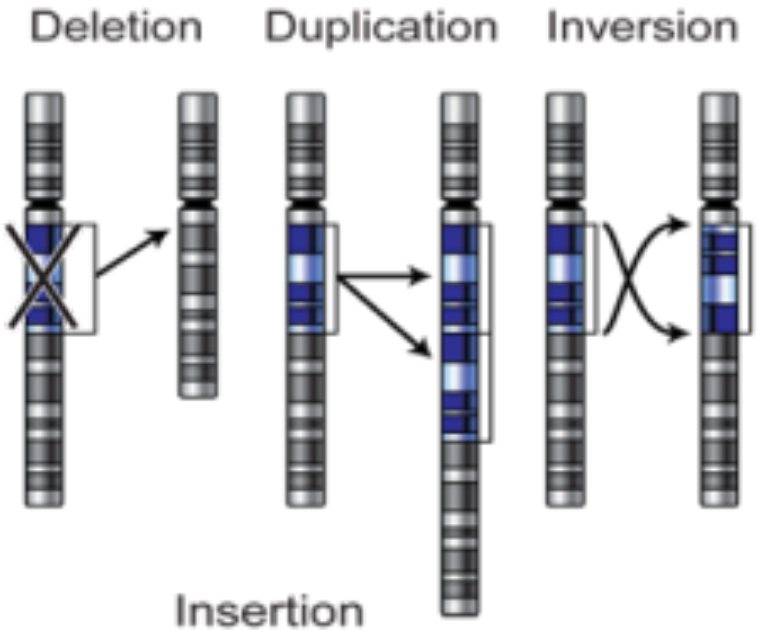
E-value: The expectation value is the number of different alignments with scores equivalent to or better than a score S that one can expect to occur in a database search by chance. The lower the E-value, the more significant the score S and, thus, the alignment.

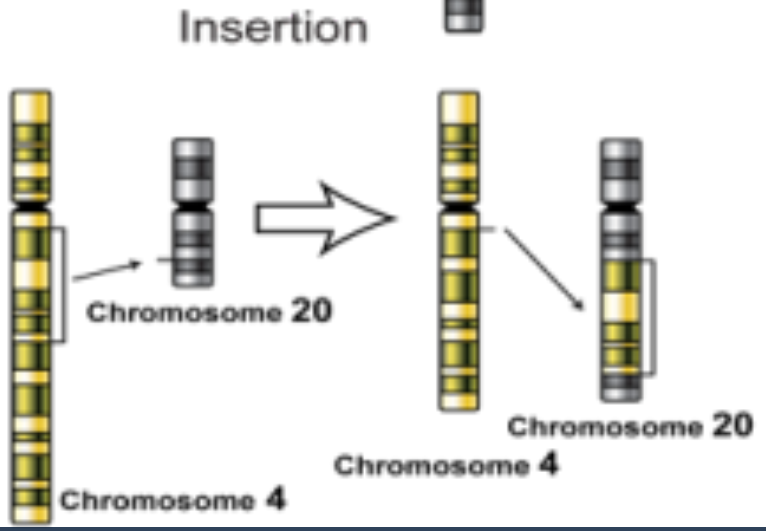
Cancer

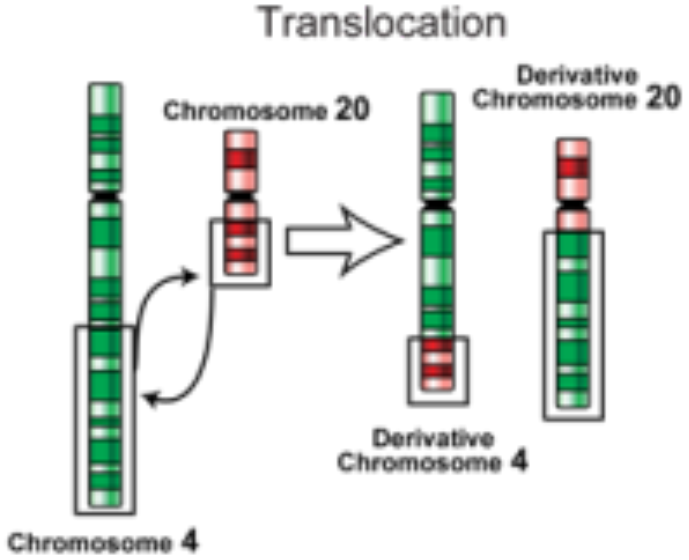
Cancer


Cancer Growth and Progression
Normal Cell
Normal Cell Growth
Mutation results in Cancer
Cancerous Growth forms Tumor
Tumors bud and spread throughout the body

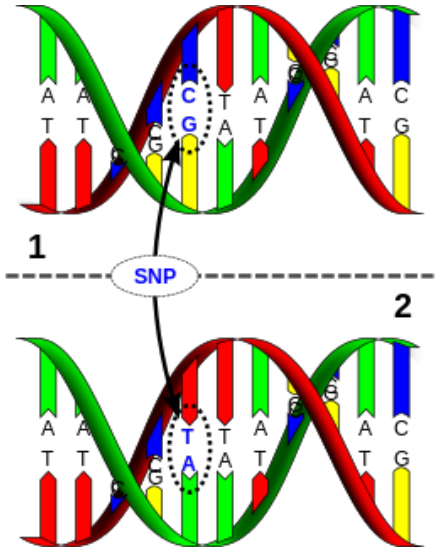
Mutations: Genome is NOT static!



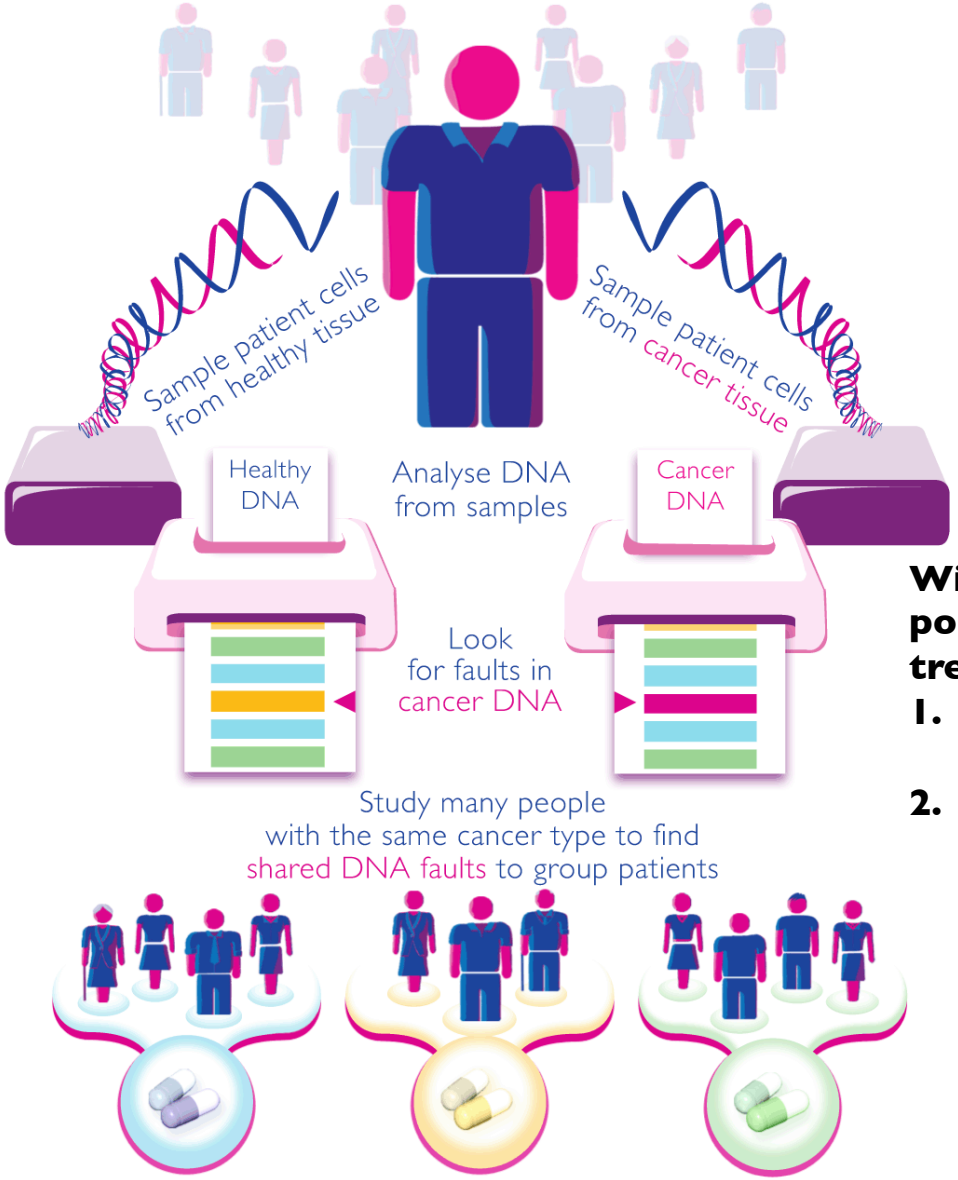




SNP



The International Cancer Genome Consortium

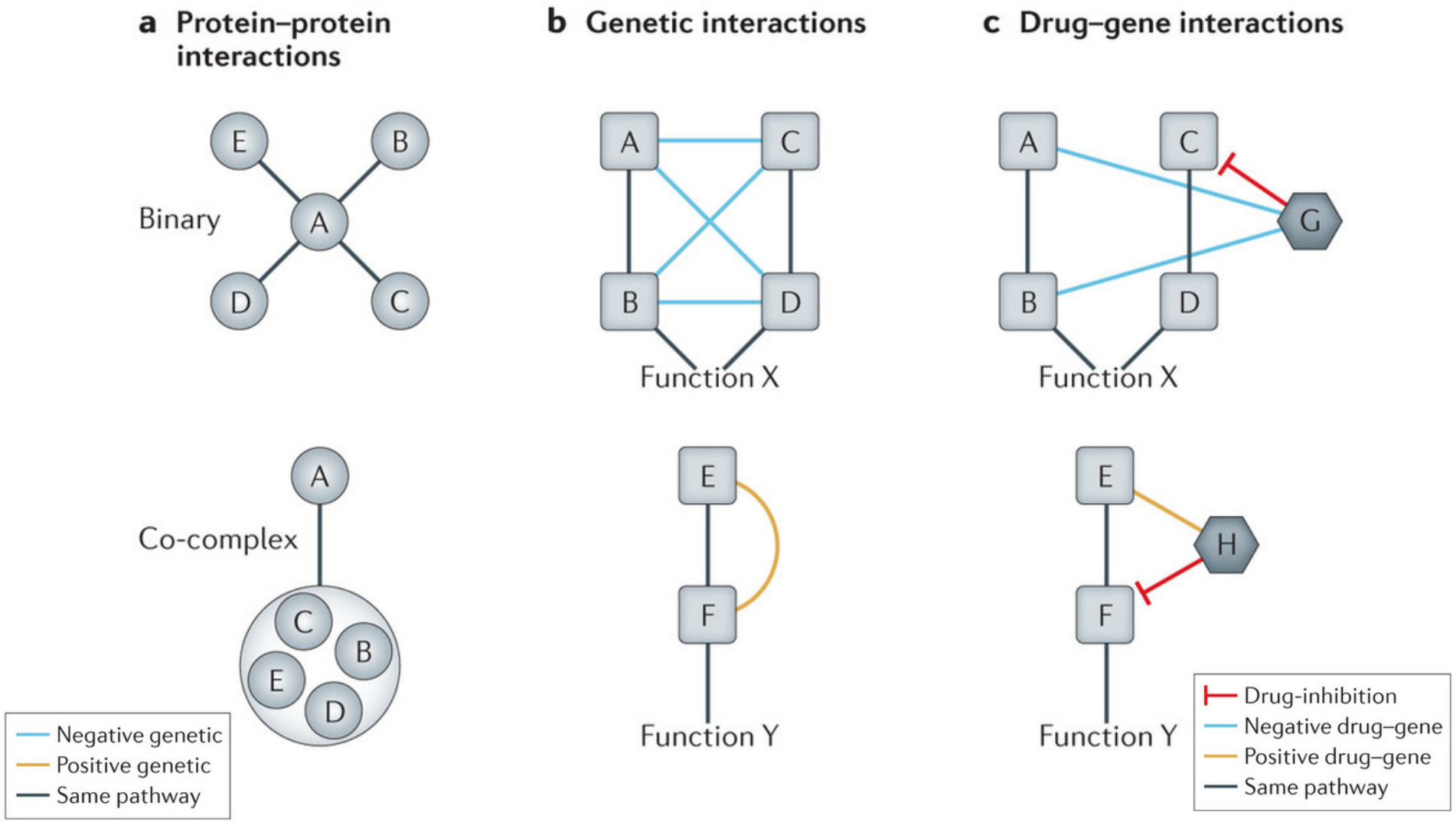


Within a decade, it will be possible to better tailor treatment :

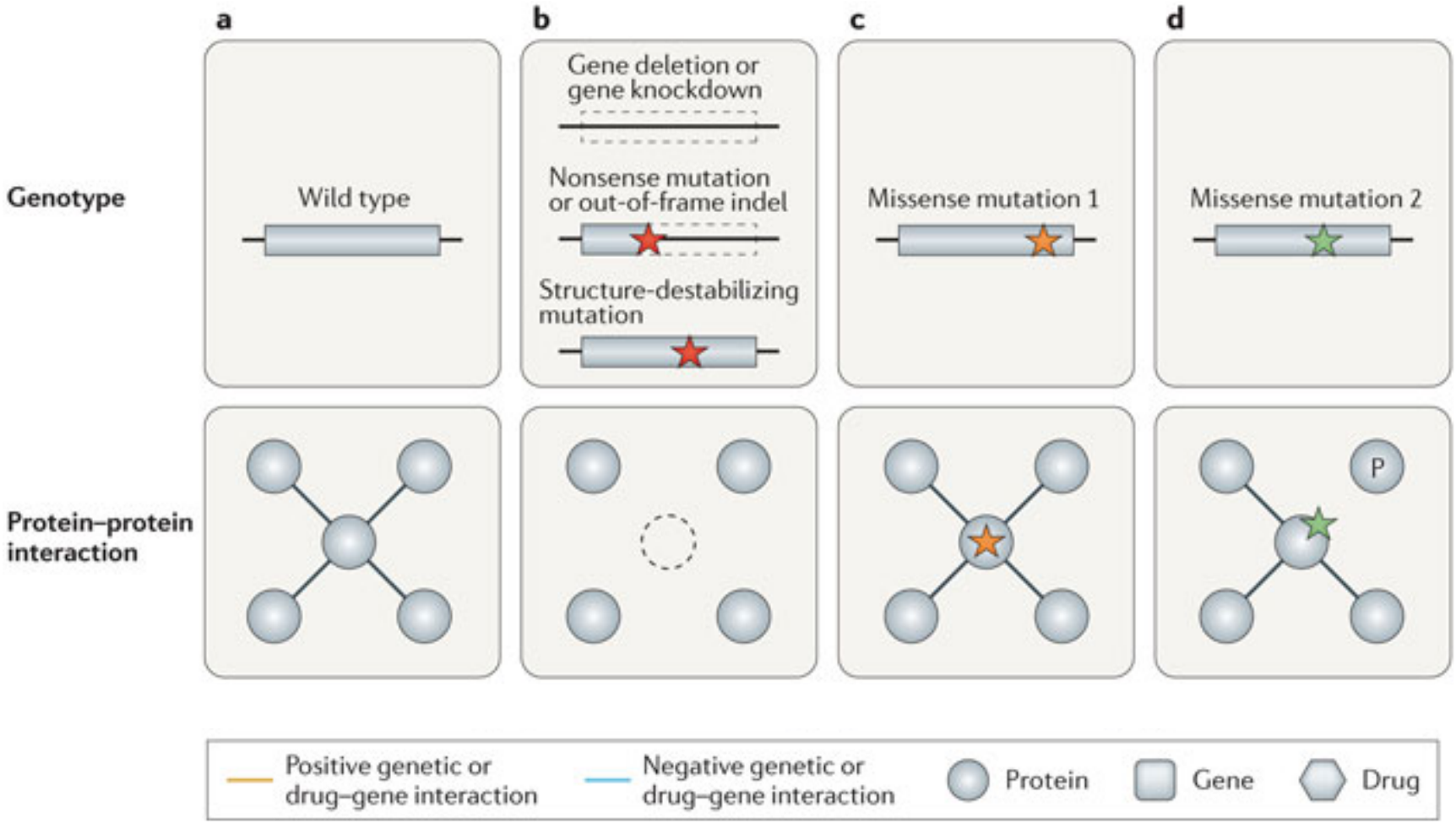
1. **Develop gene test to routinely group patients**
2. **Find new drugs that target specific groups better**

Networks

Interaction networks

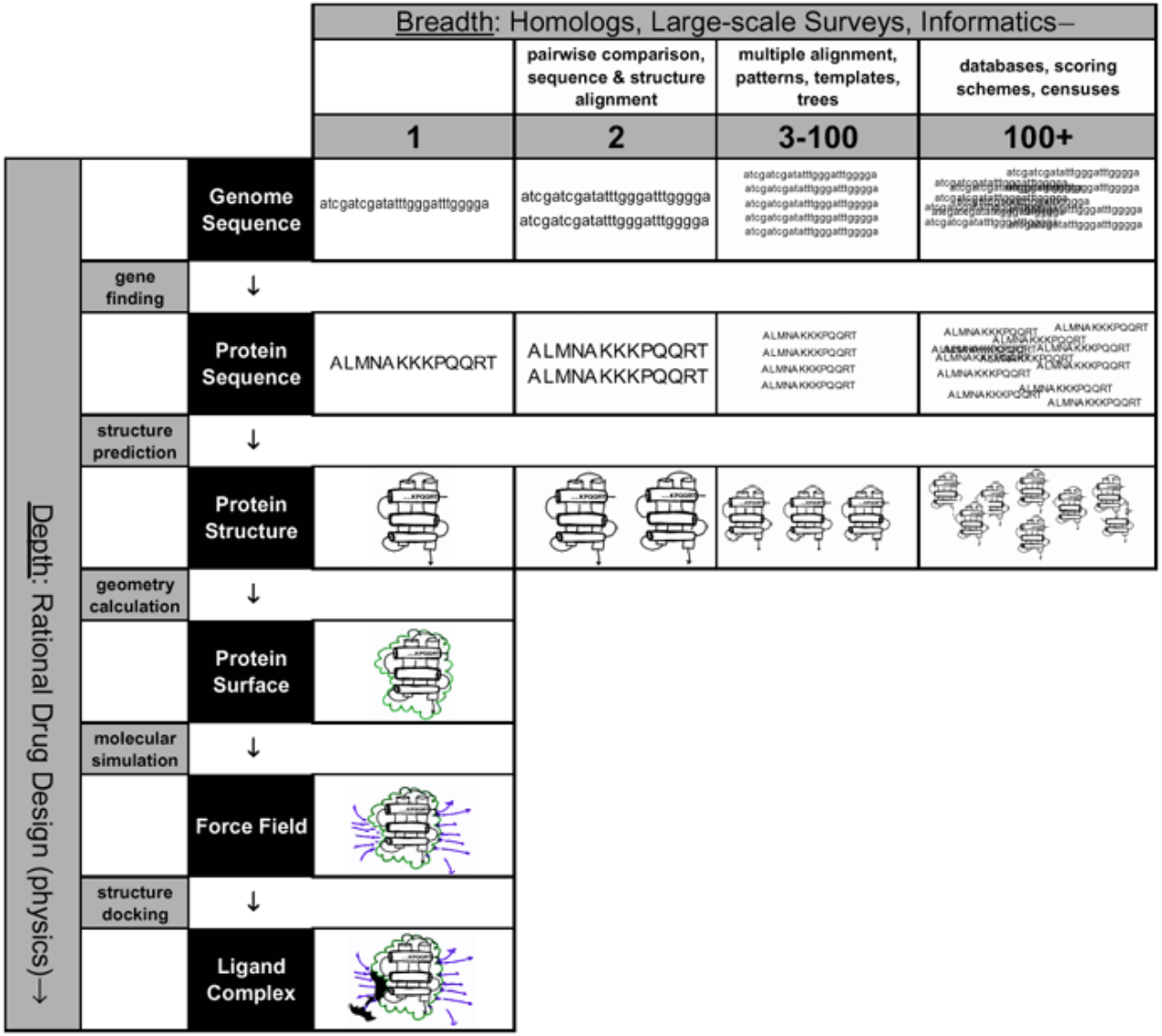


Consequences of mutations

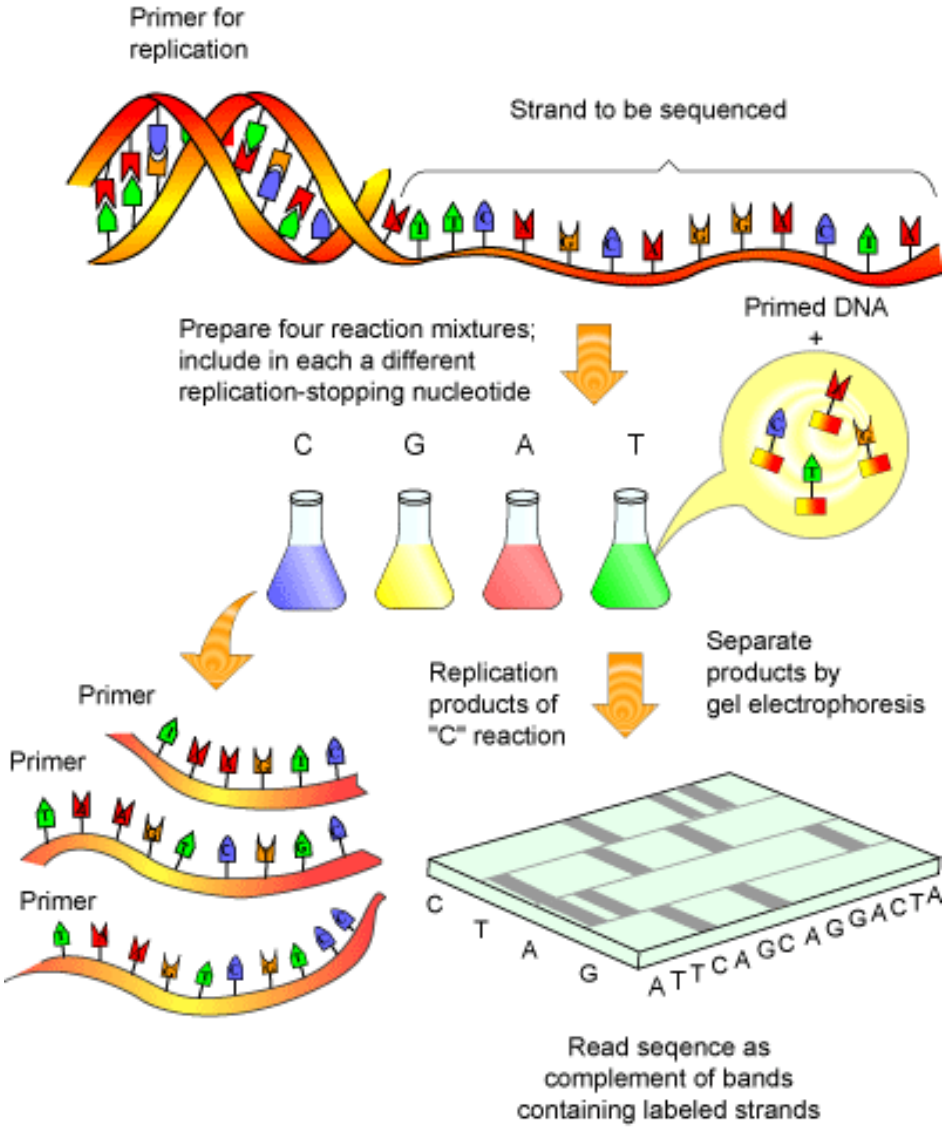


Thank you!

Bioinformatics Spectrum



DNA sequencing

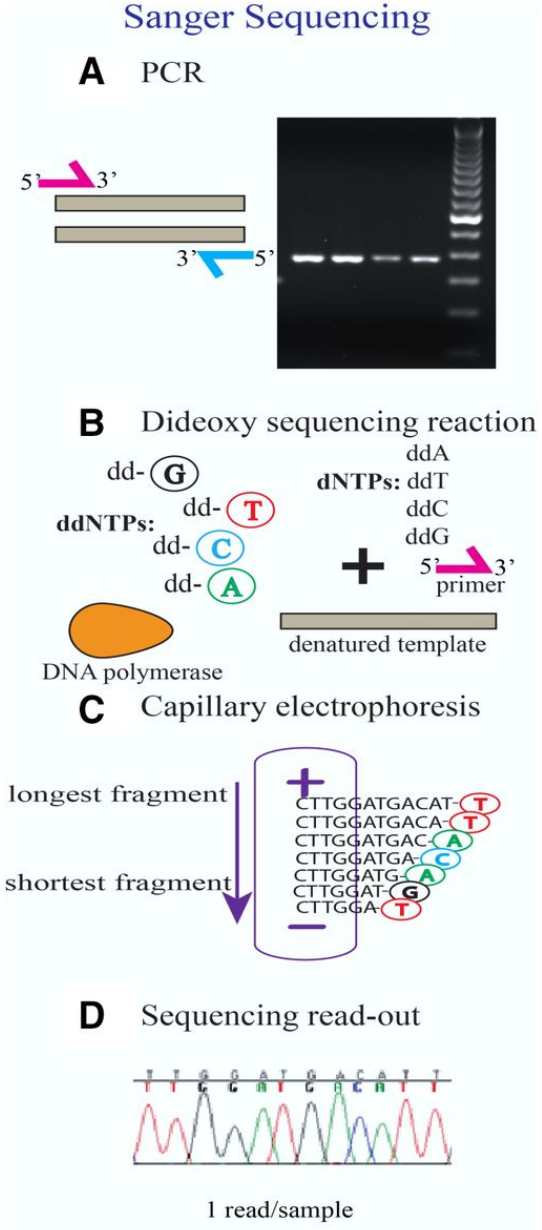


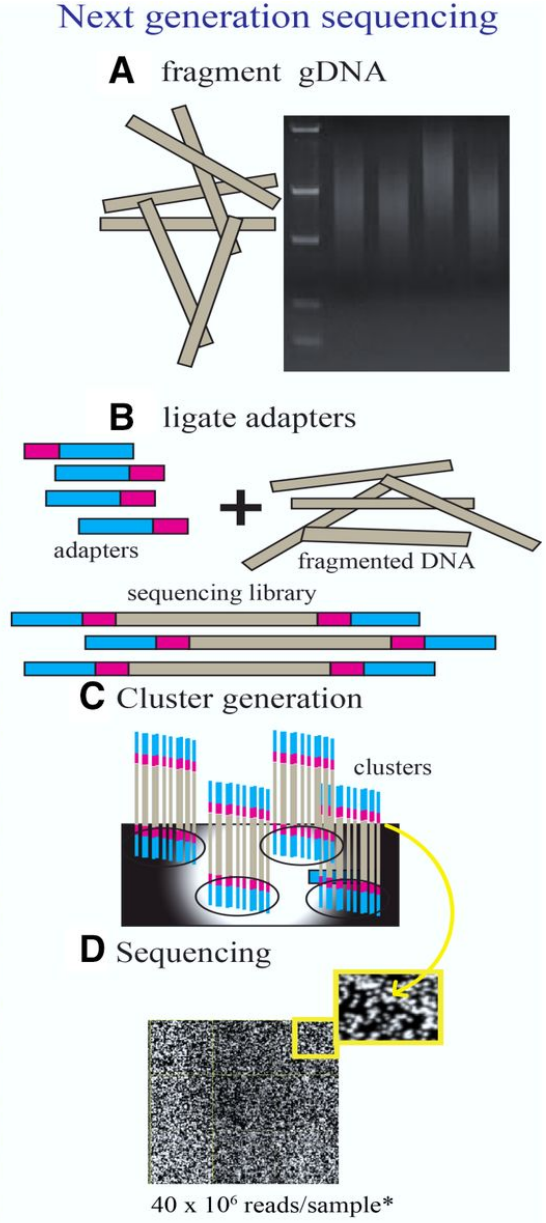
DNA sequencing: " to determine the exact sequence of bases (A, C, G, and T) in a DNA segment"

The DNA base sequence carries the information a cell needs to assemble protein and RNA molecules.

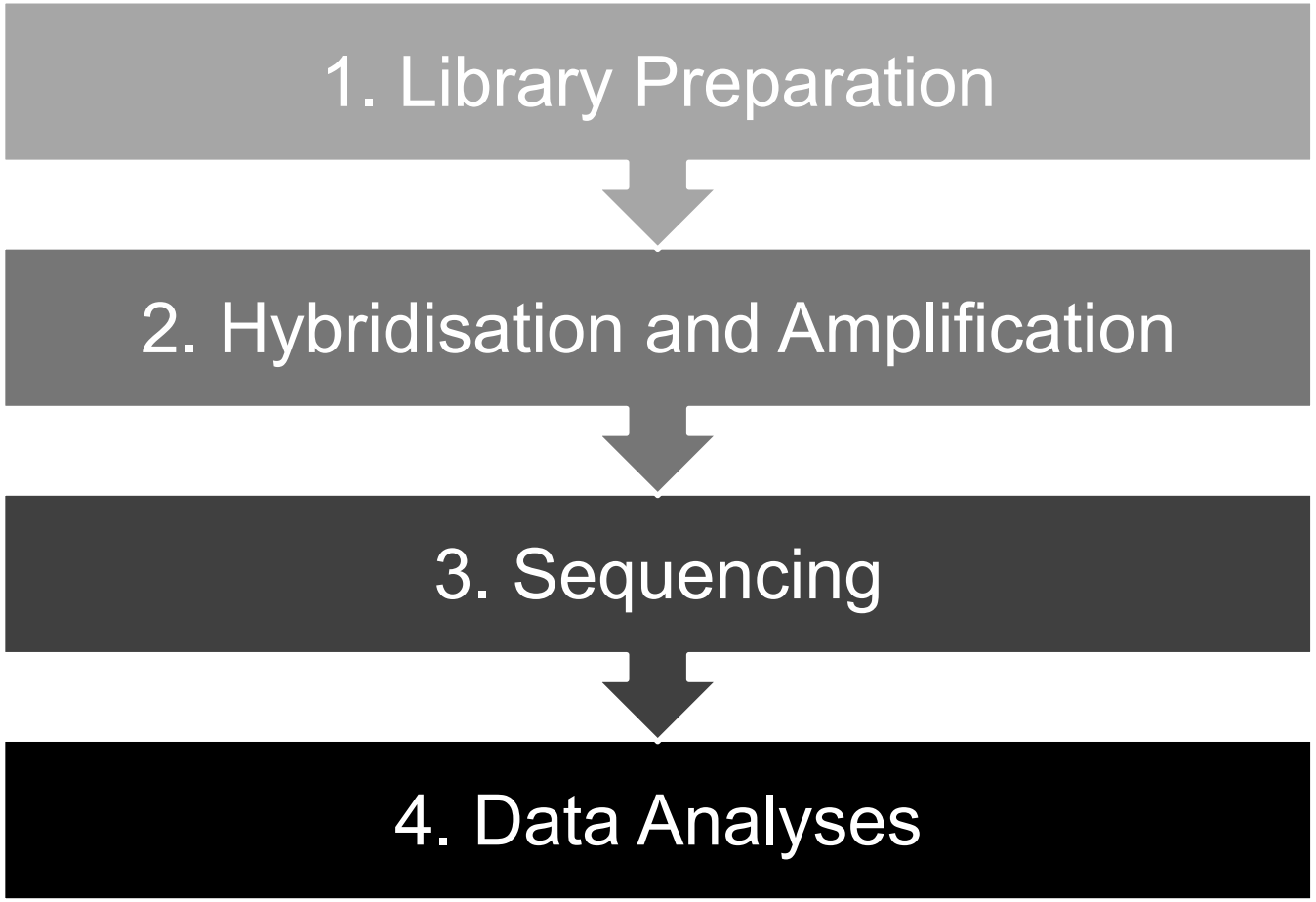
DNA sequence information is important to scientists investigating the functions of genes

Sequencing


Sanger Sequencing
A PCR
5'
3'
3'
5'
B Dideoxy sequencing reaction
dd- G
dd- T
ddNTPs:
dd- C
dd- A
DNA polymerase
ddA
dNTPs: ddT
ddC
ddG
5'
3'
primer
denatured template
C Capillary electrophoresis
longest fragment
shortest fragment
CTTGGATGACAT-T
CTTGGATGACA-T
CTTGGATGAC-A
CTTGGATGA-C
CTTGGATG-A
CTTGGAT-G
CTTGGA-T
D Sequencing read-out
1 read/sample


Next generation sequencing
A fragment gDNA
B ligate adapters
adapters
fragmented DNA
sequencing library
C Cluster generation
clusters
D Sequencing
40 x 10⁶ reads/sample*

Example: Illumina NGS workflow



Borrowed from
Dr Laura Emery
EMBL-EBI