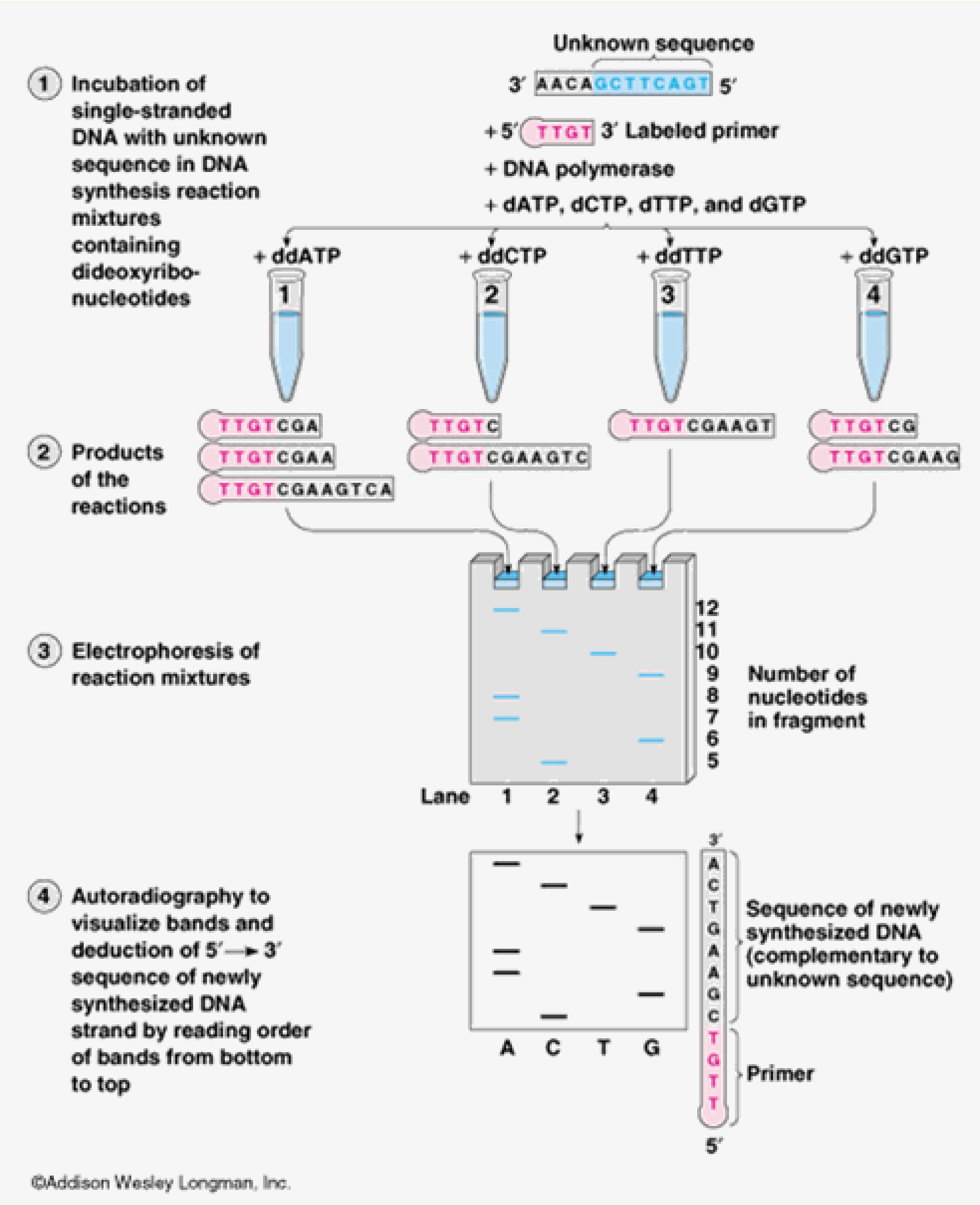
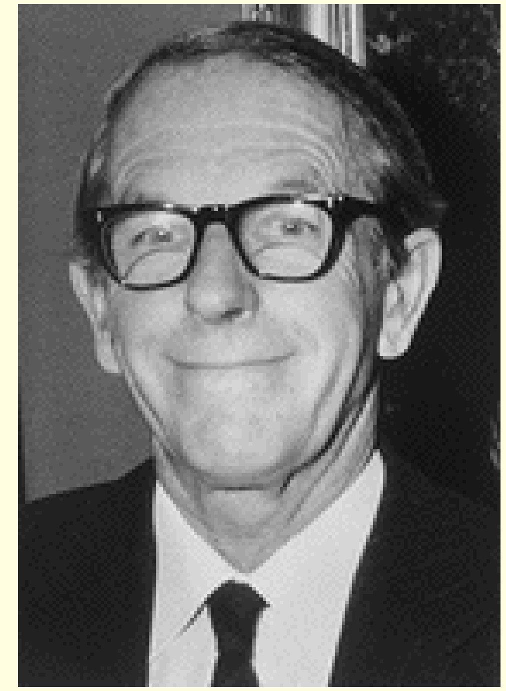
CMPS 6630: Introduction to Computational Biology and Bioinformatics

High-Throughput Sequencing and Applications

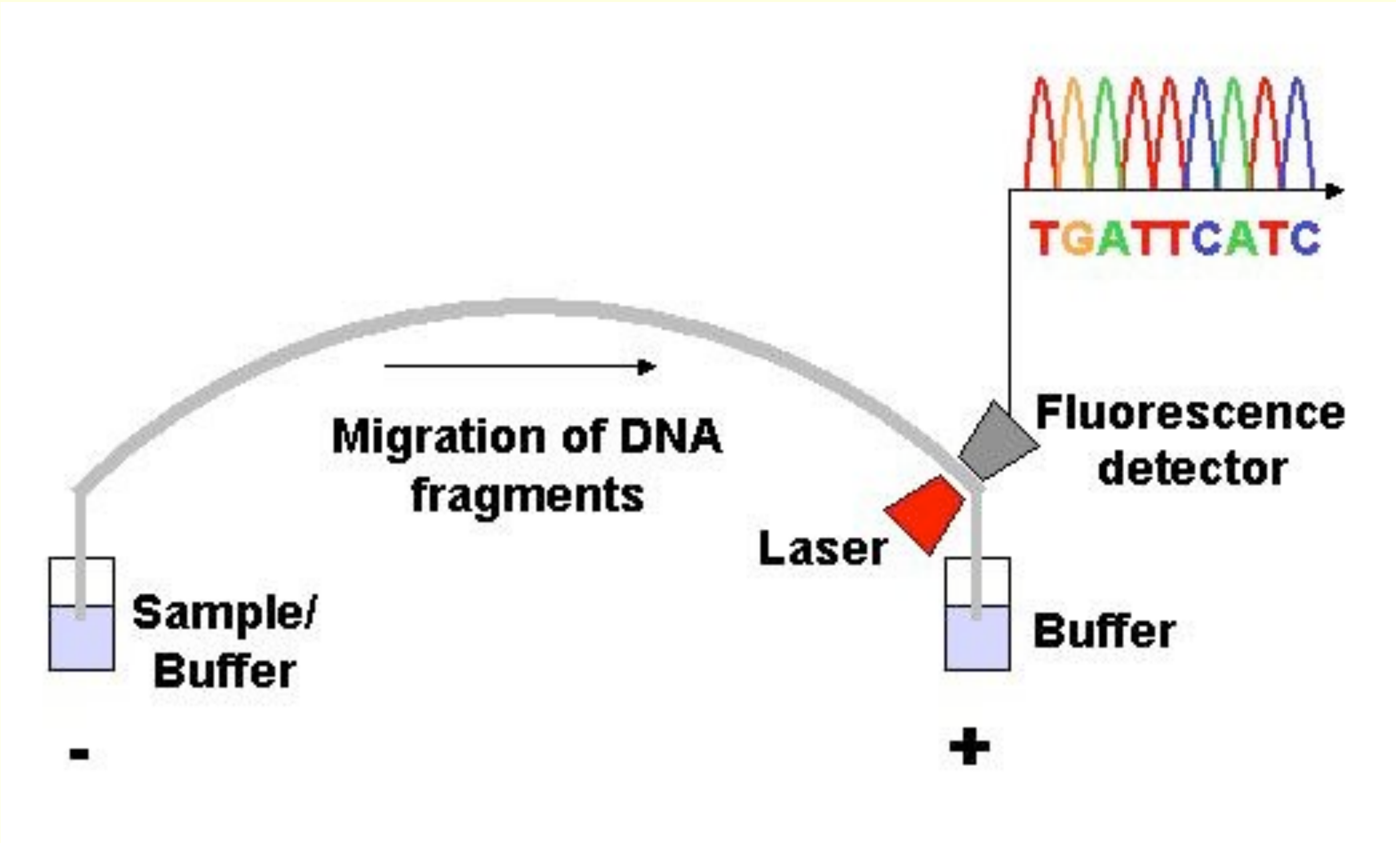


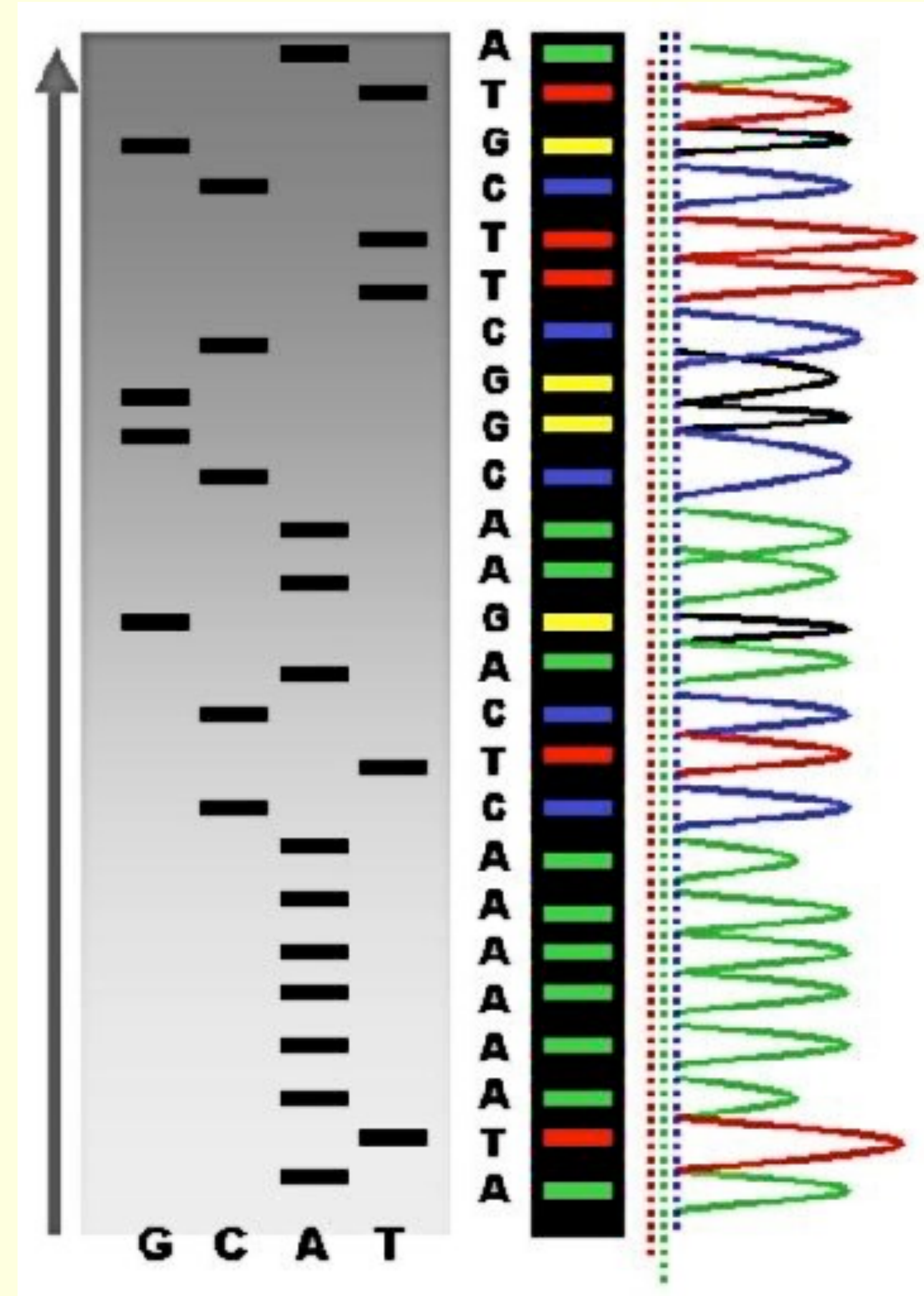


Sanger (1982) introduced chain-termination sequencing.

Main idea: Obtain fragments of all possible lengths, ending in A, C, T, G.

Using gel electrophoresis, we can separate fragments of differing lengths, and then assemble them.

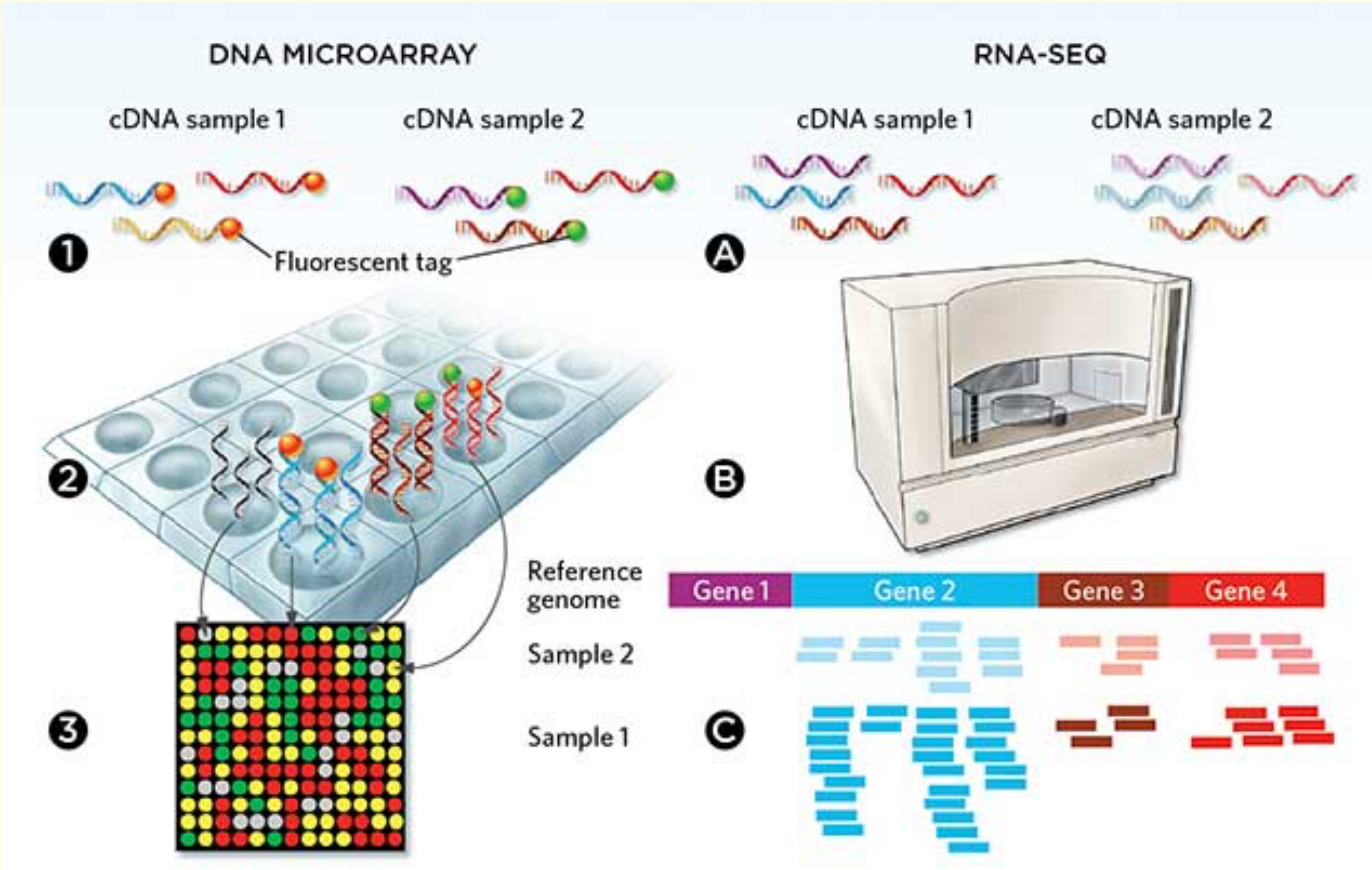




Sanger sequencing typically seeks to identify a single gene sequence.

What about sequence variation? Read coverage is an indicator…

<u>Drawback</u>: Can usually only examine variation of a single gene/sequence.



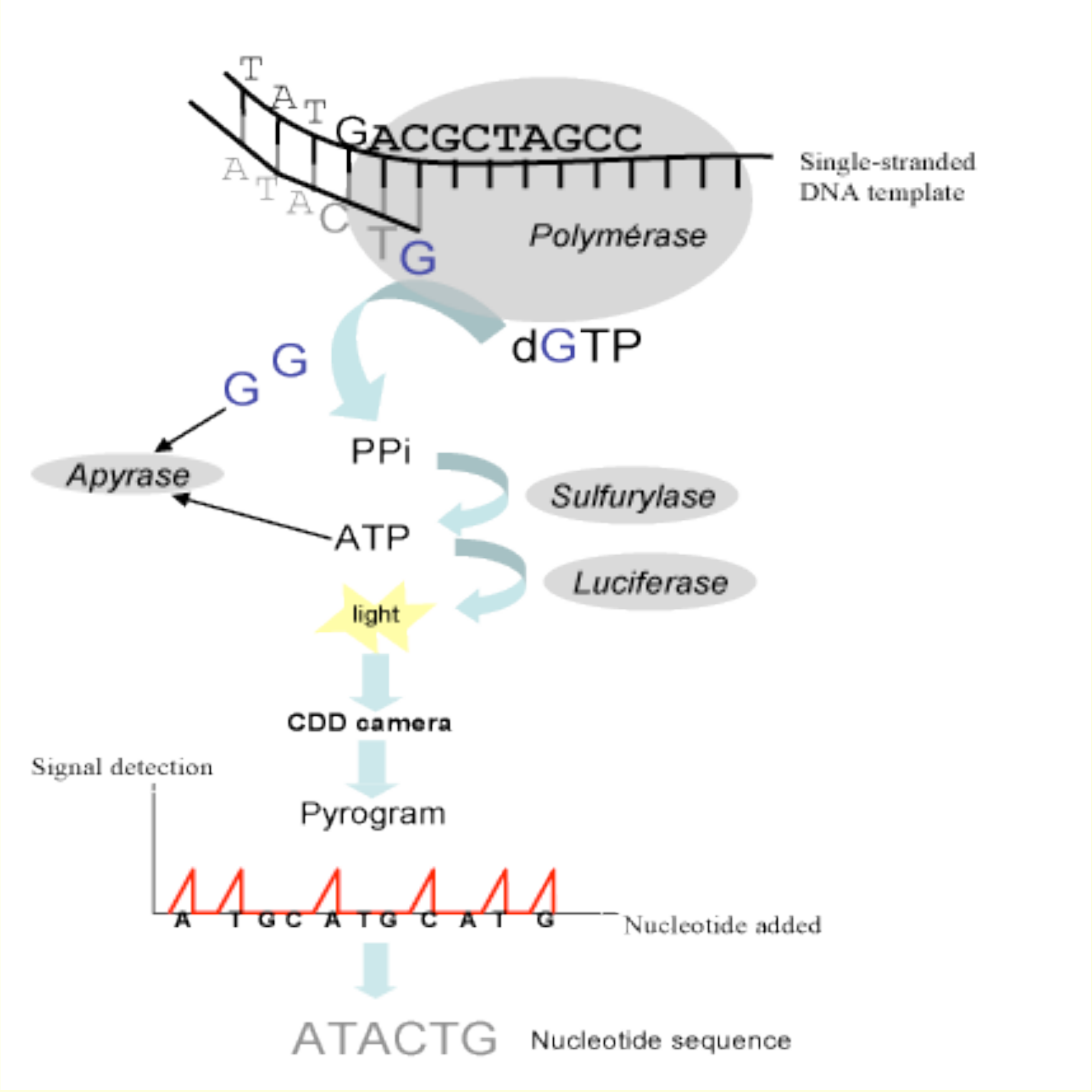
DNA microarray video

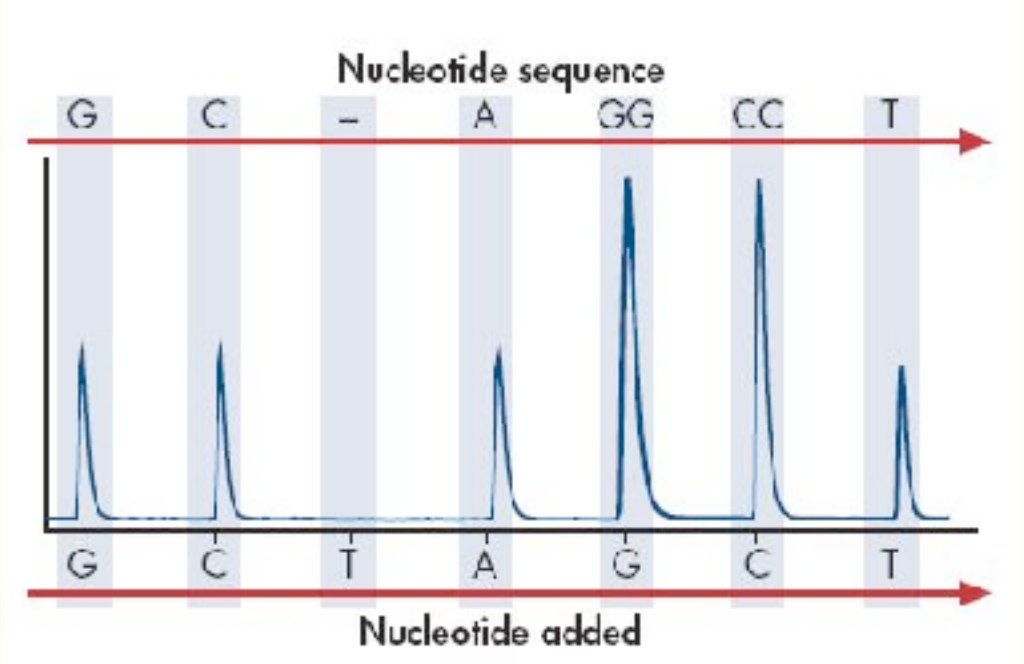
High-Throughput Sequencing

- In all HTS technologies, the idea is to perform sequencing in parallel (and at lower cost) using an “array” setup.
- The approach is essentially a combination of microarray technology and sequencing.
- Extremely high read coverage makes short reads ok; it is possible to sequence a whole genome much more quickly.

Pyrosequencing

Chemical reactions occur more quickly than capillary electrophoresis -- can we generate a signal during synthesis?

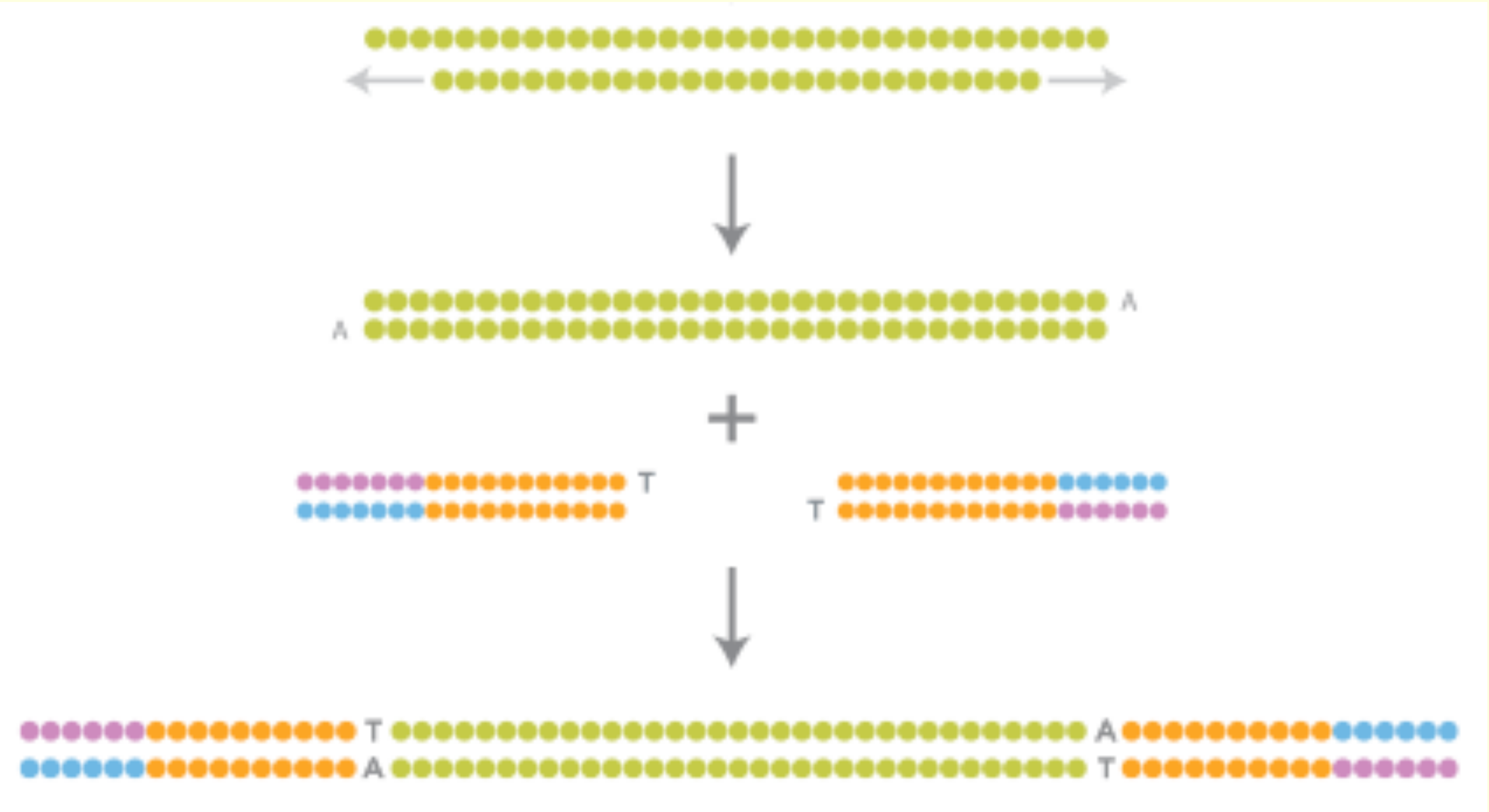




[Pyrosequencing Video]

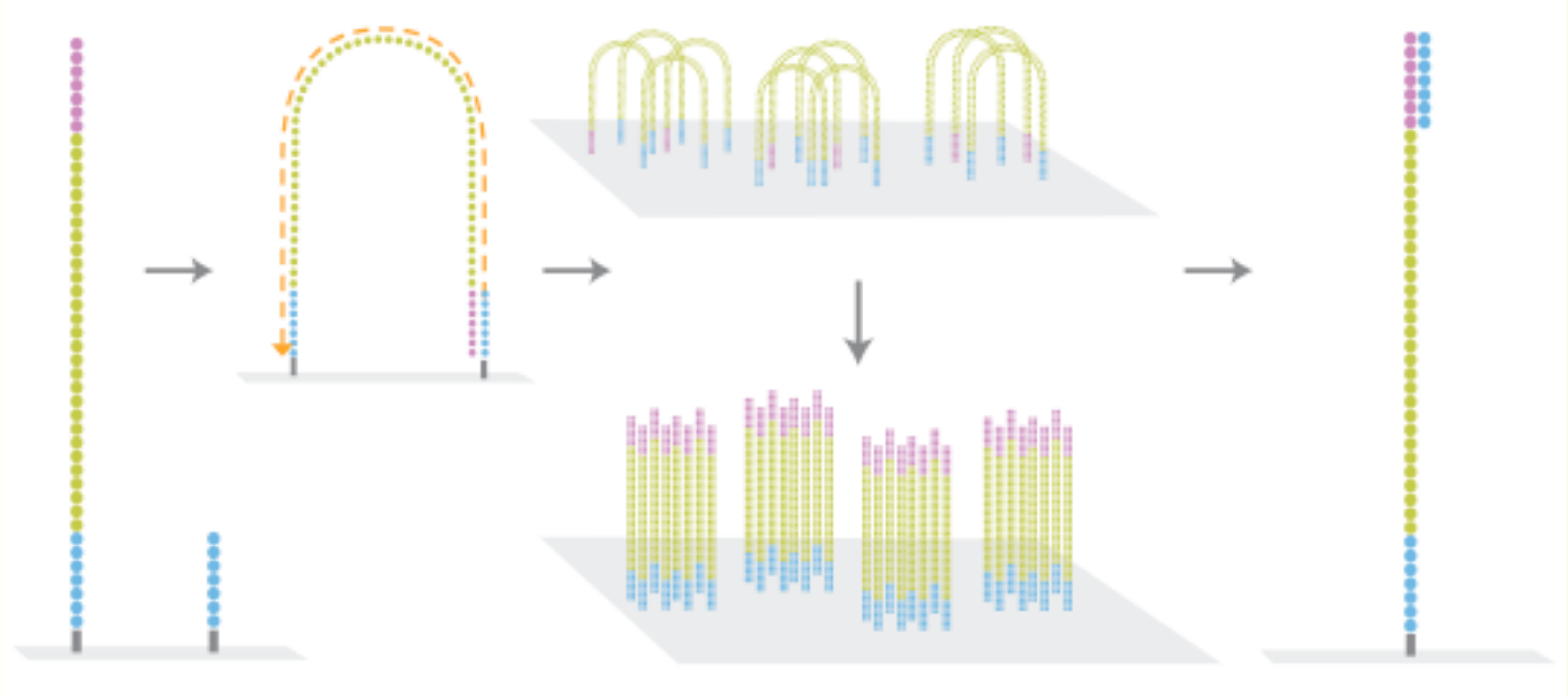
[454 Workflow]

Sequencing by Synthesis



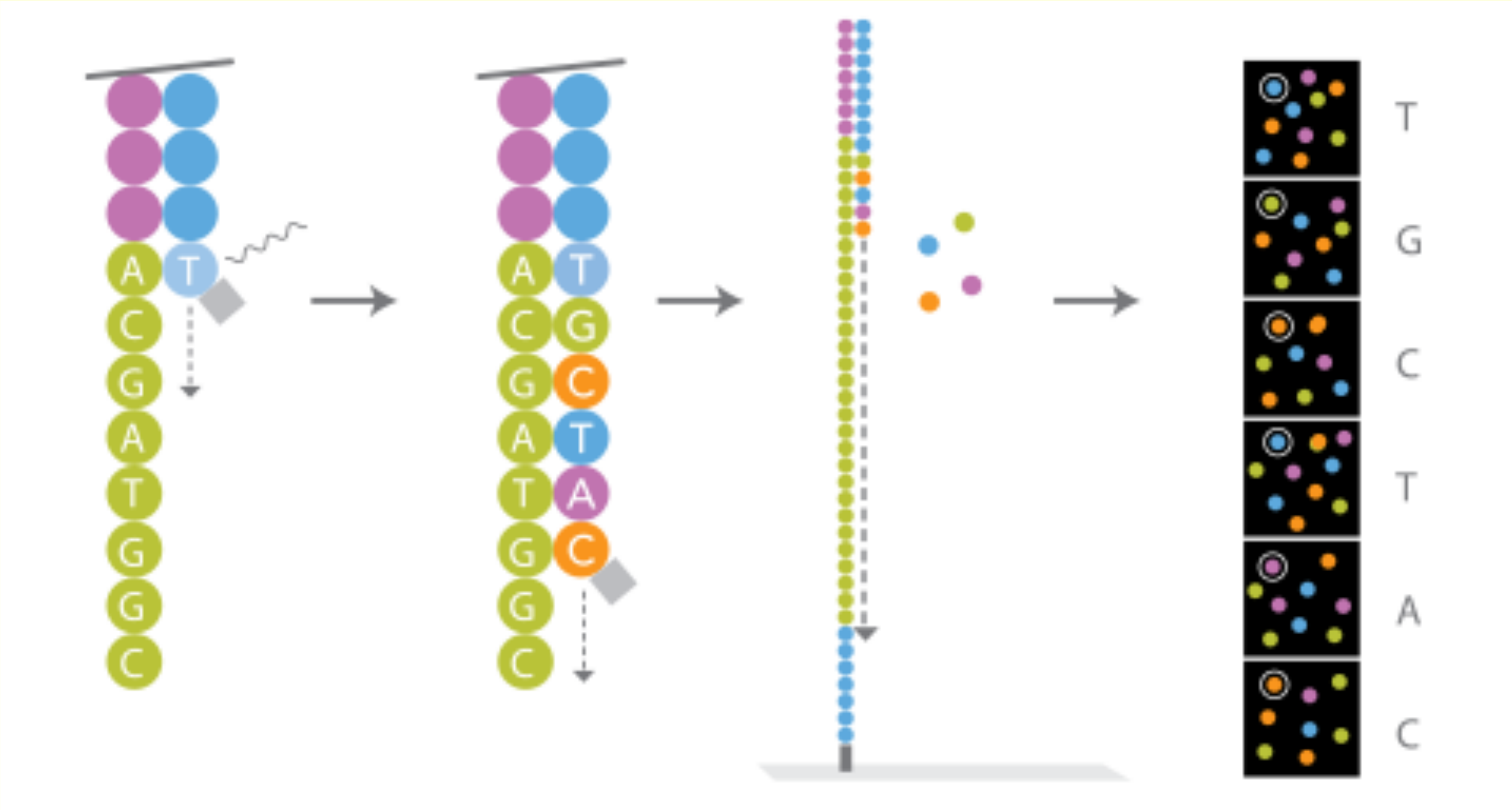
The sample is sheared and “adaptors” are added to the ends.

Sequencing by Synthesis



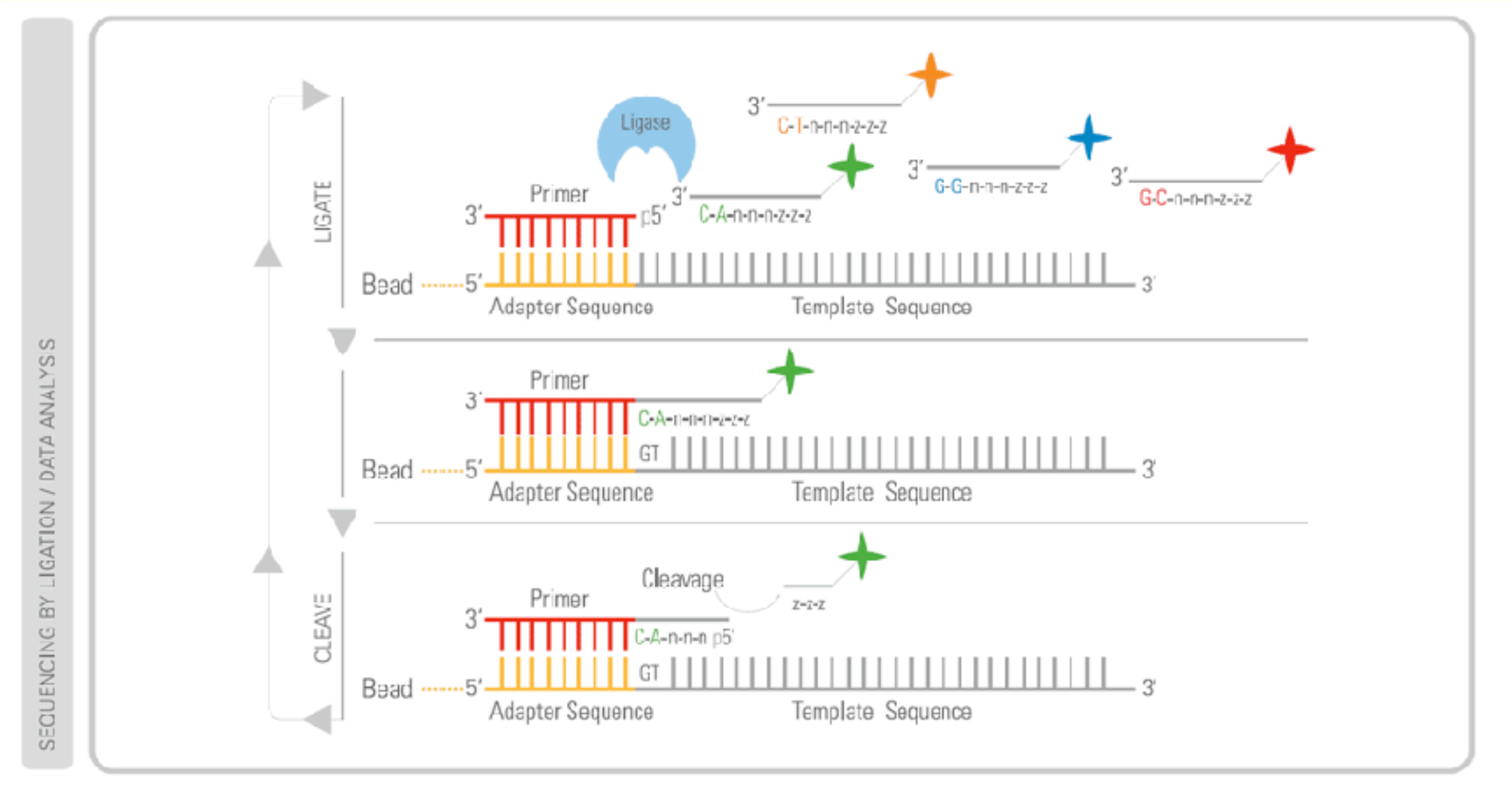
These fragments are clustered on a "flow cell", are copied, and the original fragments are cleaved and washed away.

Sequencing by Synthesis



Differentially fluorescent nucleotides are introduced and then washed away from the flow cell. Imaging is used to capture nucleotides that are synthesized. This yields the sequence of each fragment.

ABI SOLiD



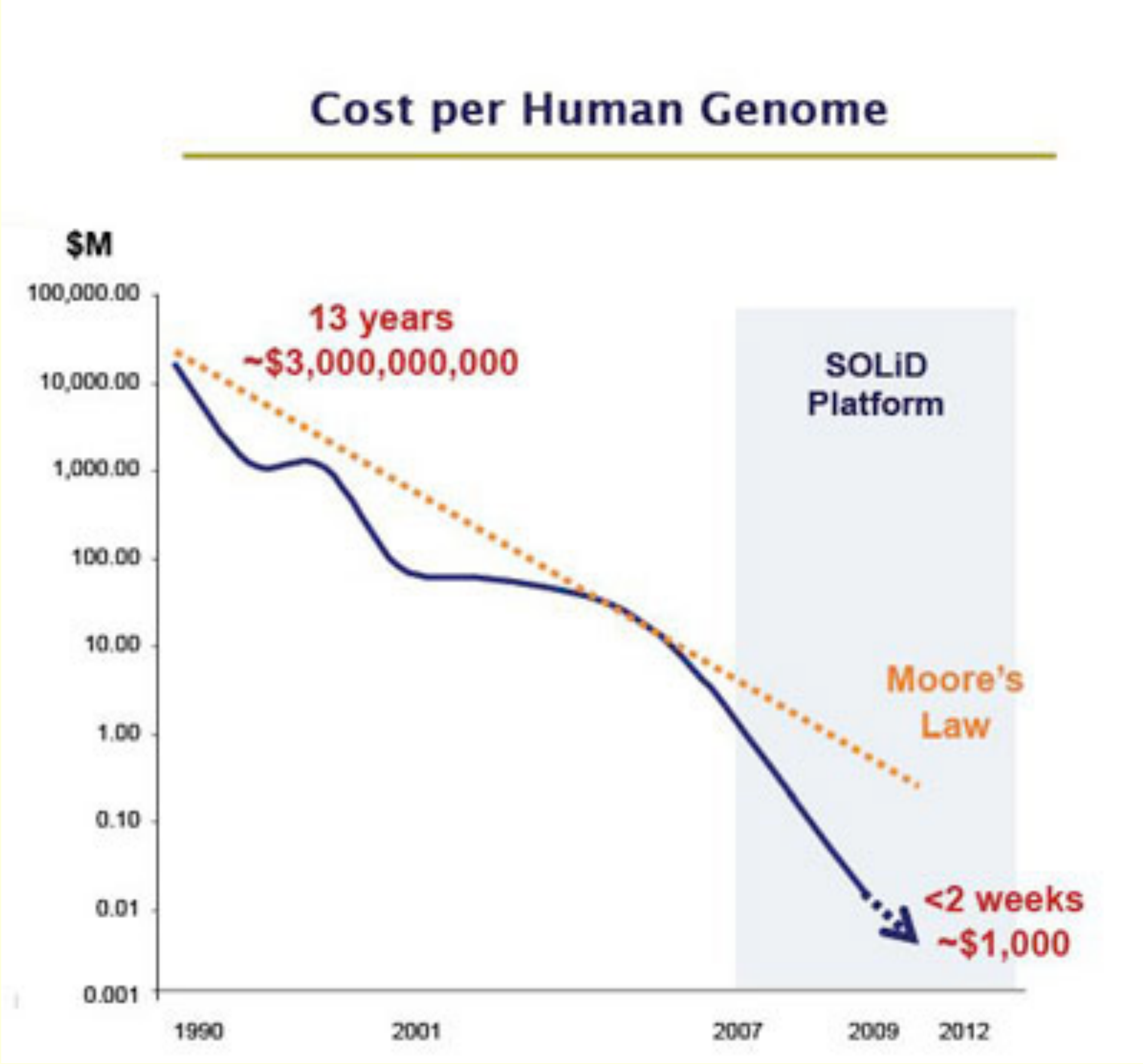
DNA Ligase is an enzyme that "fixes" DNA damage by synthesizing covalent bonds on both strands. Sequencing-by-ligation utilizes "dibase" fluorescent coding to report ligation.

Technology Summary

Method	Read Length	Sequences per Run	Utility
Dye-Terminator (Sanger)	500-1500 bp	384	*de novo* and low-throughput applications
454/Roche	120-400 bp	∼200,000	*de novo* and medium-throughput applications
Illumina/Solexa	36-60 bp	∼20,000,000	high-throughput applications

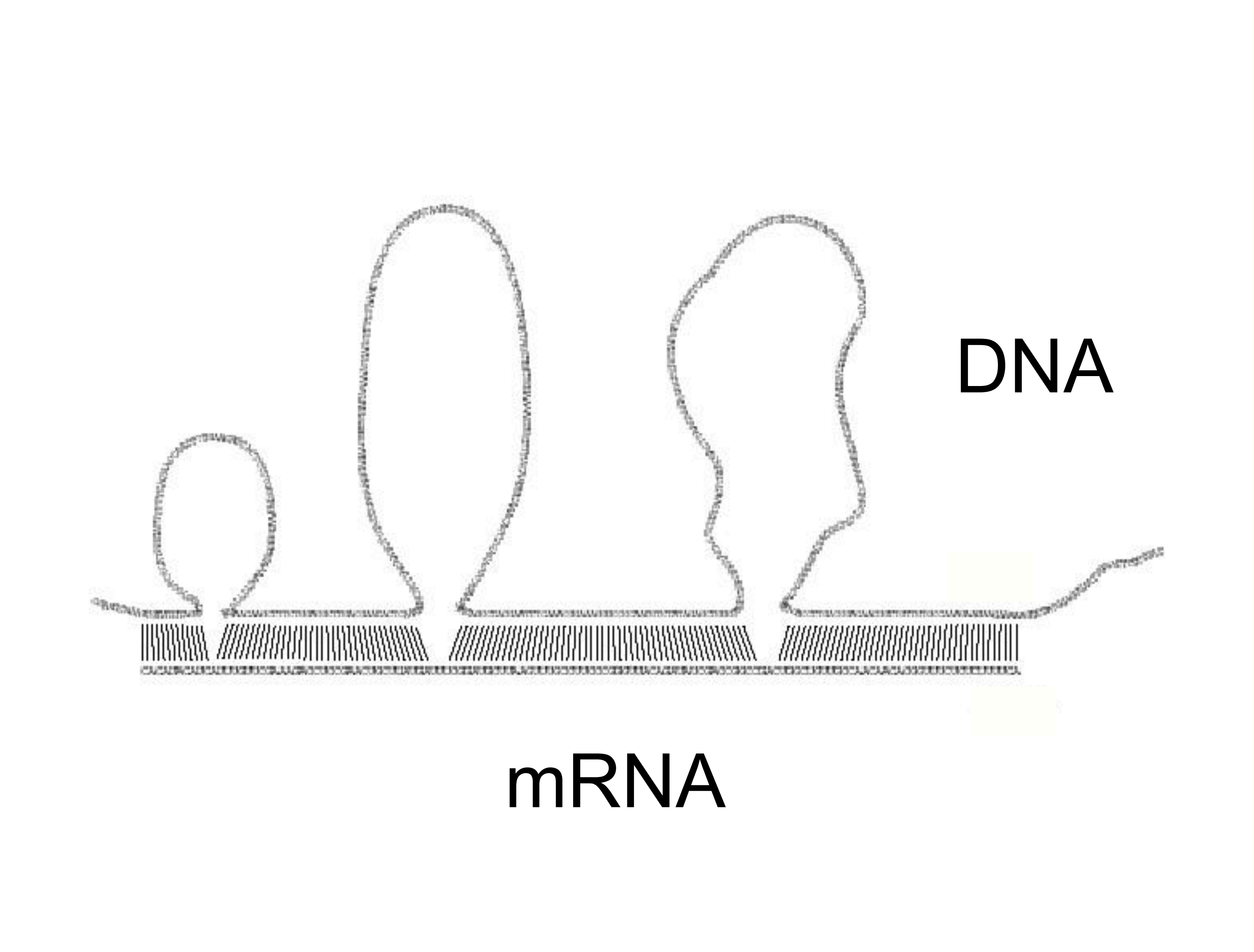
	Illumina	454	SOLiD	Helicos
Method	Rev. Term.	Pyro. Sequ.	Oligo Ligation	Single Mol.
Read Length	36-2x100	300-400	36	25-45
Error Rate	∼1%	>1%	∼0.1%	<1%
Data/Run (Gb)	1-3	0.1	2-3	8
Cost (per Gb)	$6,000	$84,000	$6,000	$2,500

Genome (Re)sequencing



Cost for sequencing an entire genome continues to drop -- this is the promise of personalized medicine.

Gene Splicing



Sharp and Roberts (1977) hybridized the mRNA for a viral protein to its corresponding "gene" and showed that transcription can be "spliced".

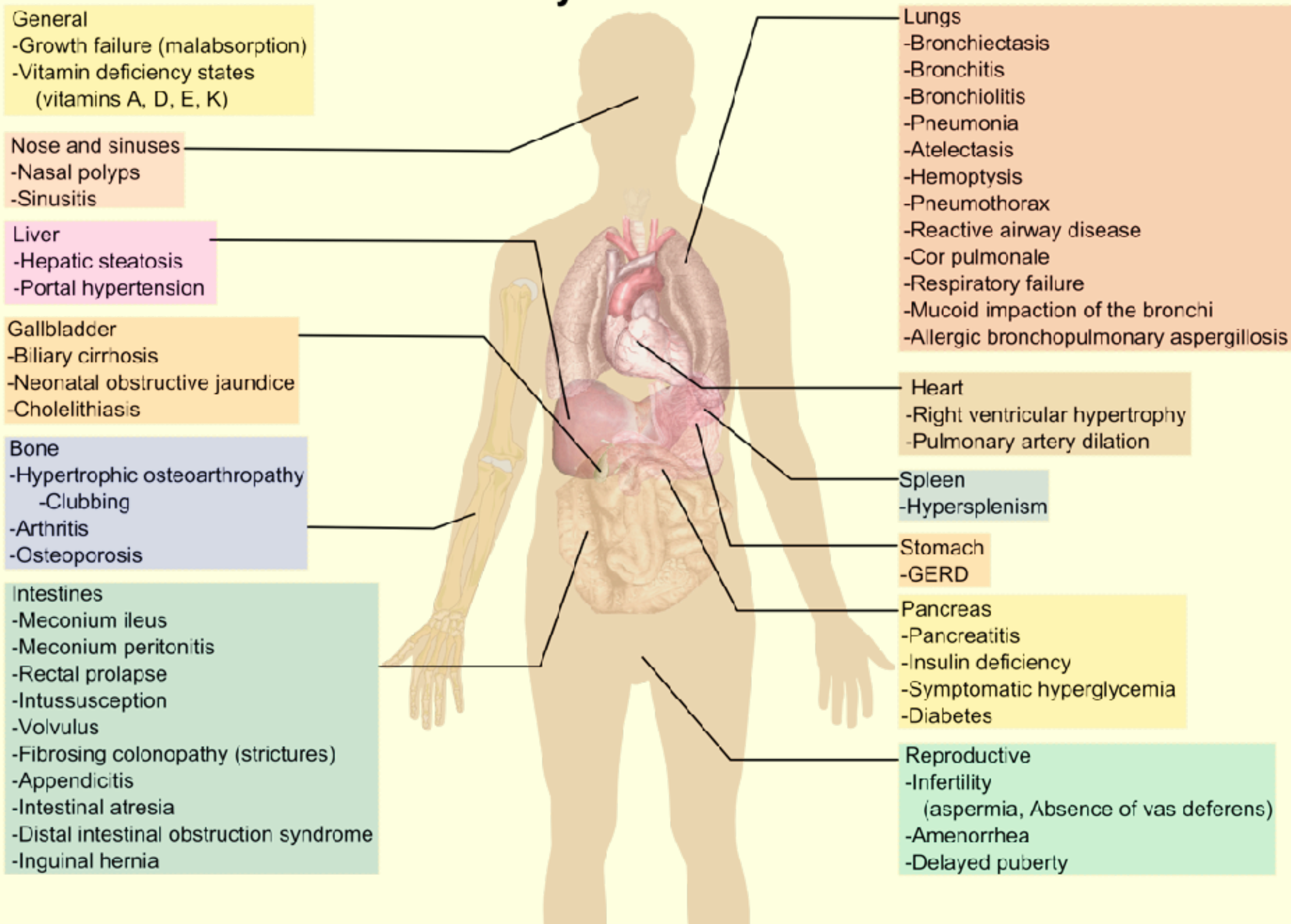
So given a genomic sequence, we need to identify fragmented exonic components (with or without mRNA).

Alternative Splicing

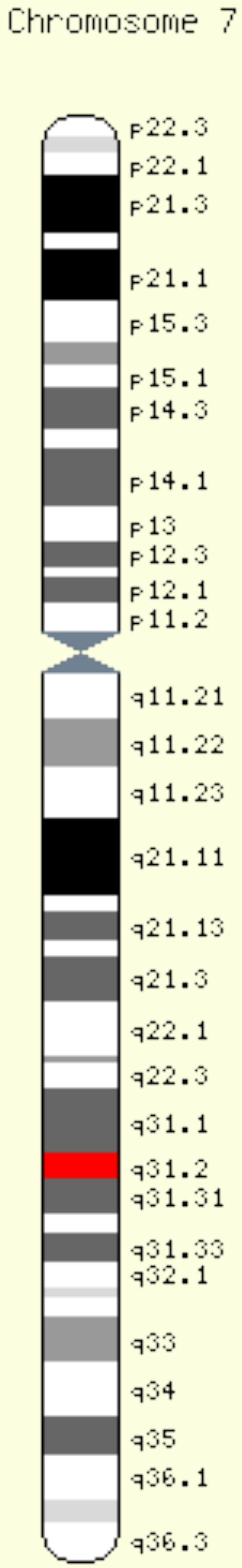
- Alternative splicing is a regulatory mechanism in different tissue types.
- What is the pattern of differential splicing, between individuals, or between tissue types?
- Numerous diseases have been shown to be splicing-related (e.g. isoform ratios, protein misfolding).
- Microarrays can be used to probe alternative splicing if the probes are designed for exons.
- With high-throughput sequencing, however, we can improve resolution, and actually discover exons.

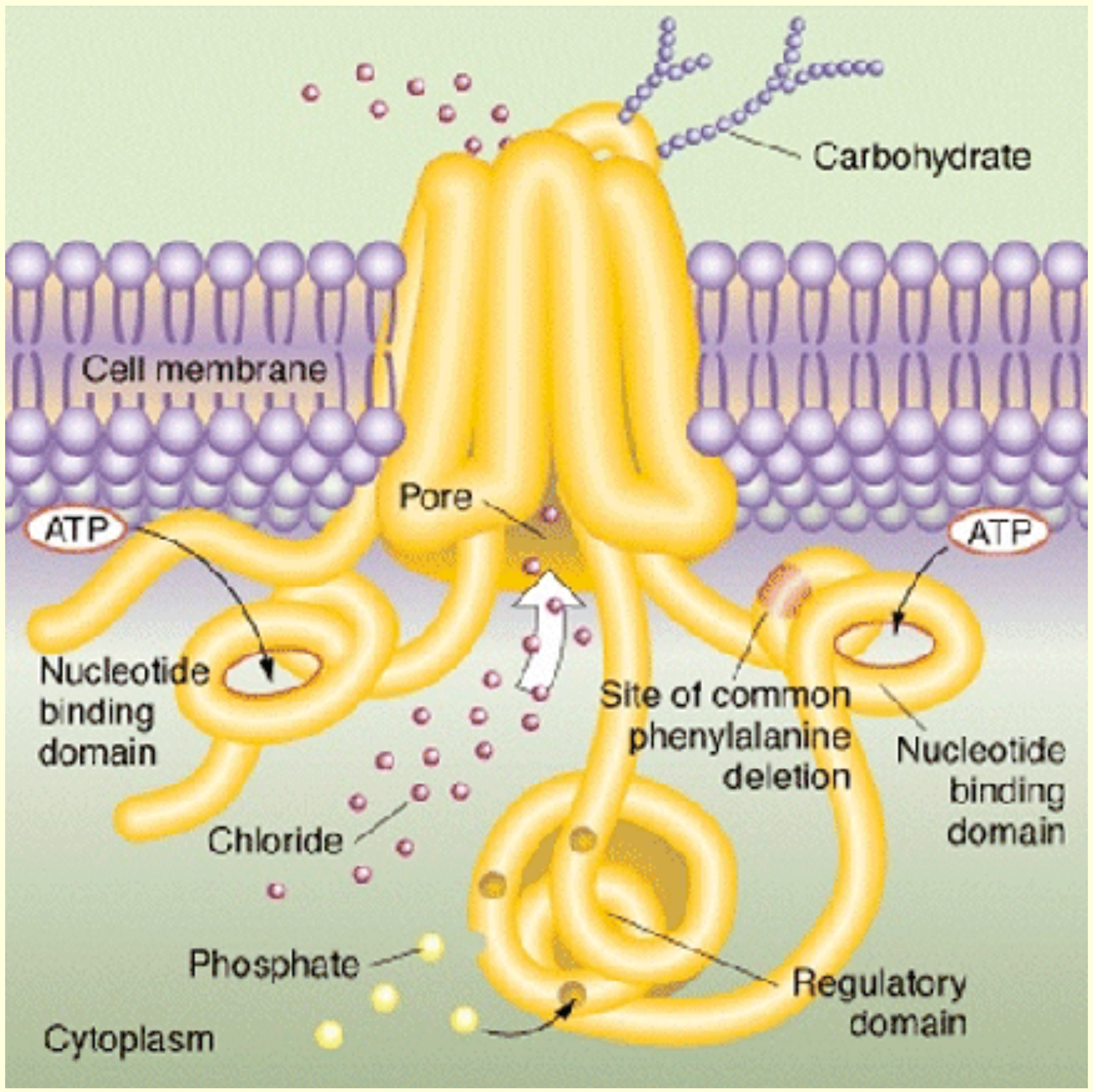
Manifestations of

Cystic Fibrosis



Cystic Fibrosis Mechanism



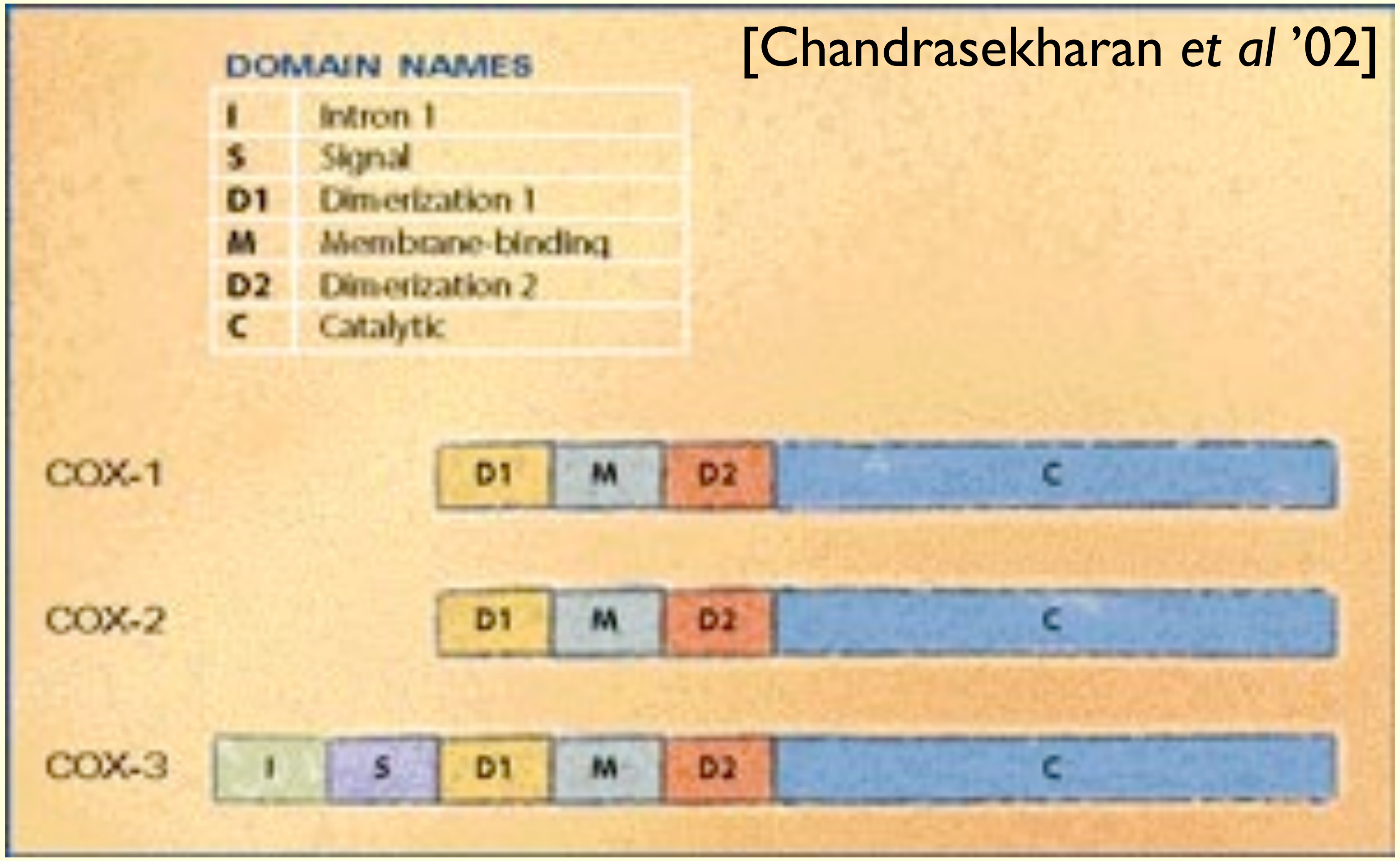


cystic fibrosis transmembrane conductance regulator (CFTR)

13-20% of CF mutations are related to "mis-splicing."

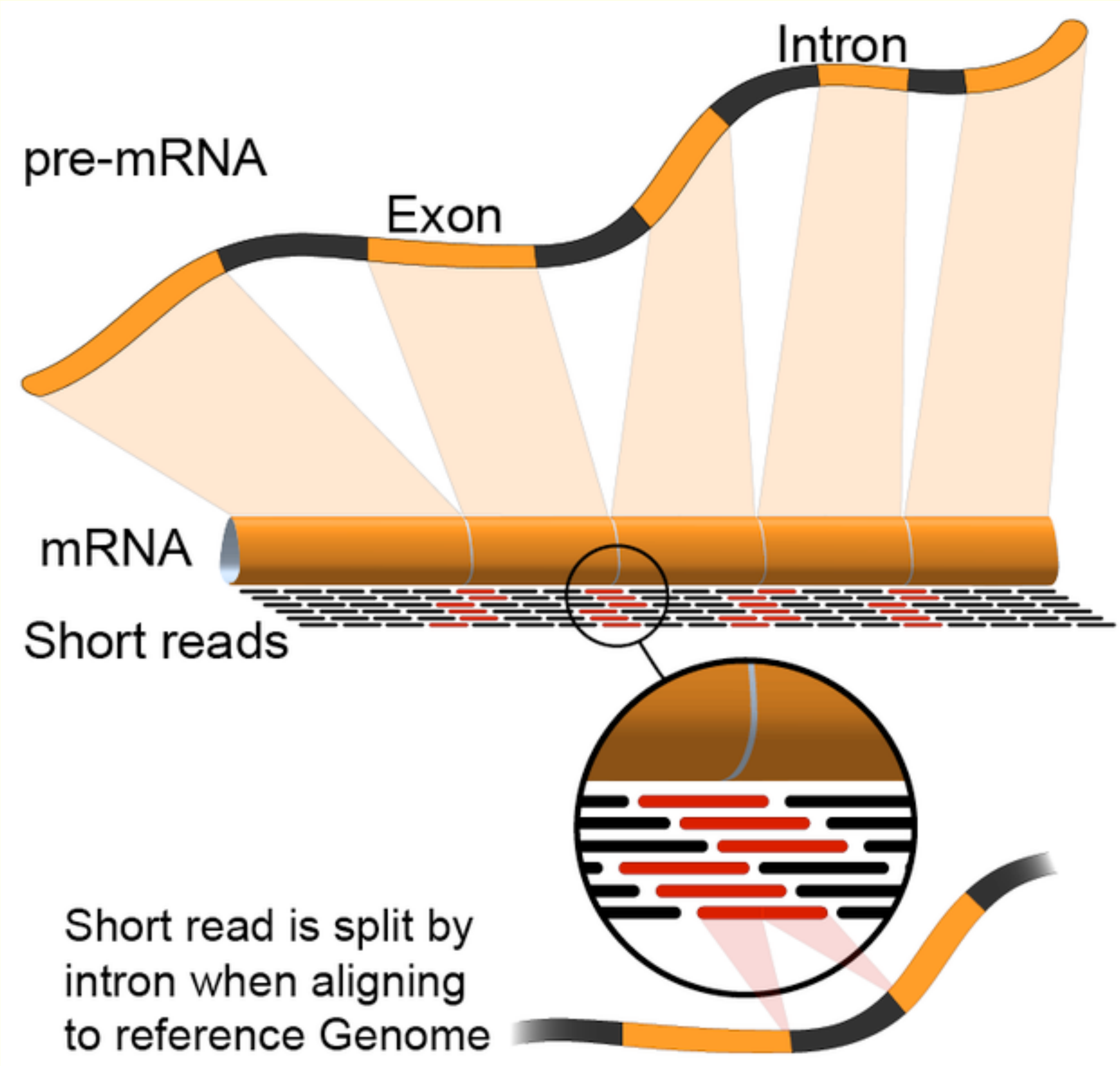
Ever use Tylenol?

Humans have been using non-steroid anti-inflammatory drugs (NSAIDS) for 3500+ years.



Cyclooxygenase (COX) enzyme regulates pain and inflammation. COX-2 is a recent target for new pain medications. Acetominophen was recently discovered to act on a COX "isozyme", COX-3, localized in the brain.

RNA-Seq



Hidden Variables:
Each exon is "on" or "off".

Observed Variables:
Reads mapped to a reference genome.

Goal: Identify most likely states of hidden variables, i.e., identify RNA isoforms.

Is there an efficient algorithm to infer isoforms?

Before HTS

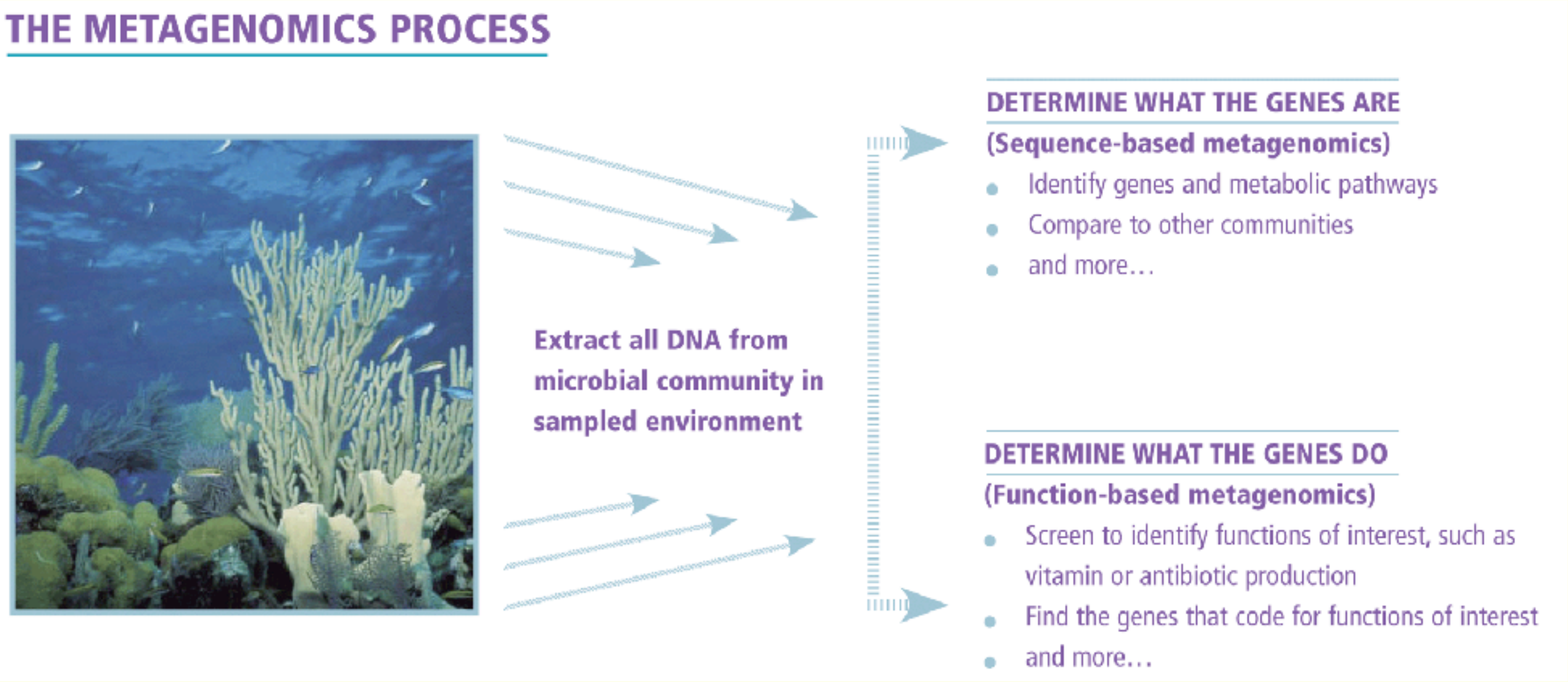
- Microarrays can be used to probe alternative splicing if the probes are designed for exons.
- What if exons are not actually known?

With HTS

- Using a reference genome, we can map short reads.
- A coverage rate that is as expected can be used to highlight exons that are "spliced in". Exons that are "spliced out" will have lower than expected read coverage.

Applications

- Can uncover differences in gene expression between tissues in one organism.
- Can uncover differences in gene expression in a given tissue type across a population.

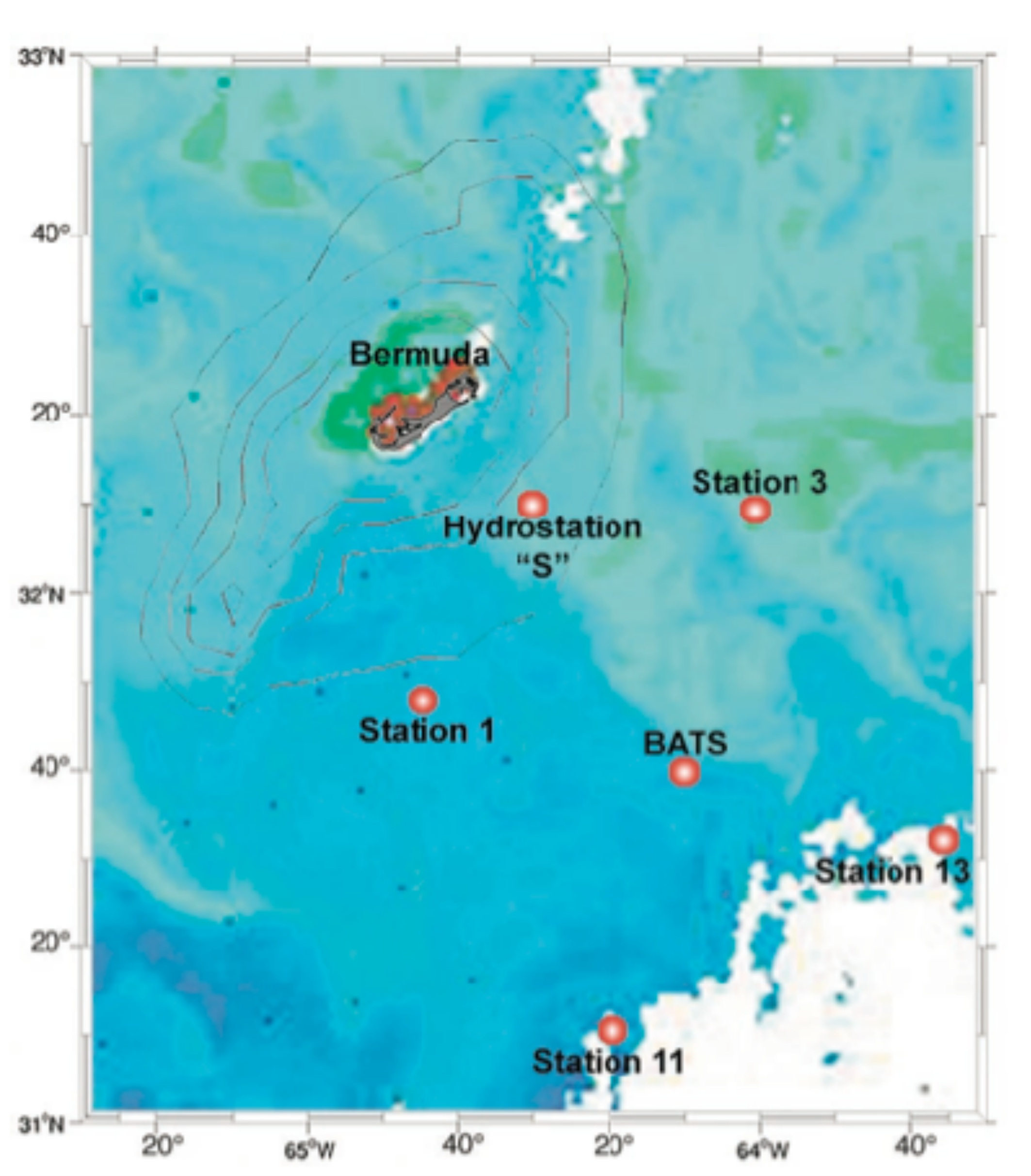


Idea: Collect an environmental sample, fragment and sequence DNA/mRNA. Map reads and try to assemble genes present in sample (not whole genomes).

Approach

- Shotgun sequencing can be used, but assembly is nearly impossible. Assembly of genes is, however, possible.
- Discovery of organisms in a particular sample is a basic task.
- [Venter *et al* ’04] showed that it is possible to reconstruct fairly complex phylogenetic information using traditional sequencing.
- High-throughput sequencing provides a method to sequence individual genes -- short reads are fine because we are not actually trying to assemble a genome.

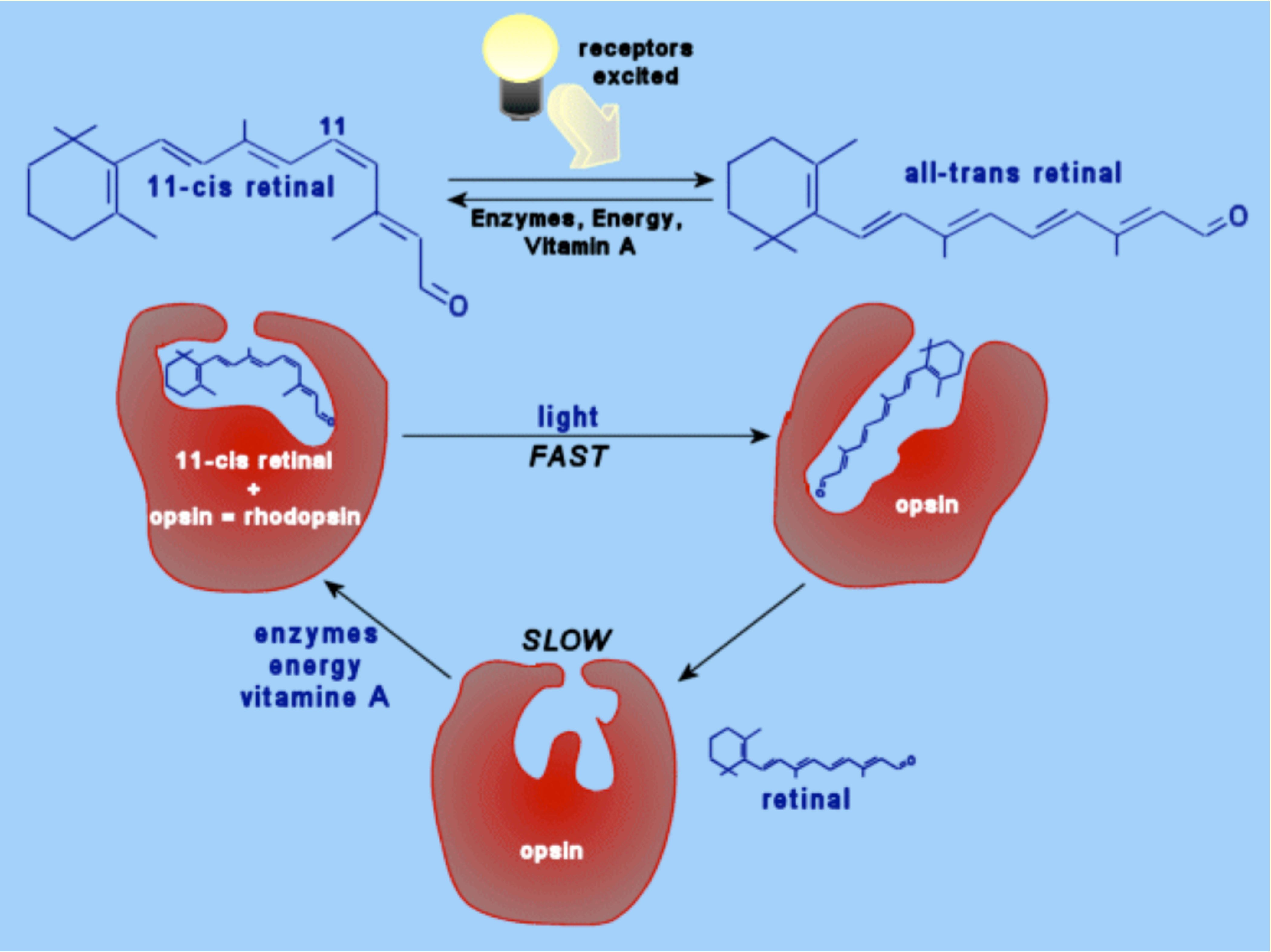
Oceanic Metagenomics



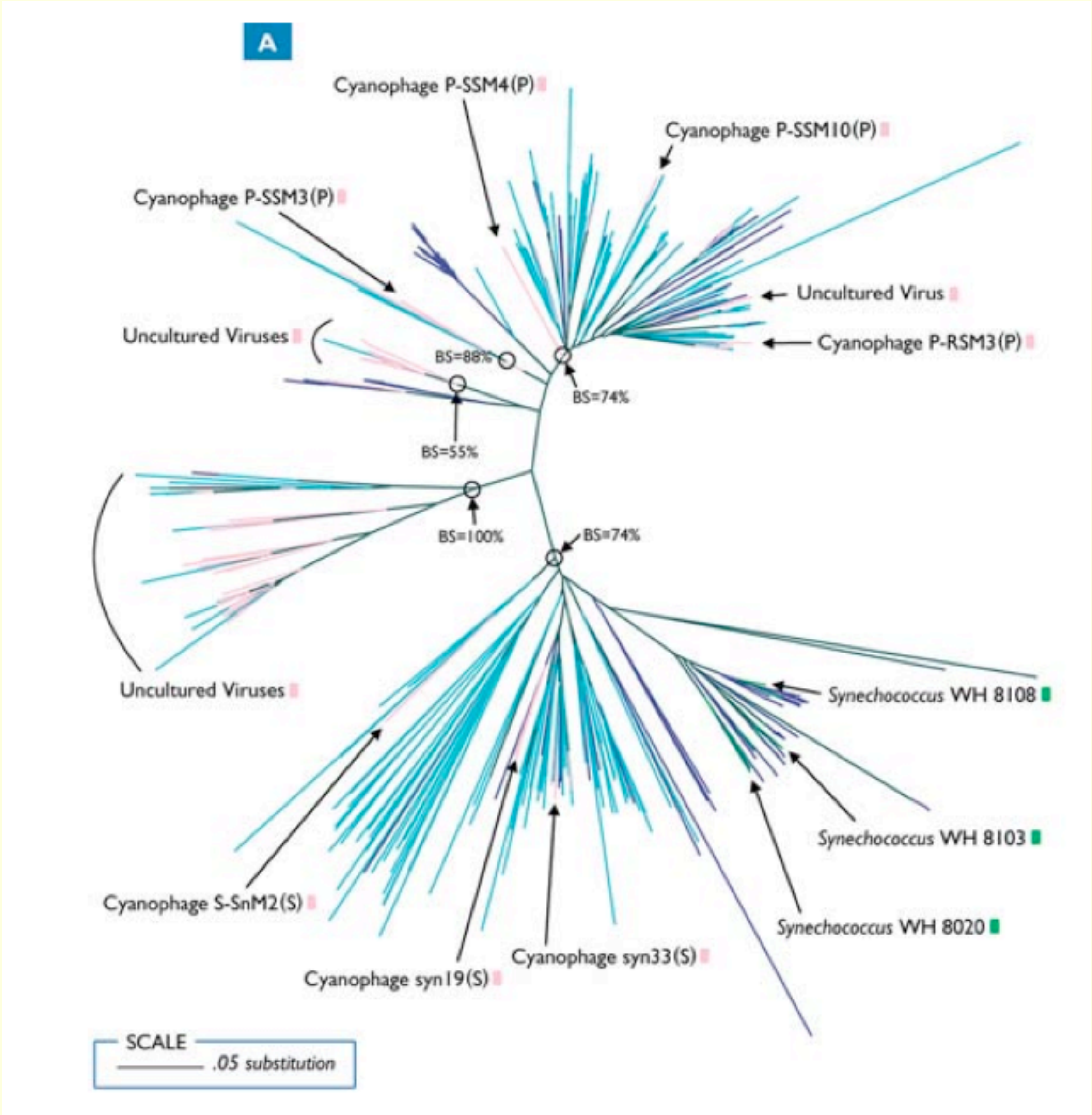
Sargasso Sea Data Set [Venter et al, '04]

An avid sailor, Venter conducted numerous expeditions to collect oceanic samples of bacteria and viruses. Metagenomic analysis resulted in millions of new genes and showed an abundance of diversity in even small oceanic regions.

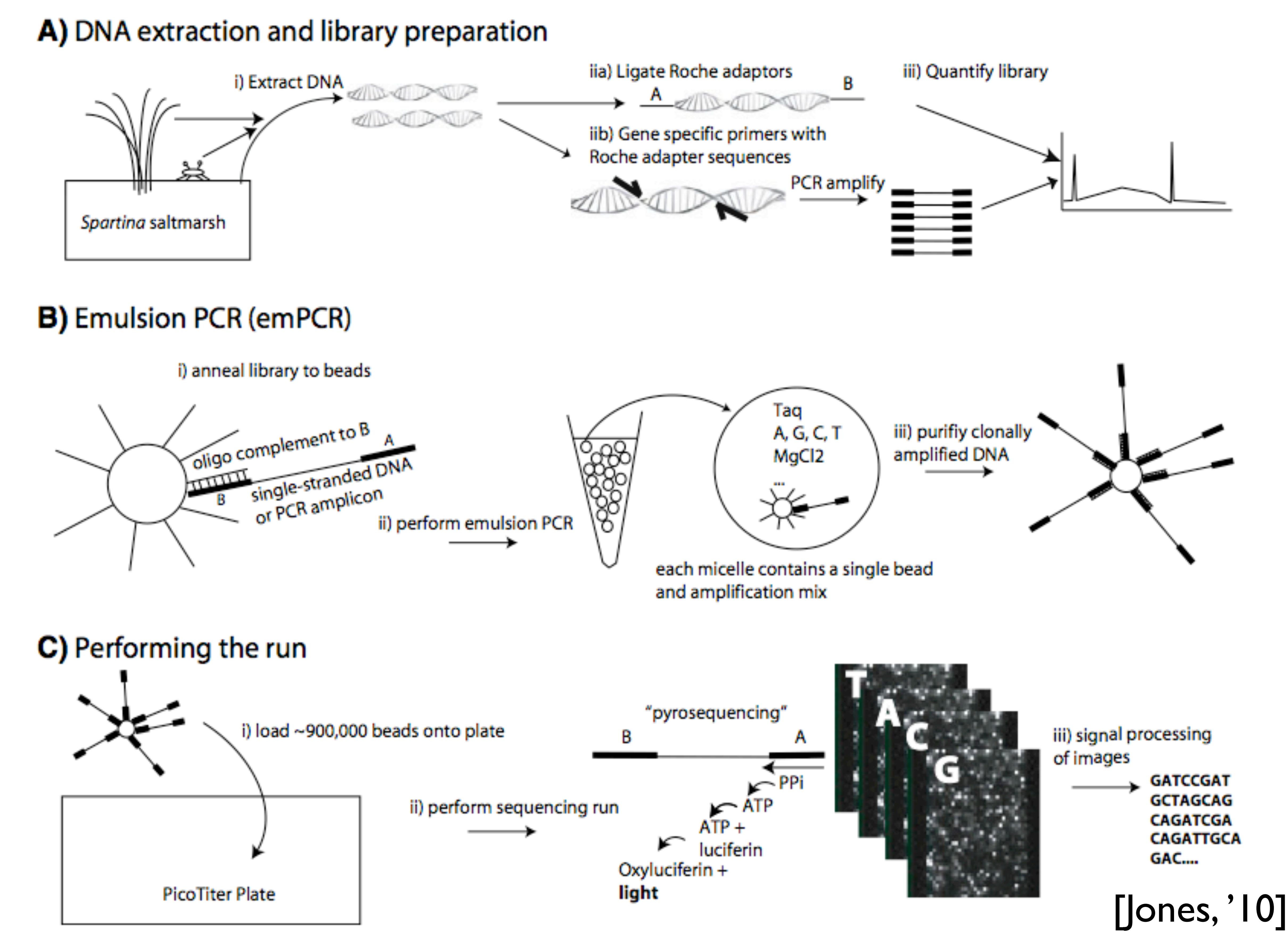
Rhodopsin



receptors
excited
11
11-cis retinal
all-trans retinal
Enzymes, Energy,
Vitamin A
O
O
light
FAST
11-cis retinal
+
opsin = rhodopsin
opsin
SLOW
enzymes
energy
vitamine A
opsin
retinal



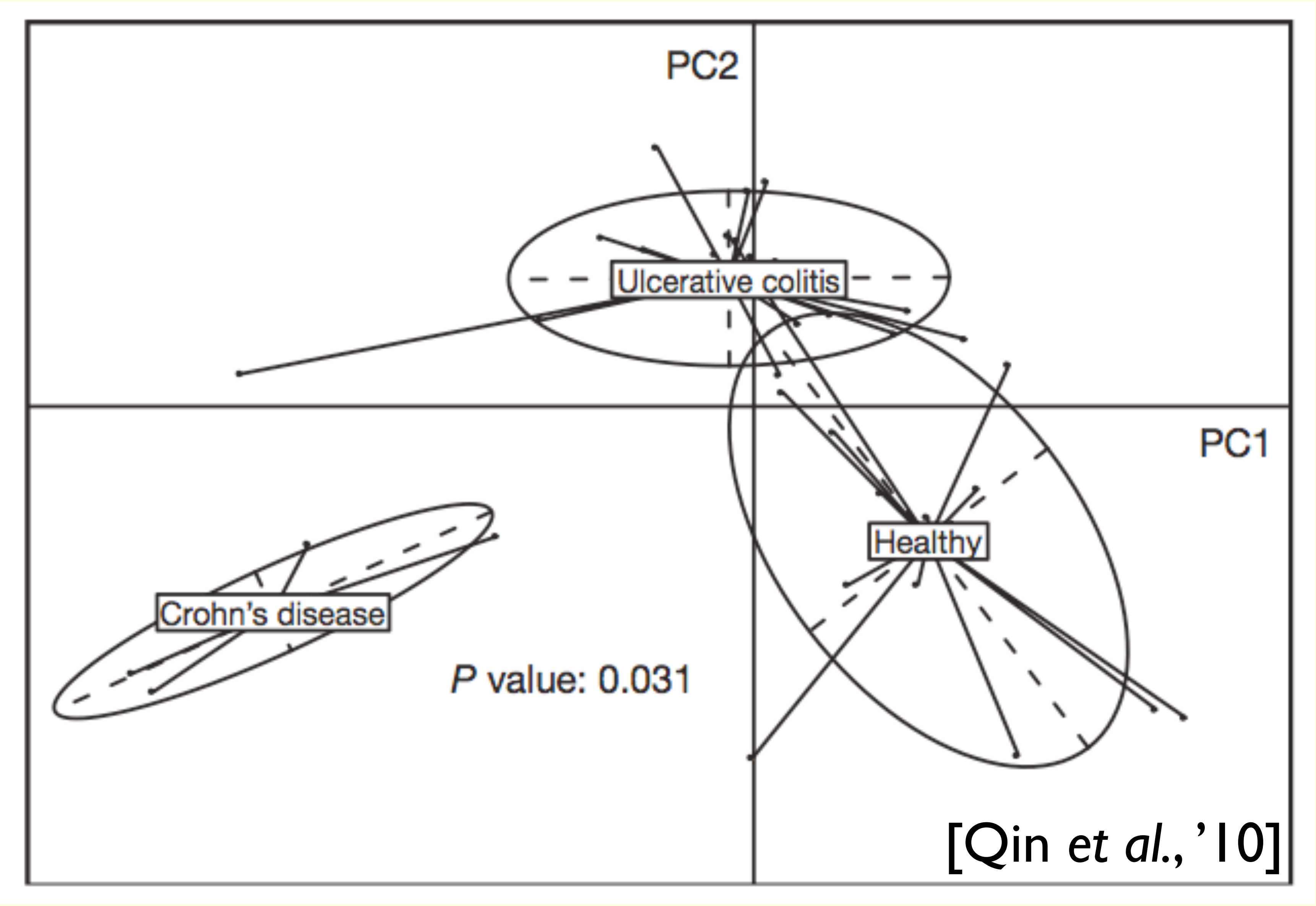
There is also an abundance of viral diversity in aquatic samples. [Venter et al, ’08]



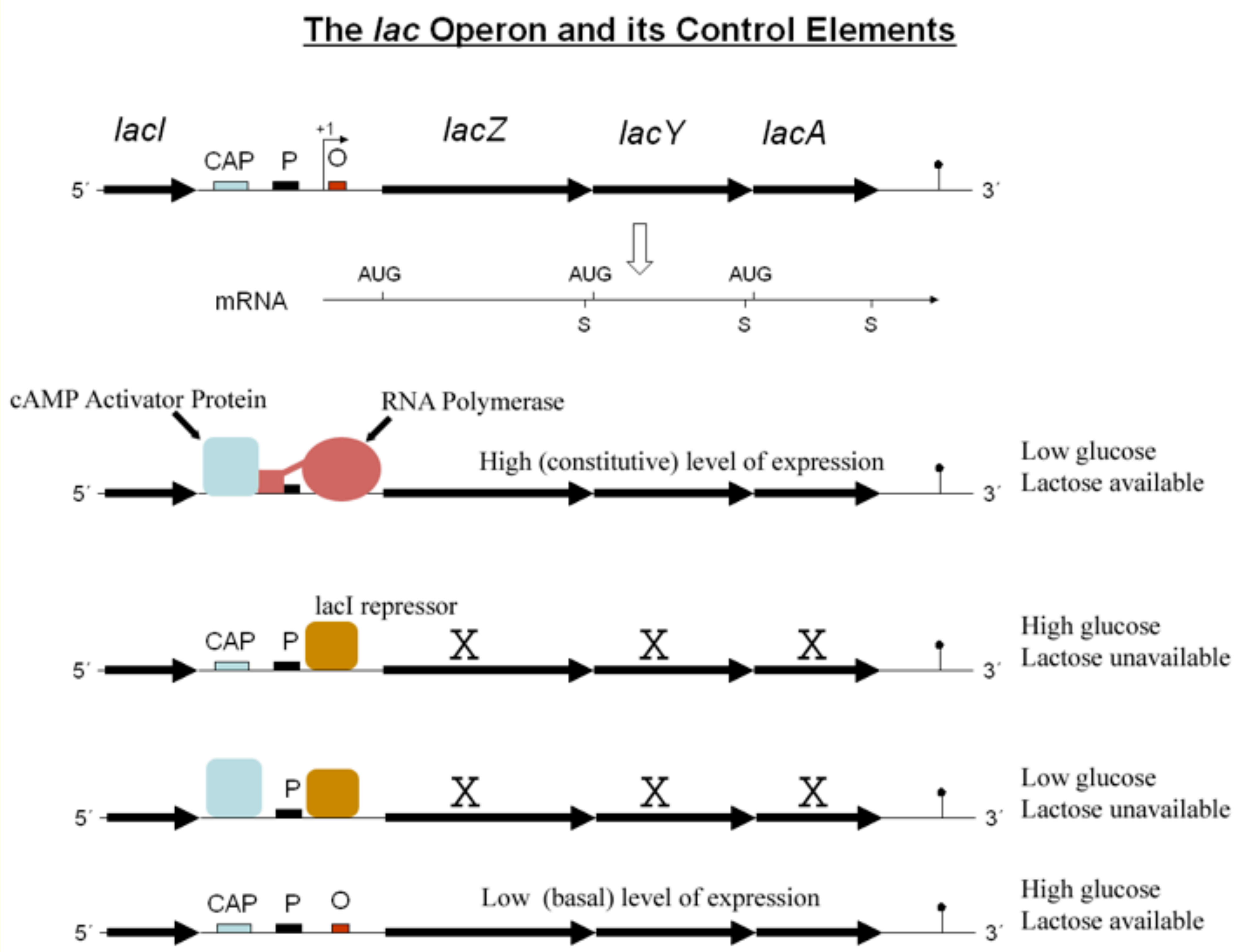
We can perform very rapid genomic analysis of environmental samples due to the parallel nature of high-throughput sequencers.

Microbiomics

The human gut is a complex ecosystem consisting of numerous bacteria and viruses - can it give us insights into disease?



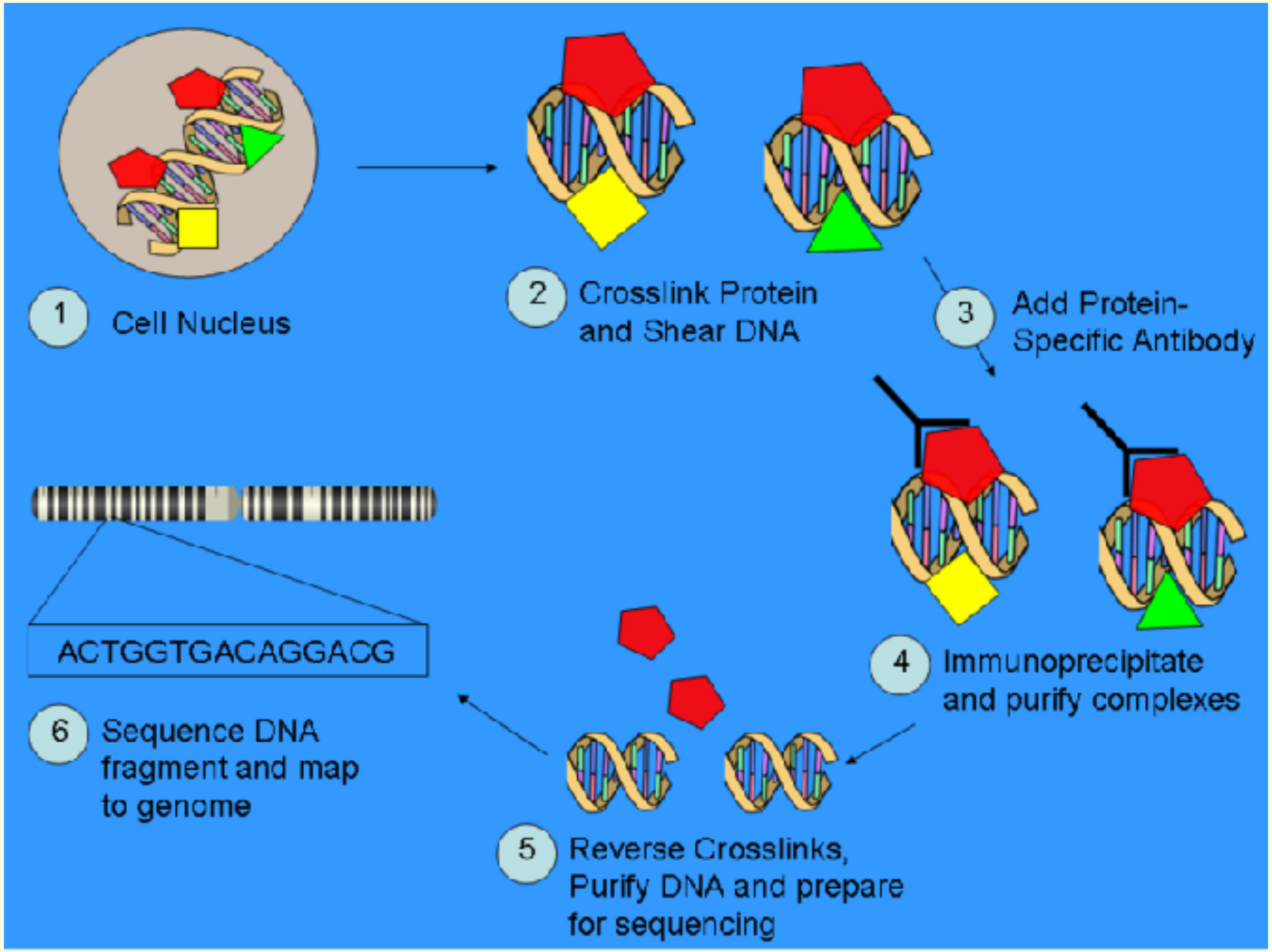
Metagenomic analysis on gut flora shows that we can classify patients based on GI disorders. KEGG analysis can then be performed to reveal functional differences.



Wikipedia

Jacques Monod discovered the Lac operon (1950) which controls beta-galactosidase production in *E. coli*.

Gene Regulation and HTS



Sequencing DNA regulatory elements is more accurate than hybridization assays. Moreover, HTS can be used to examine other regulatory (i.e. epigenetic) aspects of DNA sequence.

modENCODE

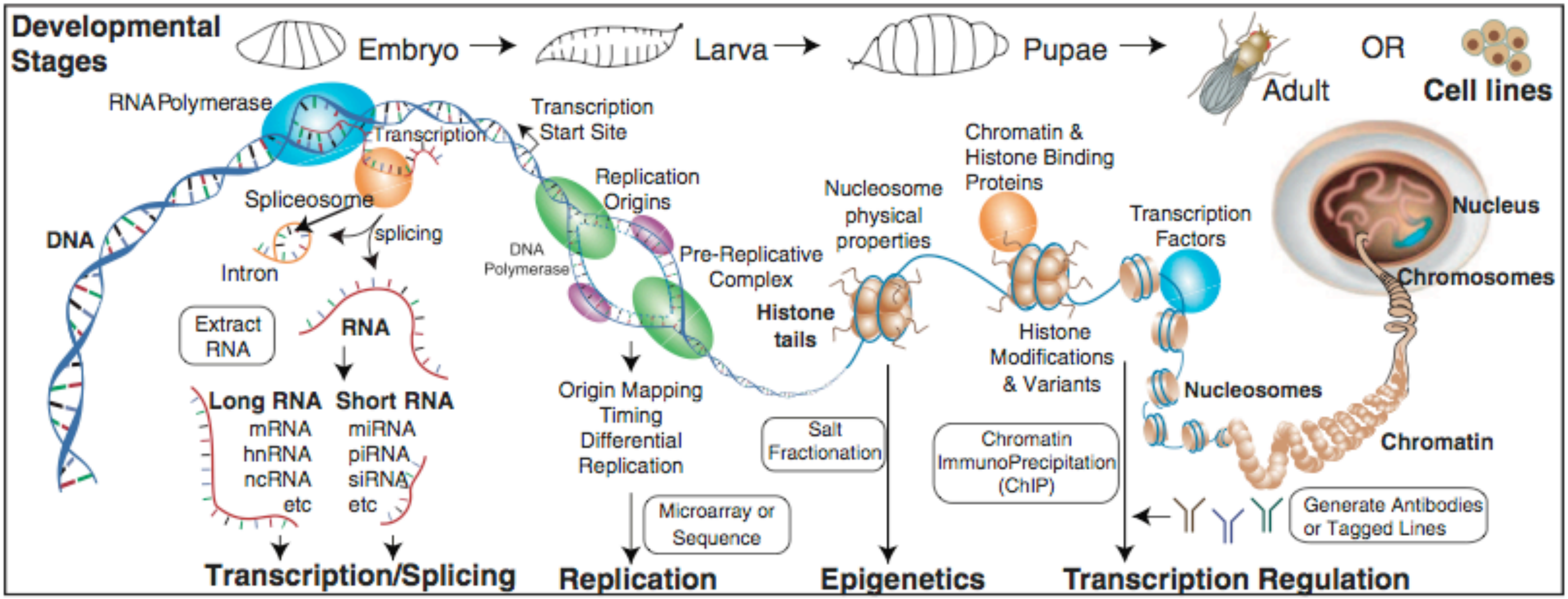


Fig. 1. Overview of *Drosophila* modENCODE data sets. Range of genomic elements and trans factors studied, with relevant techniques and resulting genome annotations. hnRNA, heterogeneous nuclear RNA.

[modEncode consortium *al.*, '10]

The *Drosophila* genome has been extensively studied -- nearly every gene has been mapped for splicing and regulation.

modENCODE

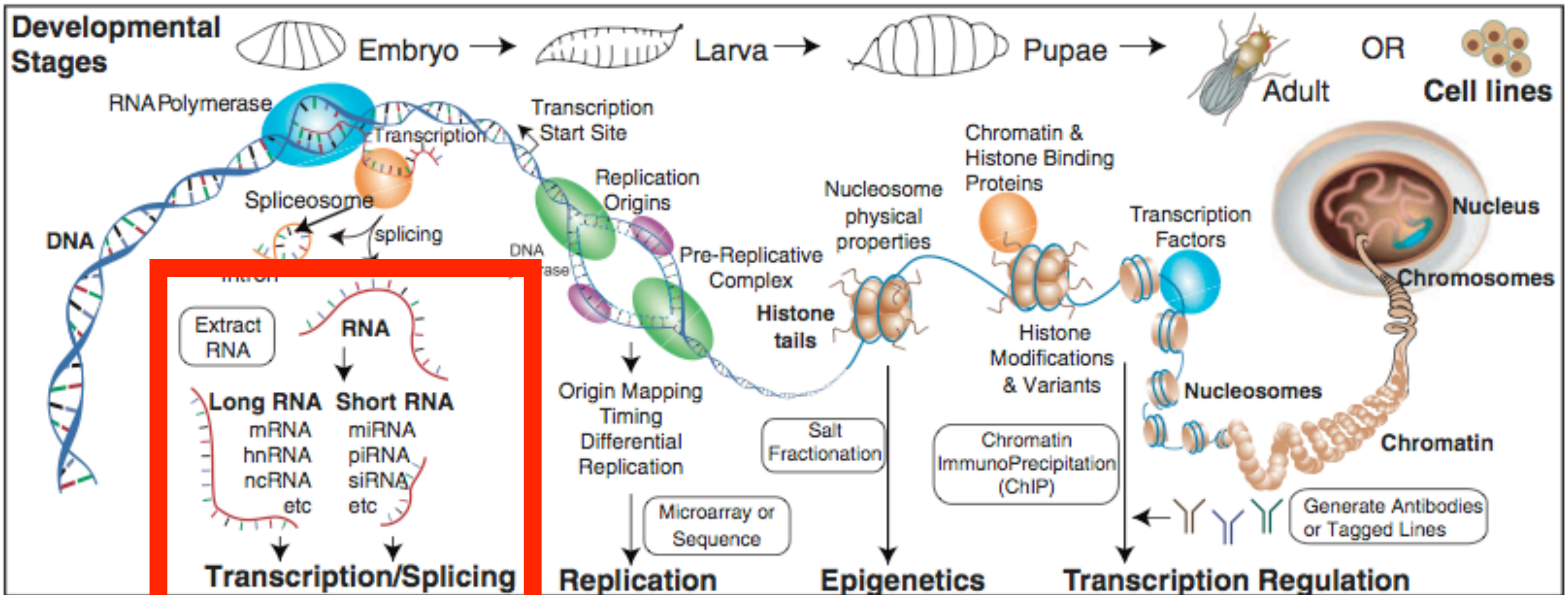
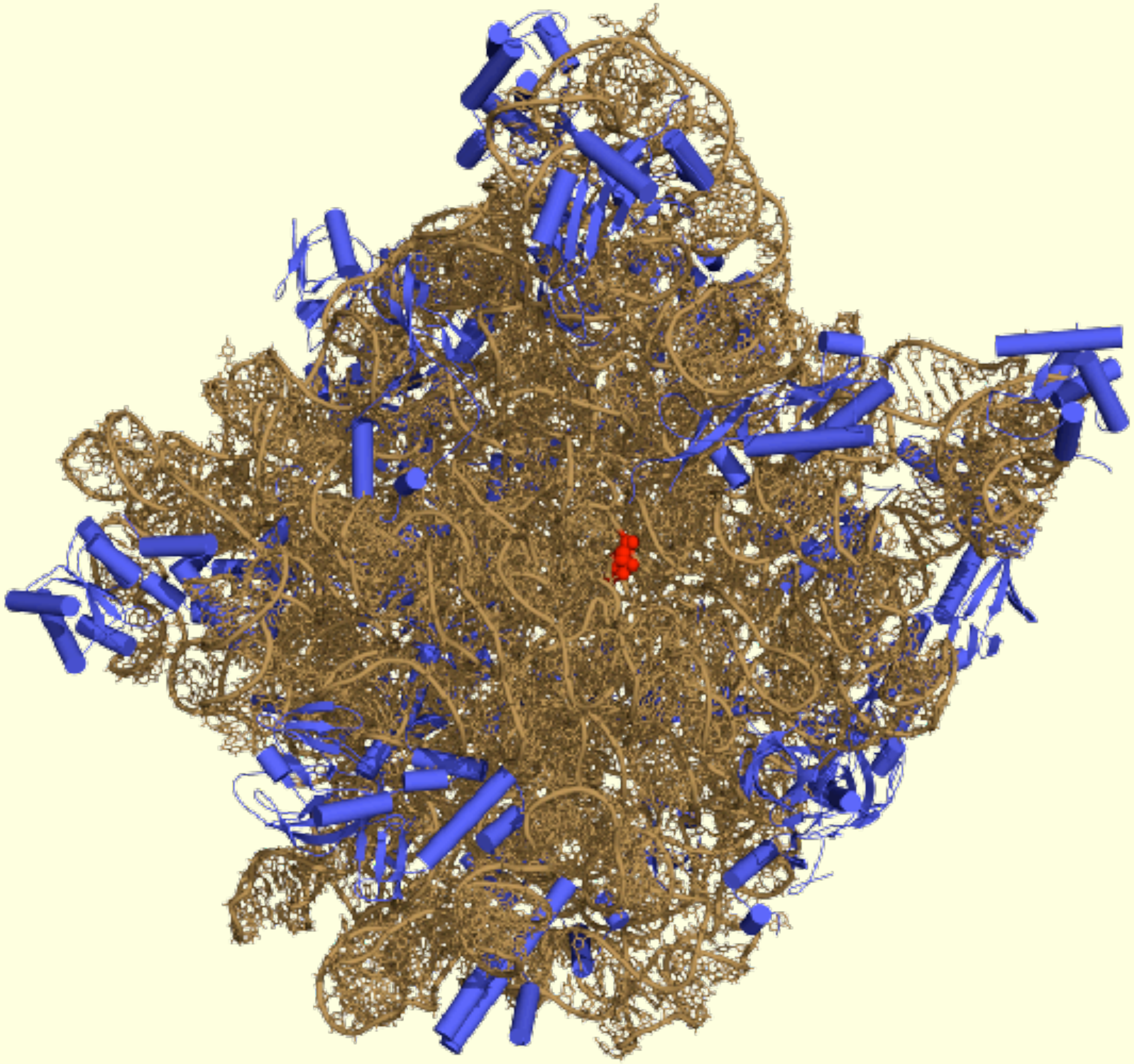


Fig. 1. Overview of *Drosophila* modENCODE data sets. Range of genomic elements and trans factors studied, with relevant techniques and resulting genome annotations. hnRNA, heterogeneous nuclear RNA.

[modEncode consortium *al.*, '10]

The *Drosophila* genome has been extensively studied -- nearly every gene has been mapped for splicing and regulation.

Regulatory RNA

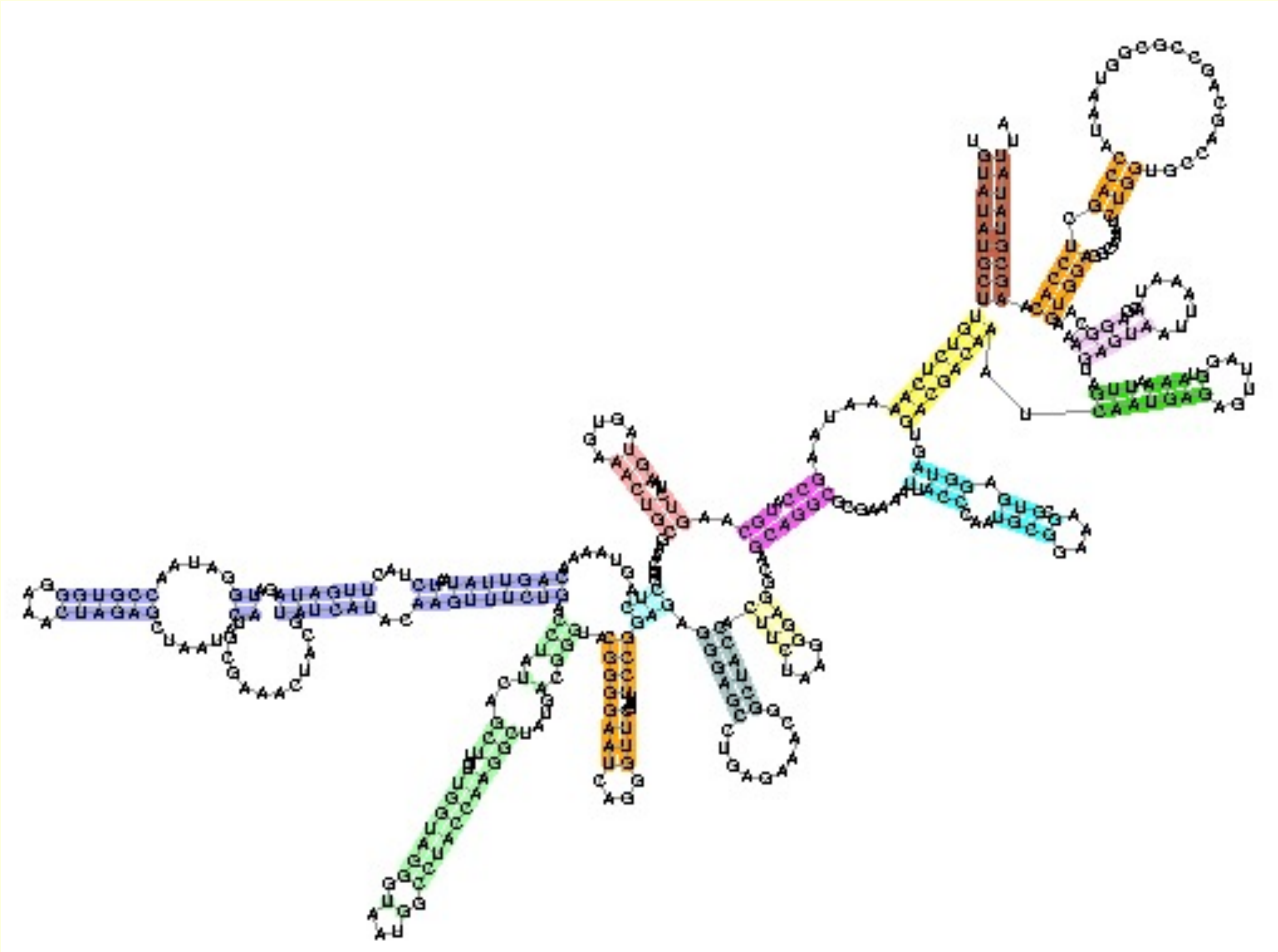


Ribosome

transfer RNA

RNA is clearly essential to gene expression -- it is a key component in protein synthesis.

RNA Structure



RNA has structure that is defined by complementarity. Given an RNA sequence, can you determine the 2D structure (using dynamic programming)?

RNA interference

- Historically, post-transcriptional silencing was observed in a number of settings. Mello and Fire (1998) showed that these phenomena could be explained by "RNA interference."
- They injected sense, anti-sense and double-stranded RNA into *C. elegans* and showed a method for <u>controlling</u> gene expression.
- They showed that even just a few molecules of <u>double-stranded</u> RNA could suppress gene expression in a cell.

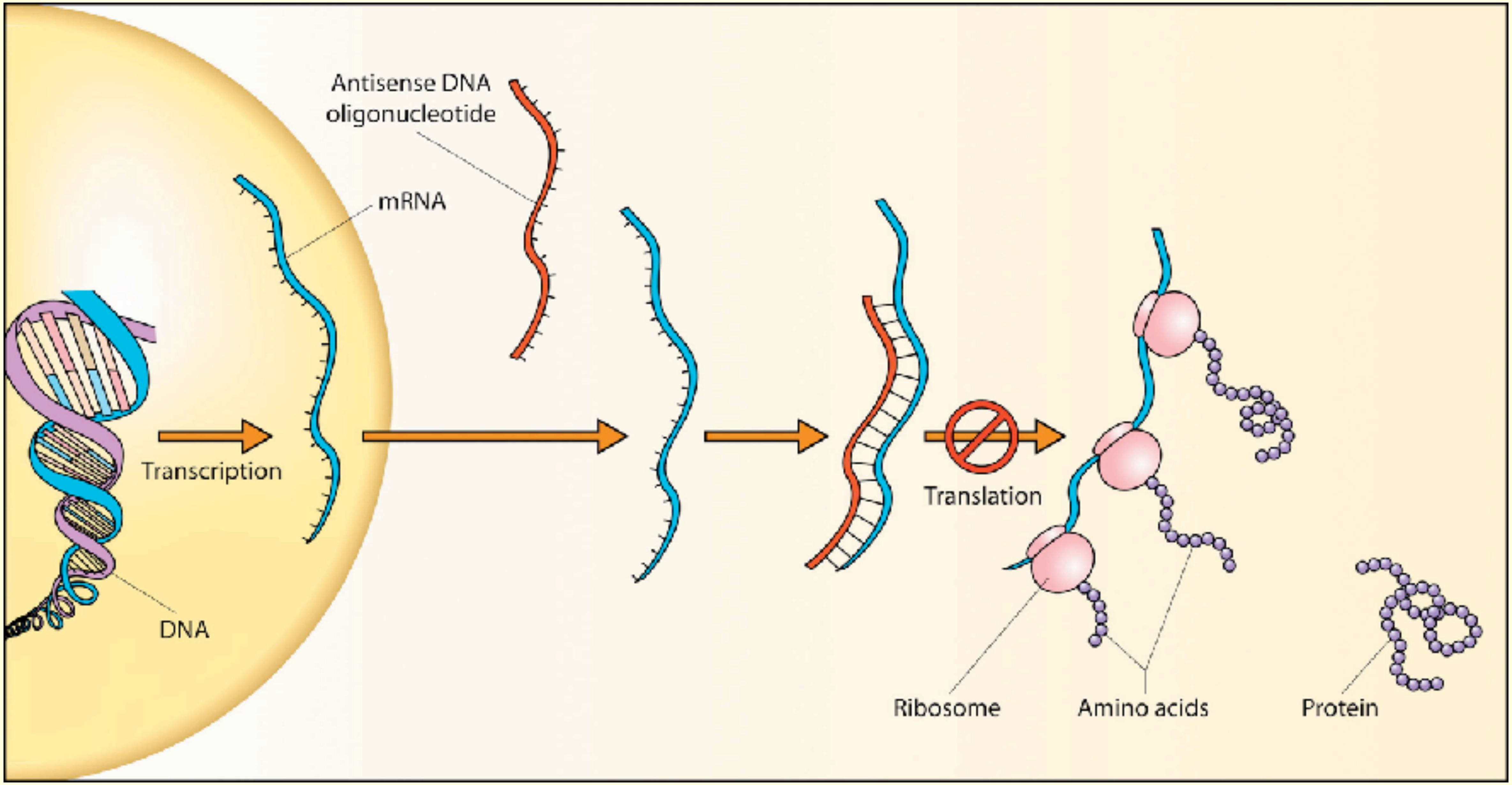
RNA interference

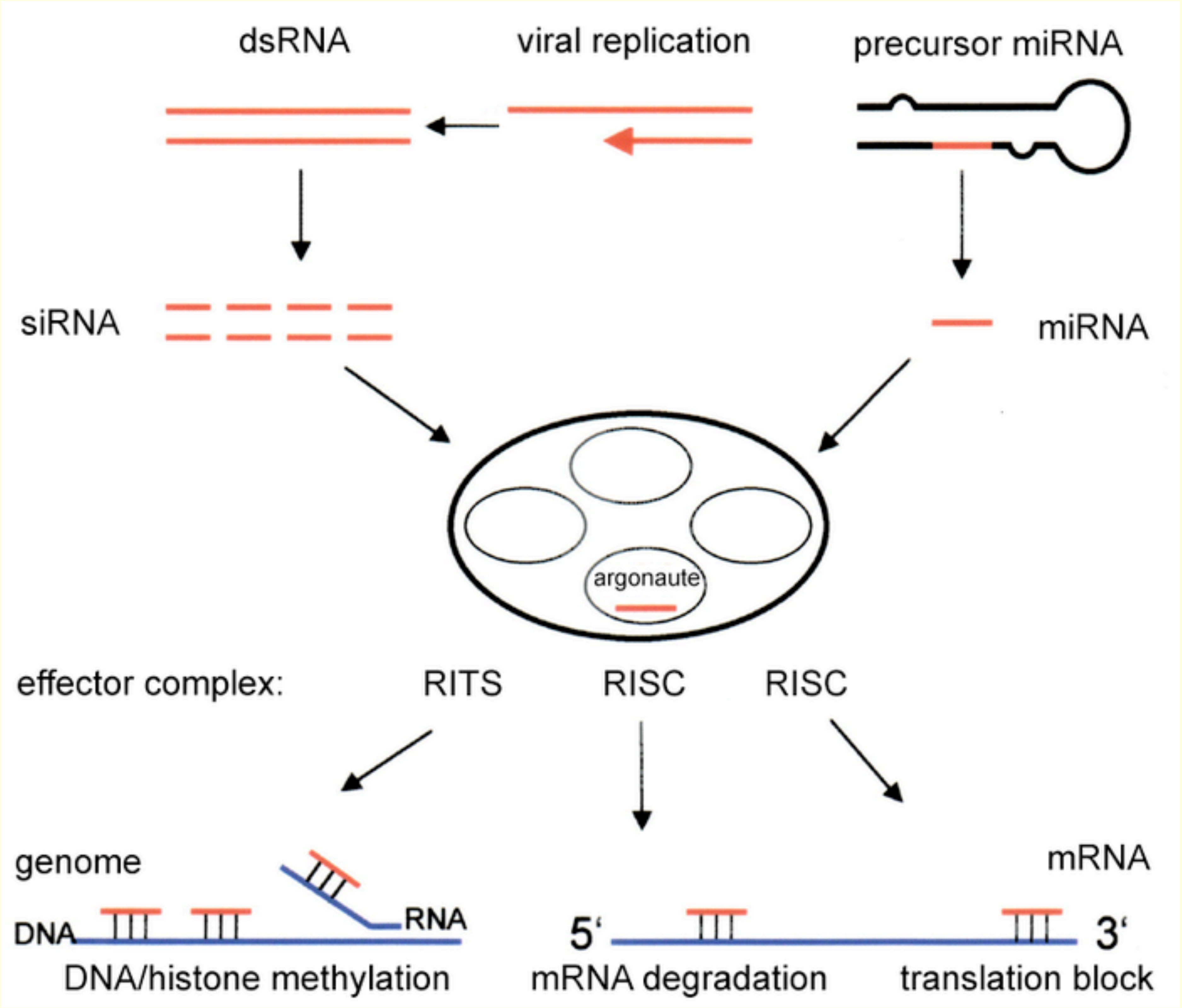
- Subsequent to the seminal work by Mello and Fire, the mechanisms of action have been further elucidated.
- The *Dicer* enzyme cleaves double-stranded RNA (dsRNA) to produce miRNAs.
- miRNAs are recognized by the "RNA-induced silencing complex" (RISC), which in turn cleaves complementary RNAs.
- In a sense, this process can be viewed as having the <u>opposite</u> effect of PCR.

Regulatory RNA

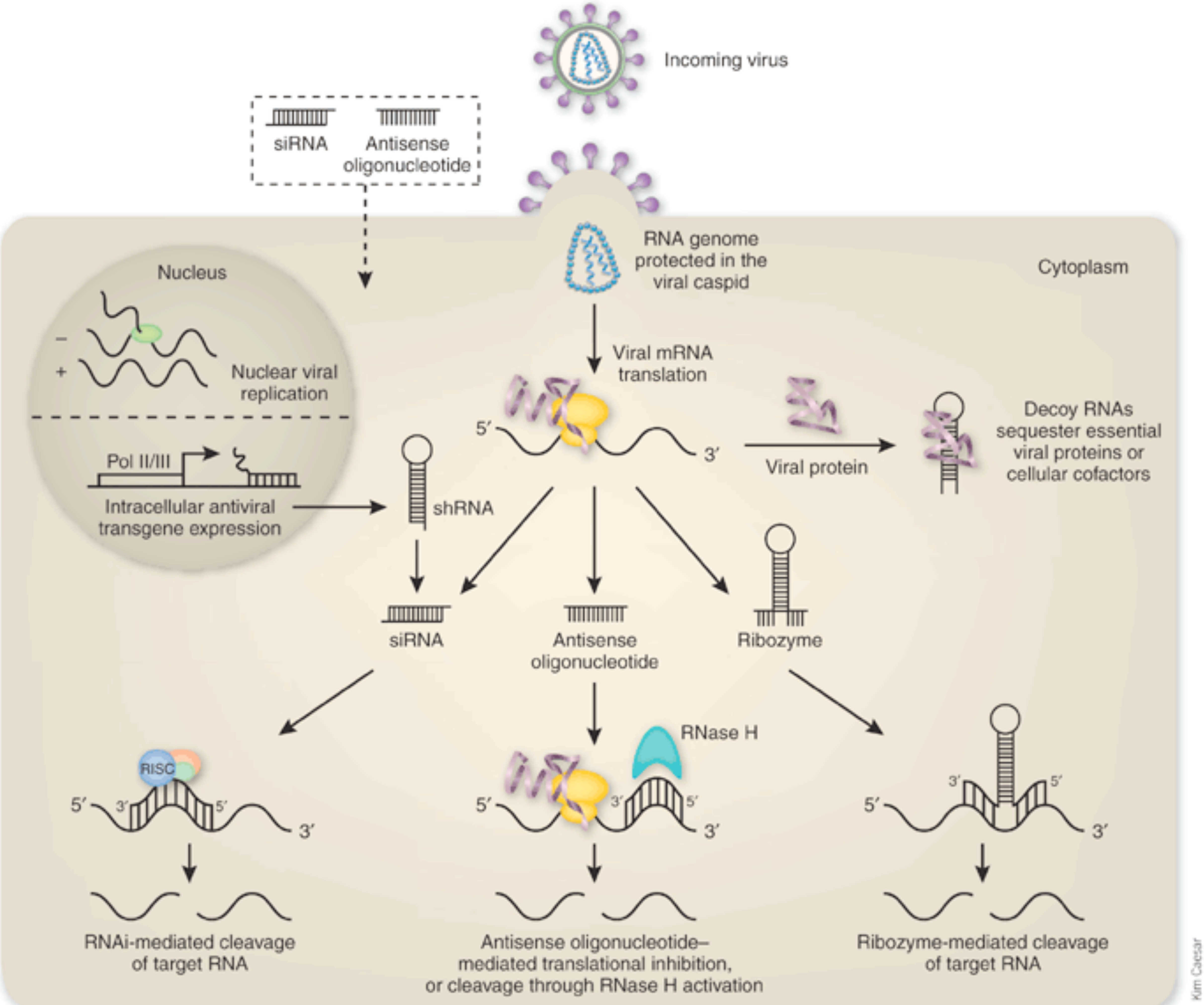
- RNAi = “RNA interference”
- miRNA = “microRNA”
- siRNA = “short interfering RNA” or “silencing RNA”
- dsRNA = “double-stranded RNA”

Antisense Regulation



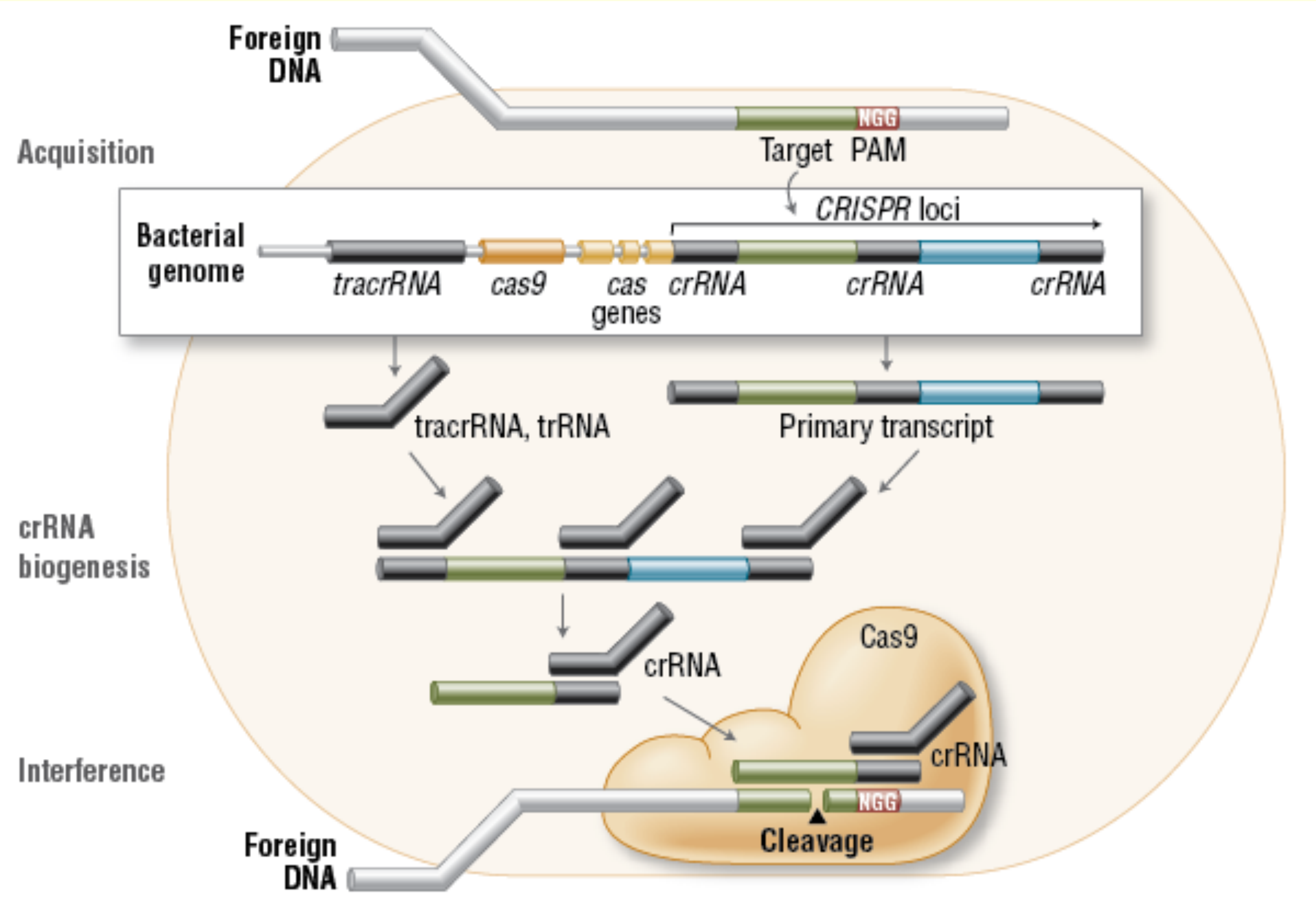


Some miRNAs are thought to have a protective effect against viral proliferation (endogenous or otherwise).



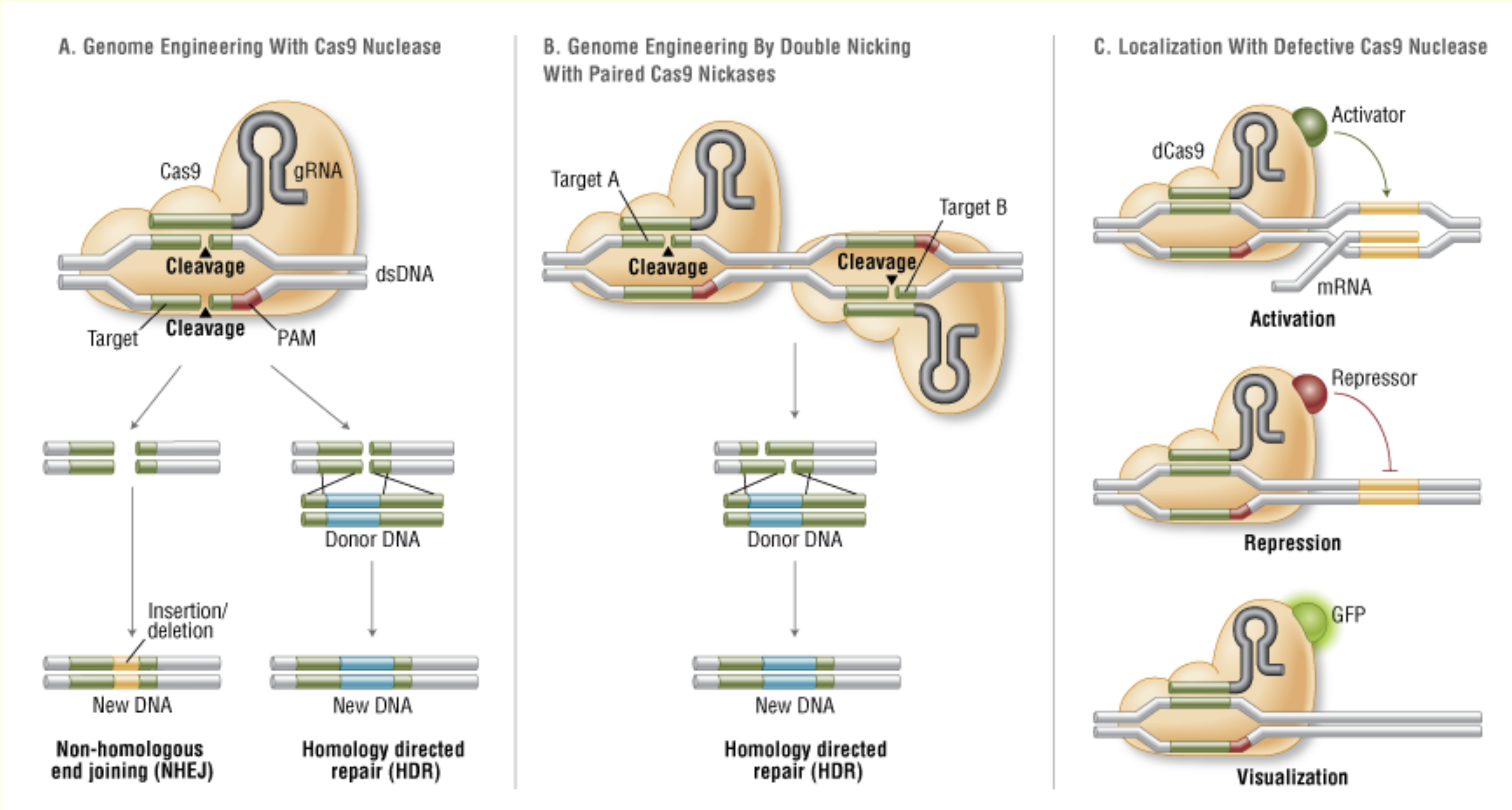
Incoming virus
siRNA
Antisense oligonucleotide
RNA genome protected in the viral caspid
Cytoplasm
Nucleus
Nuclear viral replication
Pol II/III
Intracellular antiviral transgene expression
shRNA
Viral mRNA translation
5′
3′
Viral protein
Decoy RNAs sequester essential viral proteins or cellular cofactors
siRNA
Antisense oligonucleotide
Ribozyme
RNase H
RISC
RNAi-mediated cleavage of target RNA
Antisense oligonucleotide–mediated translational inhibition, or cleavage through RNase H activation
Ribozyme-mediated cleavage of target RNA
Kim Caesar

CRISPR/Cas9



CRISPR/Cas9 was originally studied in the context of bacterial immunity. Great video on adaptation to gene editing by inventor.

CRISPR/Cas9



The CRISPR/Cas9 system is generally recognized as a breakthrough in gene editing, and can be used for a variety of tasks.

More RNA regulation

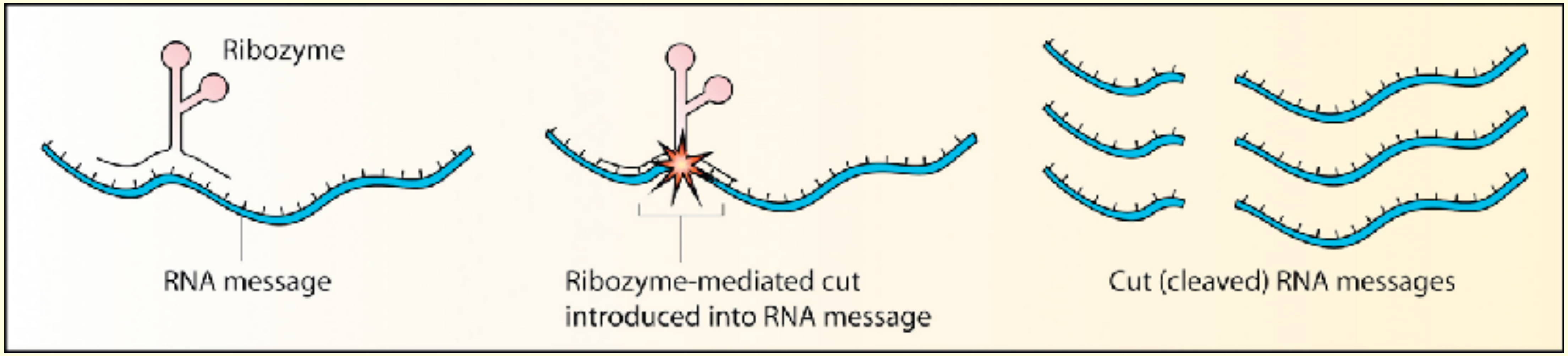
Lysine riboswitch

A "riboswitch" is an element contained in mRNA that binds a small molecule.

Riboswitches are usually consist of an "aptamer" that performs small-molecule recognition, and an "expression platform" that regulates gene expression.

The structure of the riboswitch in the "apo" versus "holo" controls expression, and can either up- or down-regulate a gene.

Even more RNA function




Hammerhead ribozyme (X-ray)

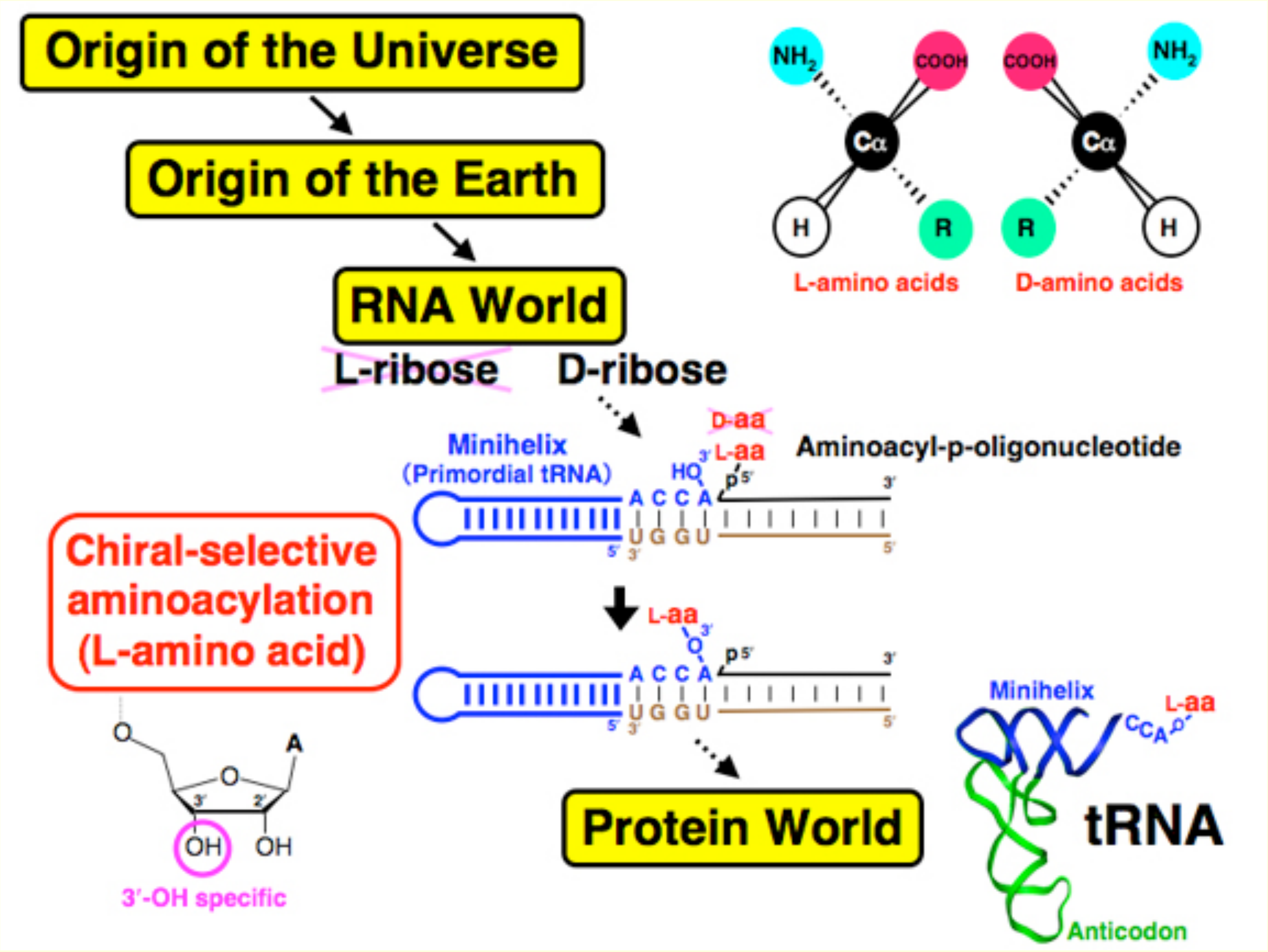
RNA structures, or "ribozymes" can in fact have catalytic activity. The Nobel Prize for Chemistry was awarded to Cech and Altman in 1989 for this discovery.

Ribozymes fold so as to recognize specific RNA sequence and cleave it.

The RNA bundles in ribosomes are fact ribozymes that assist in protein synthesis.

Which came first? The enzymes that do the work of the cell, or the RNA that codes for it?

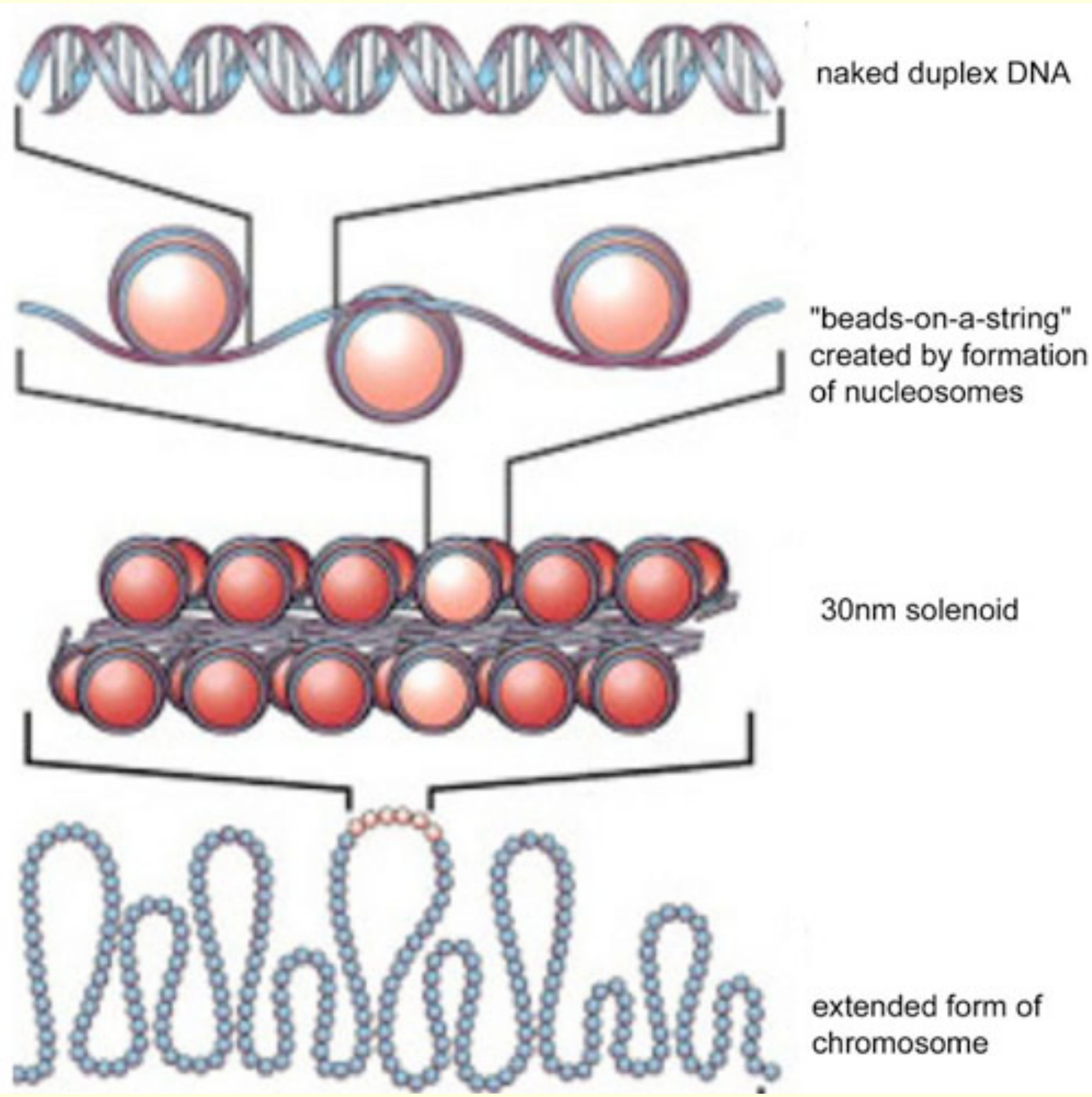
RNA World Hypothesis

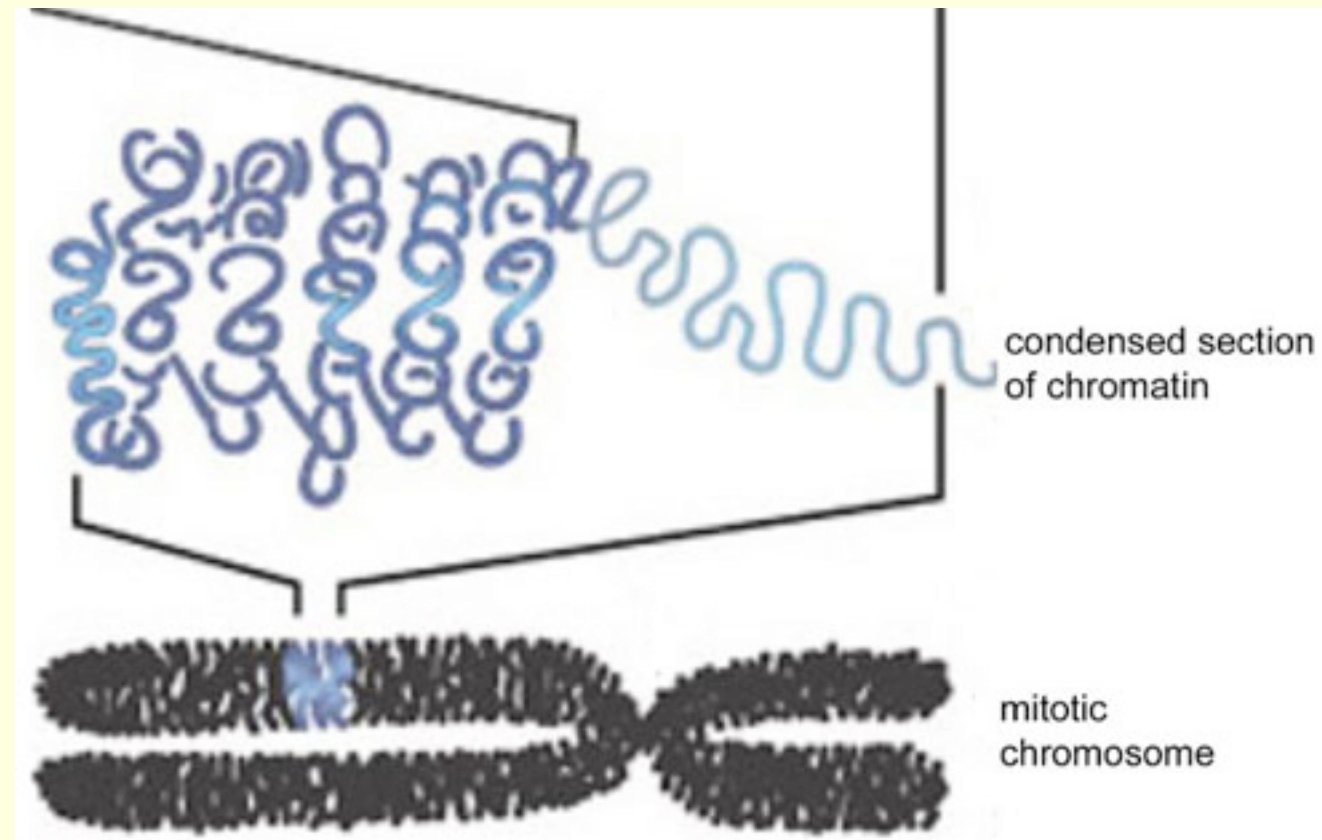


Epigenetics

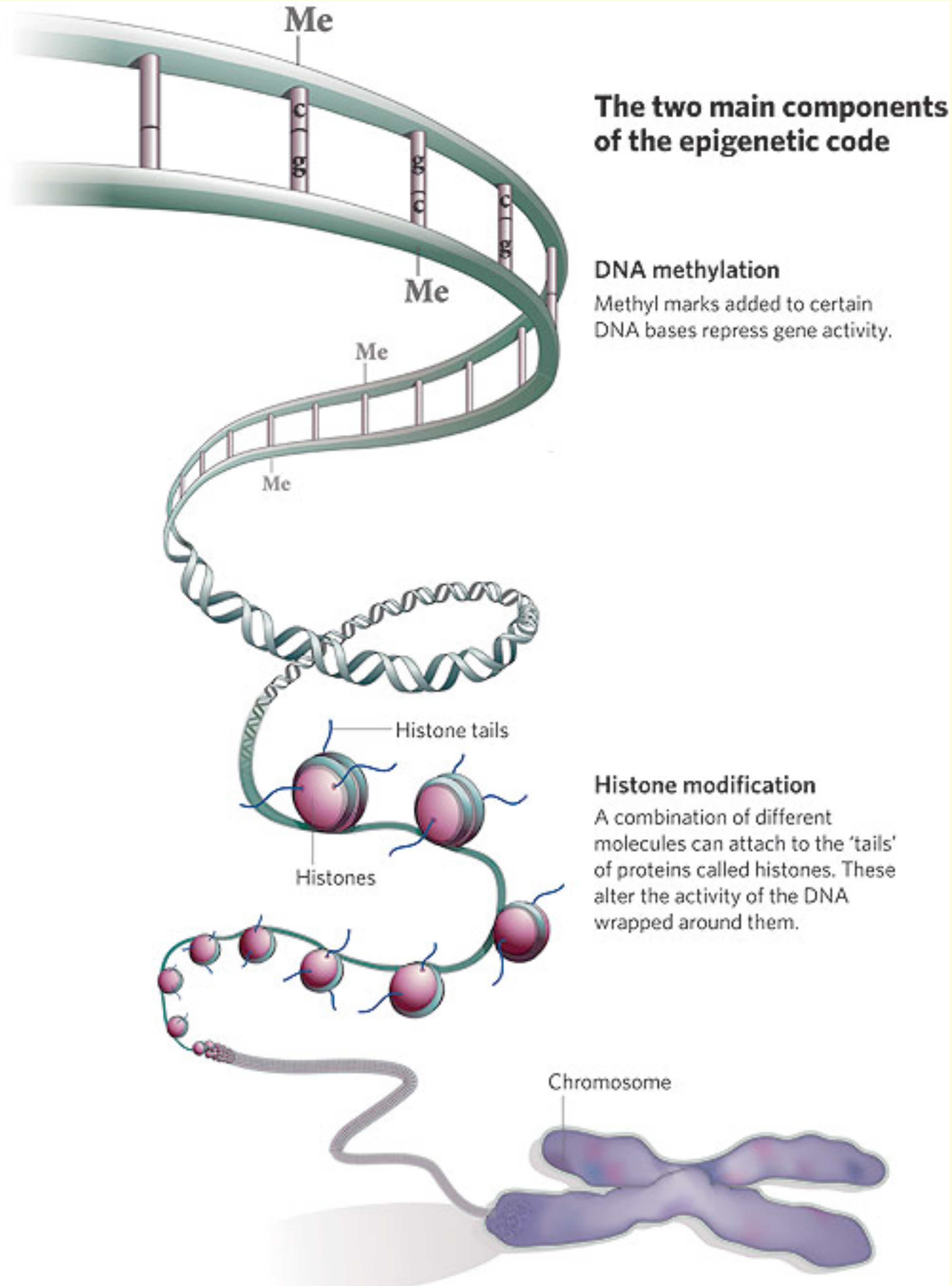
- **Epigenetics** is the study of how gene expression changes in response to external factors (disease, environment, etc.)
- It has been known for some time that environment plays a role in diseases such as cancer.
- What is the genetic/molecular mechanism that mediates gene expression in these diseases?

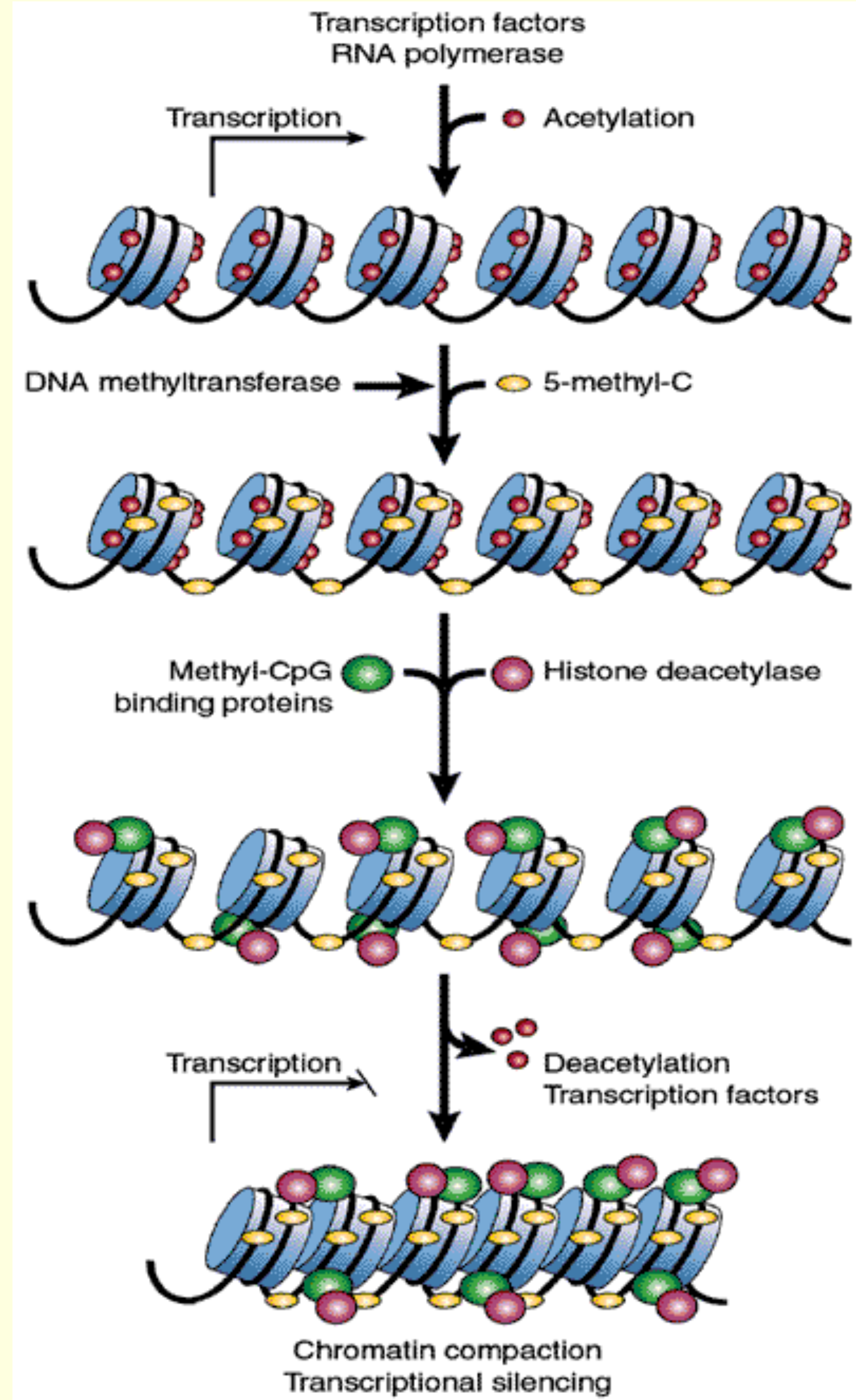
Chromosome Architecture

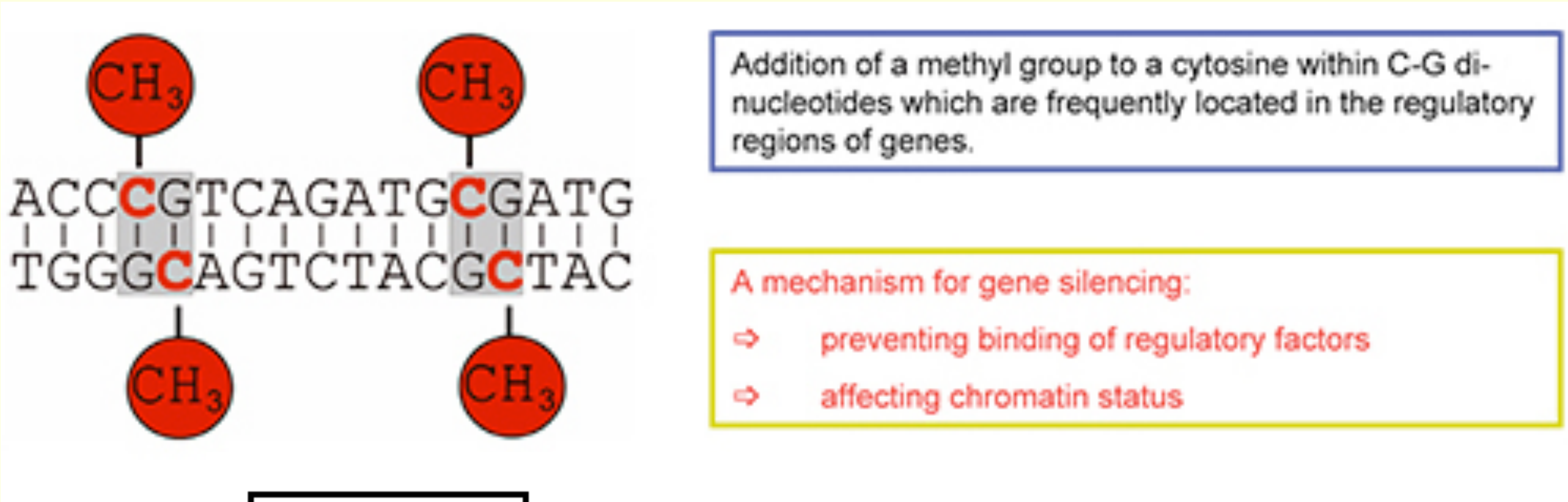


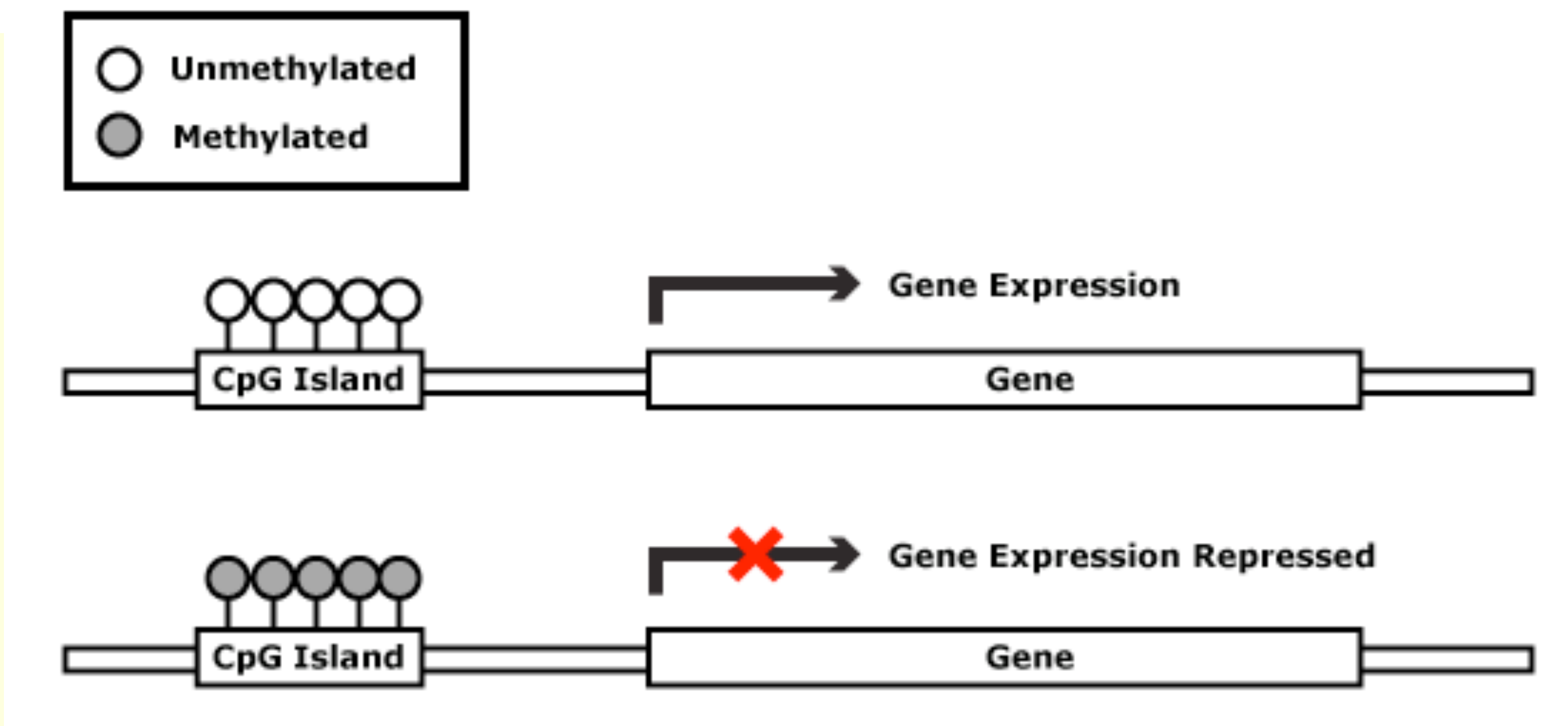


Video of Chromosome Packaging

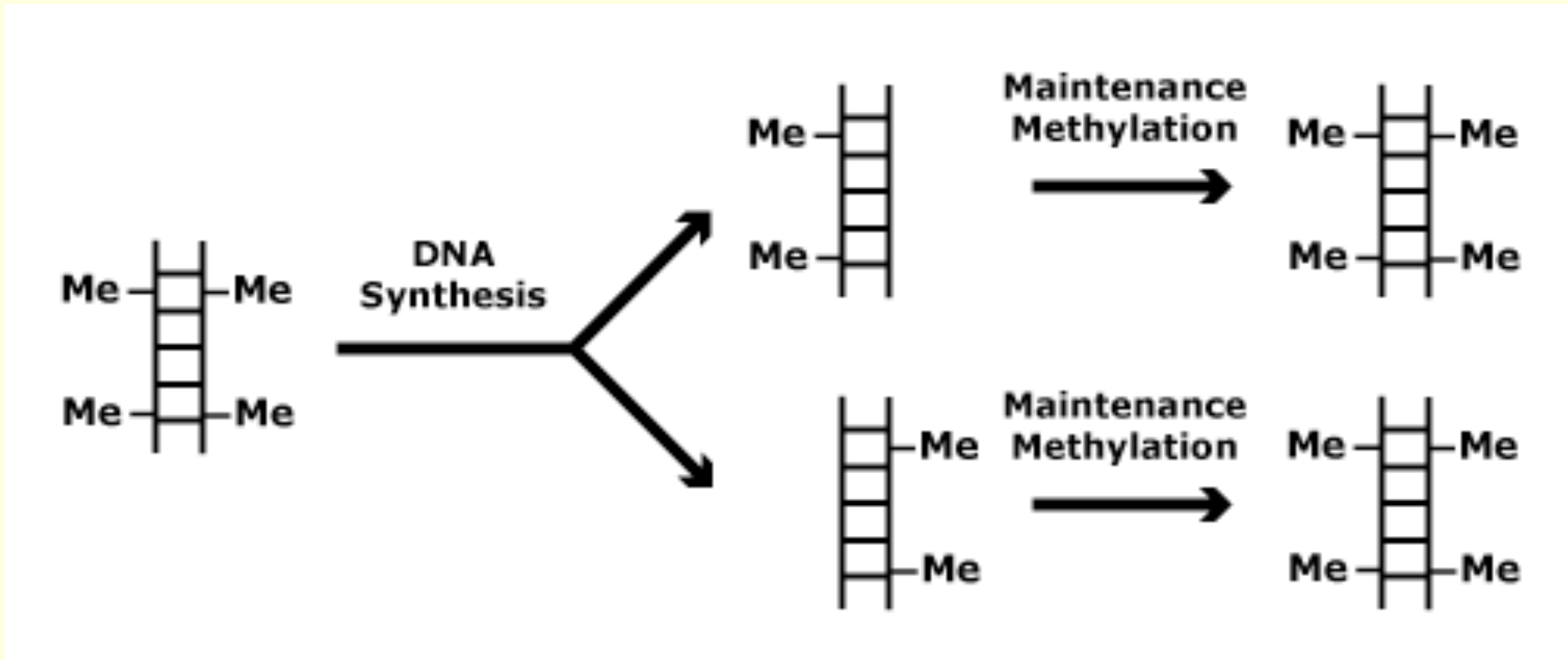

The two main components of the epigenetic code
DNA methylation
Methyl marks added to certain DNA bases repress gene activity.
Histone modification
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.
Me
Me
Me
Me
Histone tails
Histones
Chromosome


Transcription factors
RNA polymerase
Transcription
Acetylation
DNA methyltransferase
5-methyl-C
Methyl-CpG
binding proteins
Histone deacetylase
Transcription
Deacetylation
Transcription factors
Chromatin compaction
Transcriptional silencing





<u>DNA methylation</u> appears to be a basic regulatory mechanism for turning genes on and off. CpG islands are regions of DNA in which Cytosine (adjacent to a Guanine) can be methlyated to silence a downstream gene.



DNA methylation is <u>maintained</u>, suggesting the possibility of a mechanism for adaptation. Methylation also occurs *de novo*, suggesting a mechanism for disease processes.

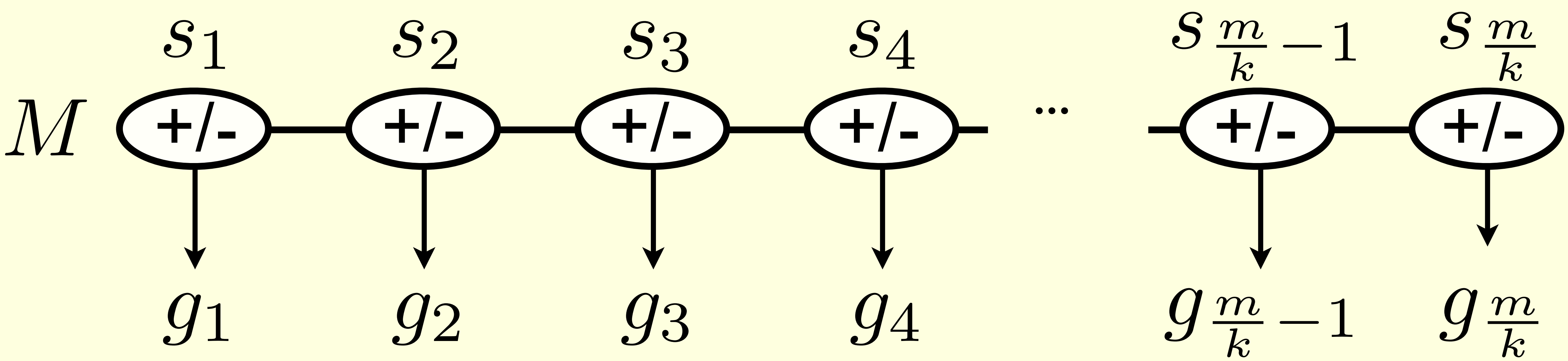
Table 1. Epigenetic Aberrations among Different Tumor Types.*

Type of Cancer	Epigenetic Disruption
Colon cancer	CpG-island hypermethylation (*hMLH1*, *p16^{INK4a}*, *p14^{ARF}*, *RARB2*, *SFRP1*, and *WRN*), hypermethylation of miRNAs (*miR-124a*), global genomic hypomethylation, loss of imprinting of *IGF2*, mutations of histone modifiers (*EP300* and *HDAC2*), diminished monoacetylated and trimethylated forms of histone H4
Breast cancer	CpG-island hypermethylation (*BRCA1*, E-cadherin, *TMS1*, and estrogen receptor), global genomic hypomethylation
Lung cancer	CpG-island hypermethylation (*p16^{INK4a}*, *DAPK*, and *RASSF1A*), global genomic hypomethylation, genomic deletions of *CBP* and the chromatin-remodeling factor *BRG1*
Glioma	CpG-island hypermethylation (DNA-repair enzyme *MGMT*, *EMP3*, and *THBS1*)
Leukemia	CpG-island hypermethylation (*p15^{INK4b}*, *EXT1*, and *ID4*), translocations of histone modifiers (*CBP*, *MOZ*, *MORF*, *MLL1*, *MLL3*, and *NSD1*)
Lymphoma	CpG-island hypermethylation (*p16^{INK4a}*, *p73*, and DNA-repair enzyme *MGMT*), diminished monoacetylated and trimethylated forms of histone H4
Bladder cancer	CpG-island hypermethylation (*p16^{INK4a}* and *TPEF/HPP1*), hypermethylation of miRNAs (*miR-127*), global genomic hypomethylation
Kidney cancer	CpG-island hypermethylation (*VHL*), loss of imprinting of *IGF2*, global genomic hypomethylation
Prostate cancer	CpG-island hypermethylation (*GSTP1*), gene amplification of polycomb histone methyltransferase *EZH2*, aberrant modification pattern of histones H3 and H4
Esophageal cancer	CpG-island hypermethylation (*p16^{INK4b}* and *p14^{ARF}*), gene amplification of histone demethylase *JMJD2C/GASC1*
Stomach cancer	CpG-island hypermethylation (*hMLH1* and *p14^{ARF}*)
Liver cancer	CpG-island hypermethylation (*SOCS1* and *GSTP1*), global genomic hypomethylation
Ovarian cancer	CpG-island hypermethylation (*BRCA1*)

[Esteller ’08]

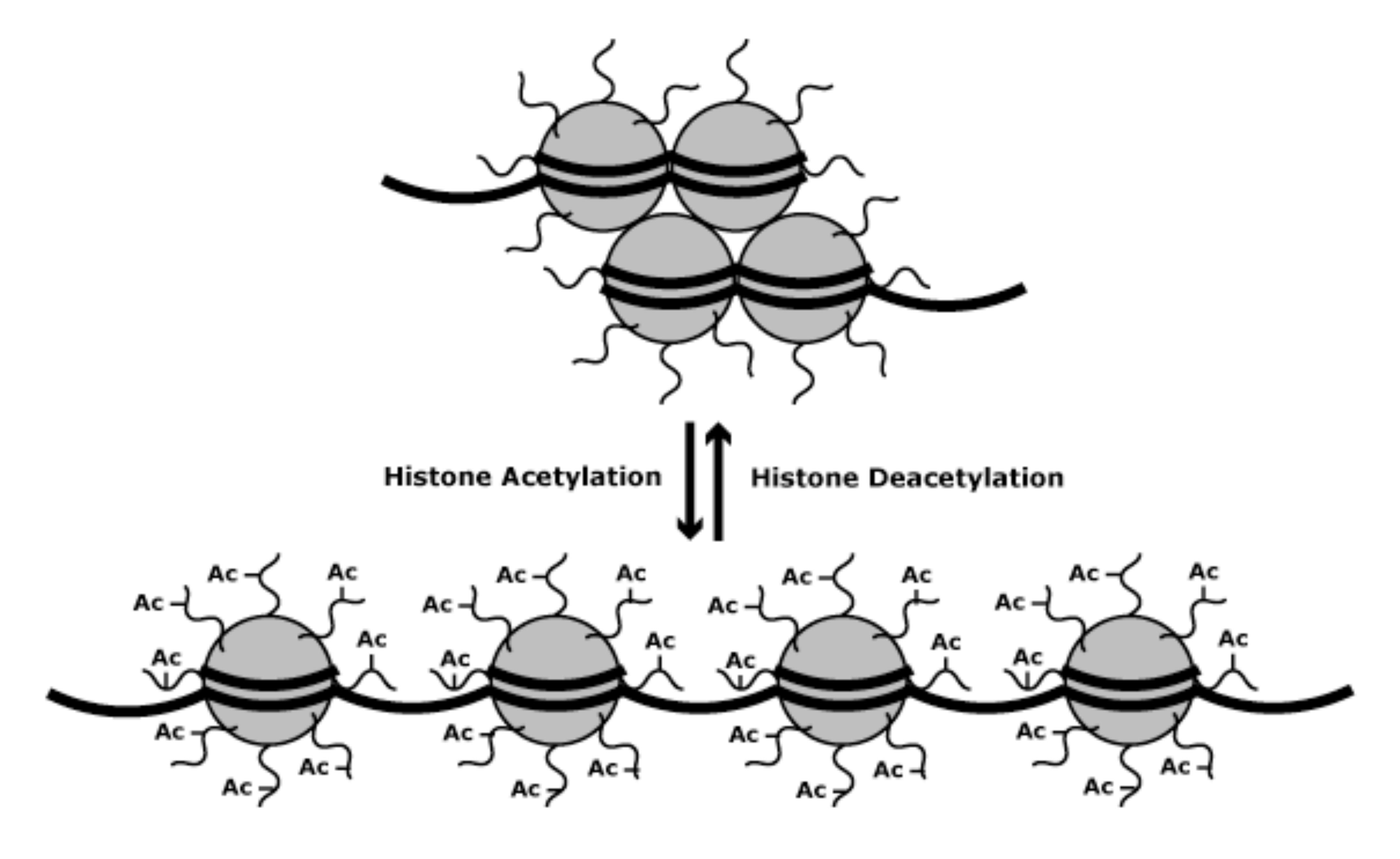
Finding CpG Islands

"Train" the HMM using known CpG islands. Then, given a new sequence, identify whether each nucleotide is in an island or not.



Each g_i represents a block (or single nucleotide) of sequence, and is annotated +/-. Then, blocks of "+" in the most likely state sequence give us the CpG islands.

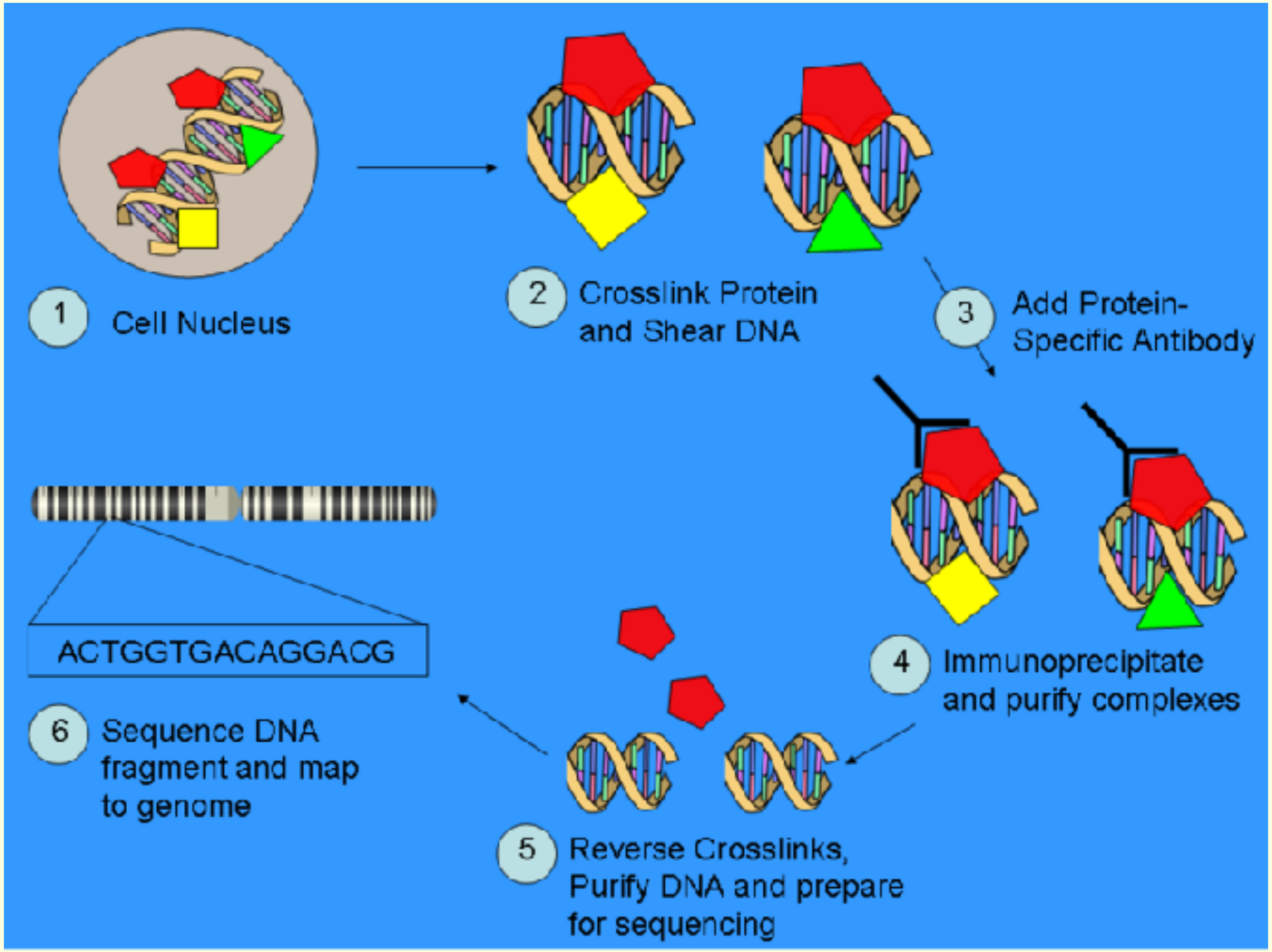
Euchromatin/Heterochromatin



Terminal regions of histone proteins are amenable to modifications which control whether DNA is accessible for transcription.

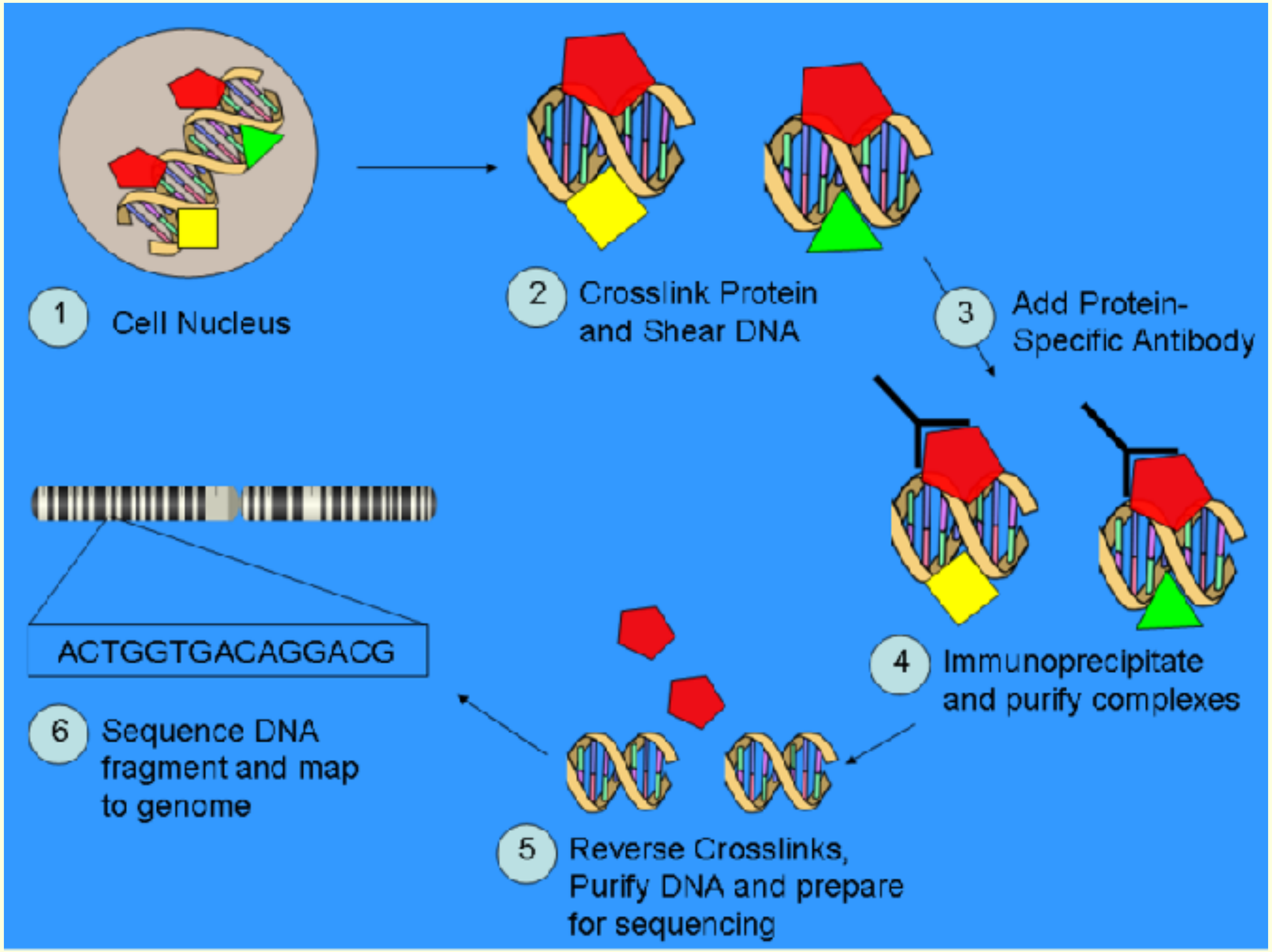
Enzymes (HAT, HDAC) add and remove acetyl groups to histone tails to make DNA more or less compact.

Gene Regulation and HTS

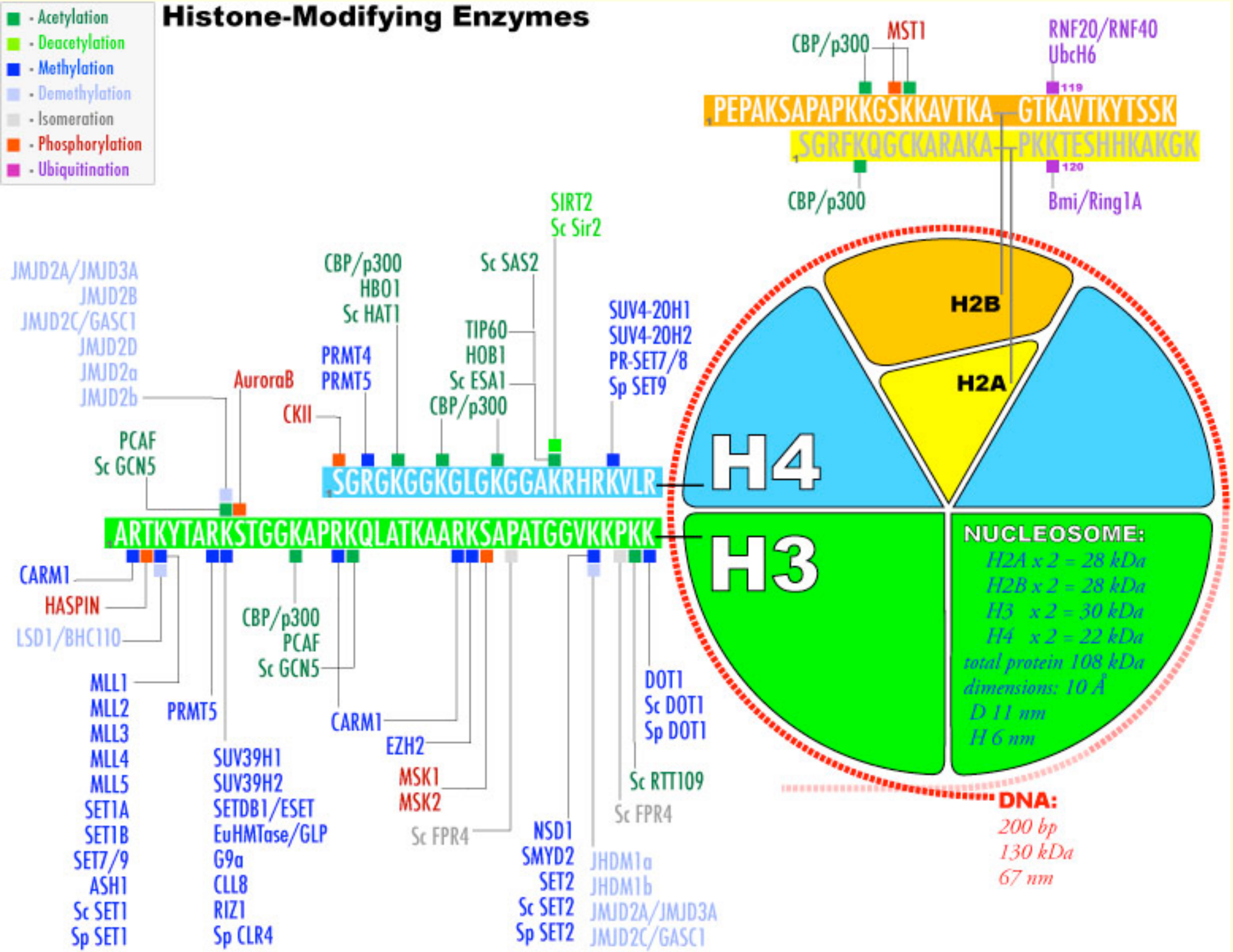


Sequencing DNA regulatory elements is more accurate than hybridization assays. Once antibodies ‘select’ for DNA-binding proteins, the resulting DNA can be sequenced and mapped.

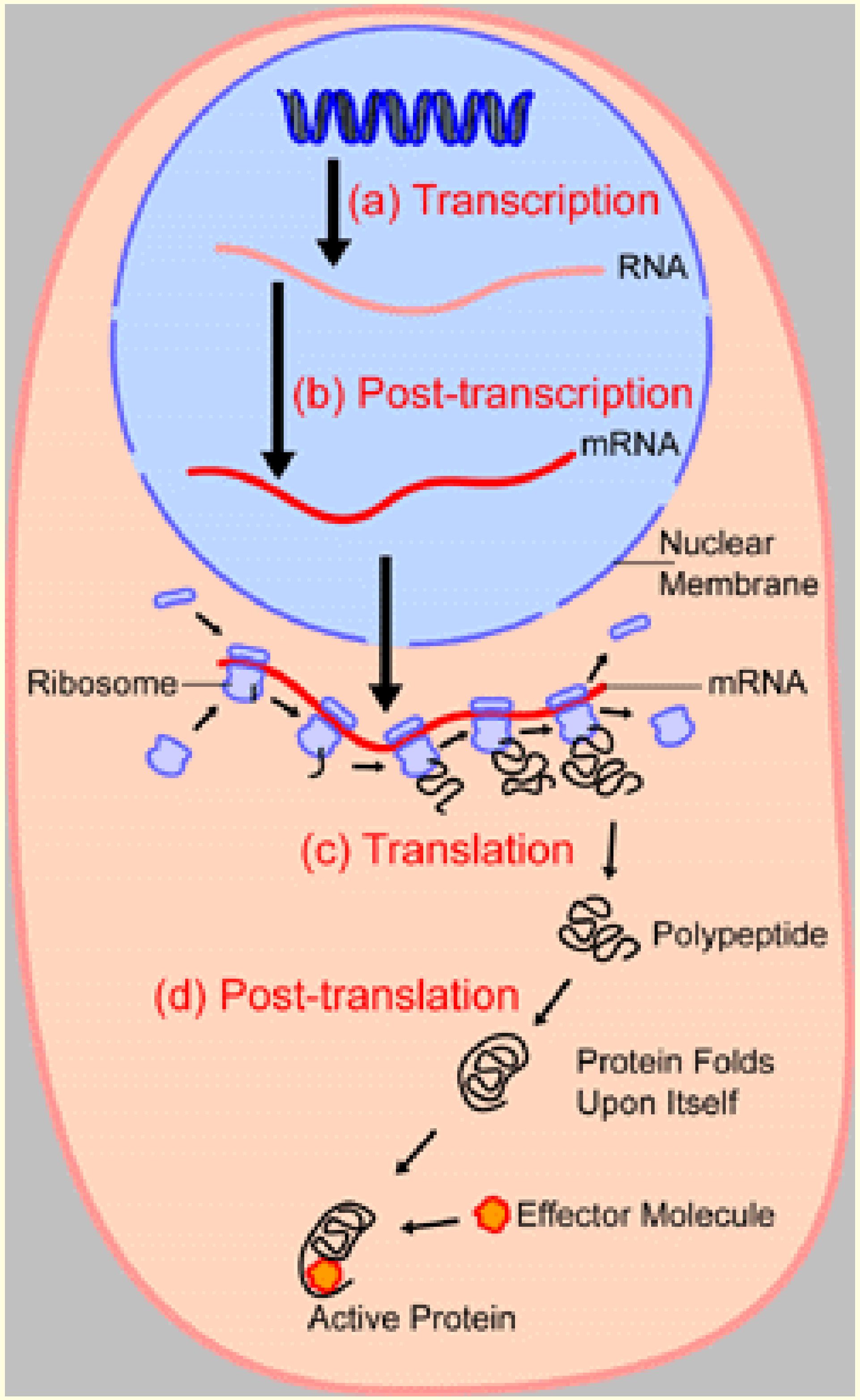
HTS and Chromatin “Marking”

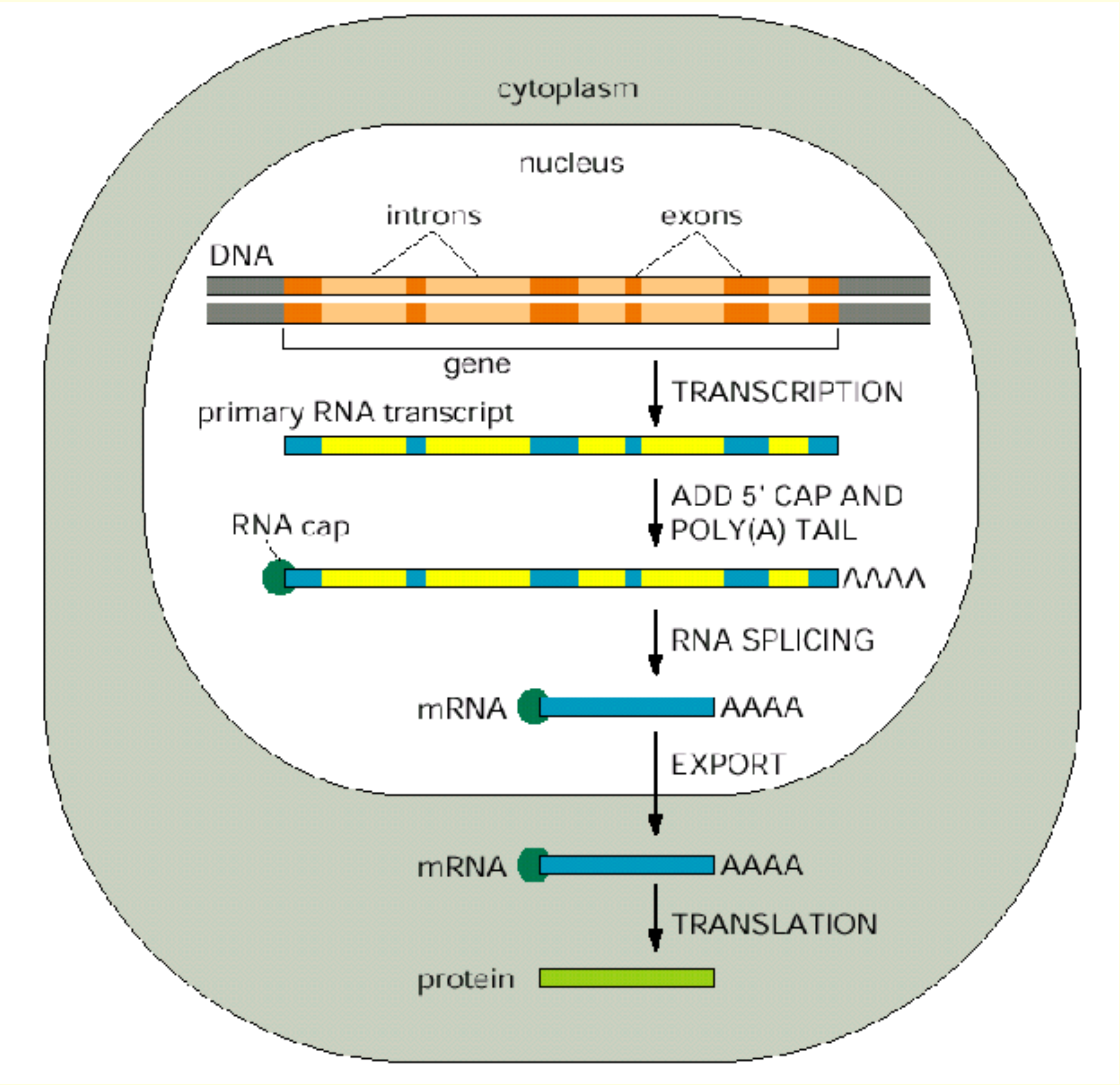


CHiP-seq can sequence bound DNA fragments. CHiP assays can select for modifications such as acetylation and methylation, and the resulting bound DNA indicates which genes are being silenced or activated.

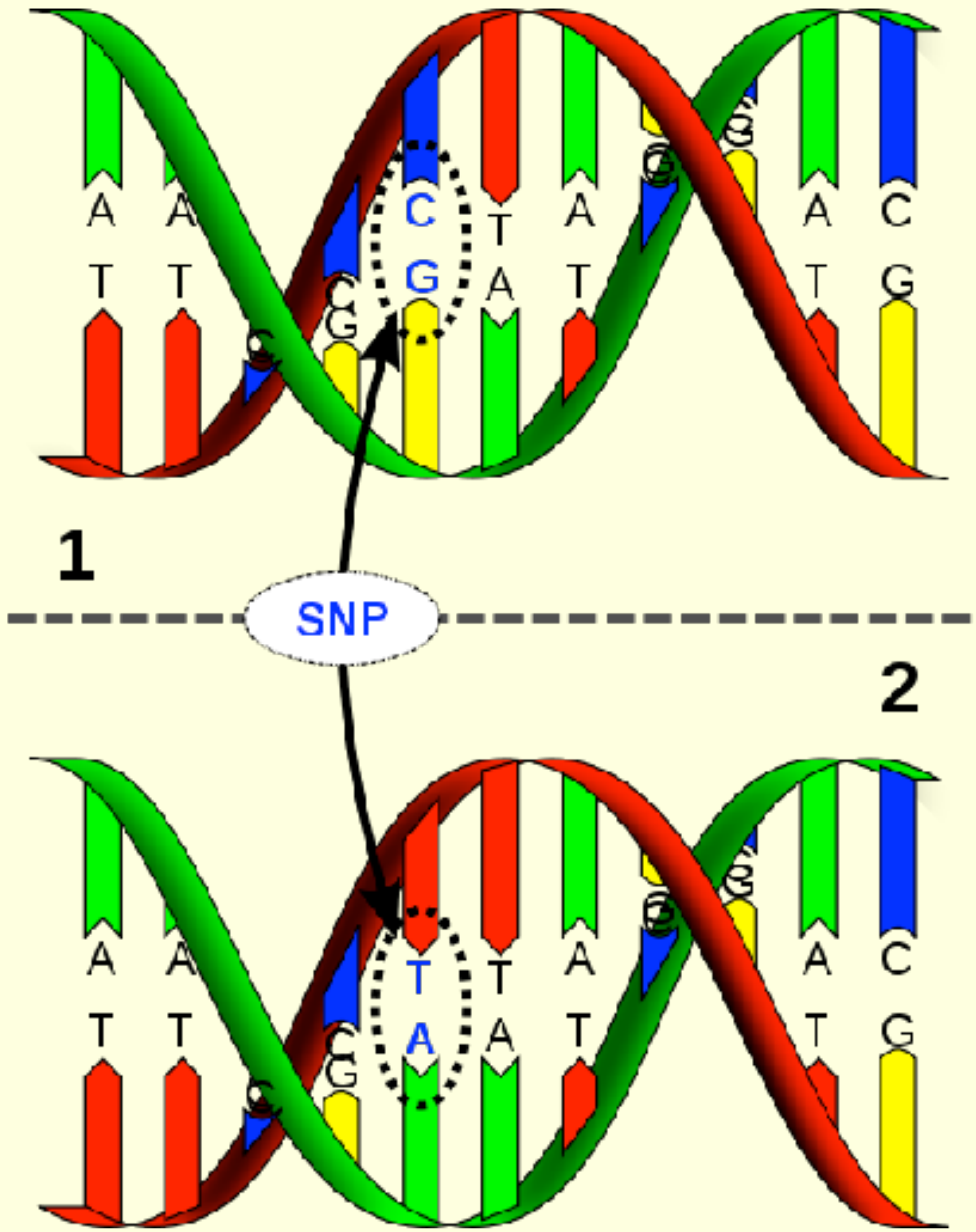

Histone-Modifying Enzymes
- Acetylation
- Deacetylation
- Methylation
- Demethylation
- Isomeration
- Phosphorylation
- Ubiquitination
CBP/p300
MST1
RNF20/RNF40
UbcH6
119
PEPAKSAPAPKKGSKKAVTKA
GTKAVTKYTSSK
SGRFKQGCKARAKA
PKKTESHHKAKGK
120
CBP/p300
Bmi/Ring1A
SIRT2
Sc Sir2
JMJD2A/JMJD3A
JMJD2B
JMJD2C/GASC1
JMJD2D
JMJD2a
JMJD2b
CBP/p300
HBO1
Sc HAT1
Sc SAS2
TIP60
HOB1
Sc ESA1
CBP/p300
PRMT4
PRMT5
AuroraB
CKII
SUV4-20H1
SUV4-20H2
PR-SET7/8
Sp SET9
H2B
H2A
PCAF
Sc GCN5
SGRGKGGKGLGKGGAKRHRKVLR
H4
ARTKYTARKSTGGKAPRKQLATKAARKSAPATGGVKKPKK
H3
NUCLEOSOME:
H2A x 2 = 28 kDa
H2B x 2 = 28 kDa
H3 x 2 = 30 kDa
H4 x 2 = 22 kDa
total protein 108 kDa
dimensions: 10 Å
D 11 nm
H 6 nm
CARM1
HASPIN
LSD1/BHC110
CBP/p300
PCAF
Sc GCN5
MLL1
MLL2
MLL3
MLL4
MLL5
SET1A
SET1B
SET7/9
ASH1
Sc SET1
Sp SET1
PRMT5
SUV39H1
SUV39H2
SETDB1/ESET
EuHMTase/GLP
G9a
CLL8
RIZ1
Sp CLR4
CARM1
EZH2
MSK1
MSK2
Sc FPR4
NSD1
SMYD2
SET2
Sc SET2
Sp SET2
JHDM1a
JHDM1b
JMJD2A/JMJD3A
JMJD2C/GASC1
DOT1
Sc DOT1
Sp DOT1
Sc RTT109
Sc FPR4
DNA:
200 bp
130 kDa
67 nm

Central "Dogma"??





Genomic Variation

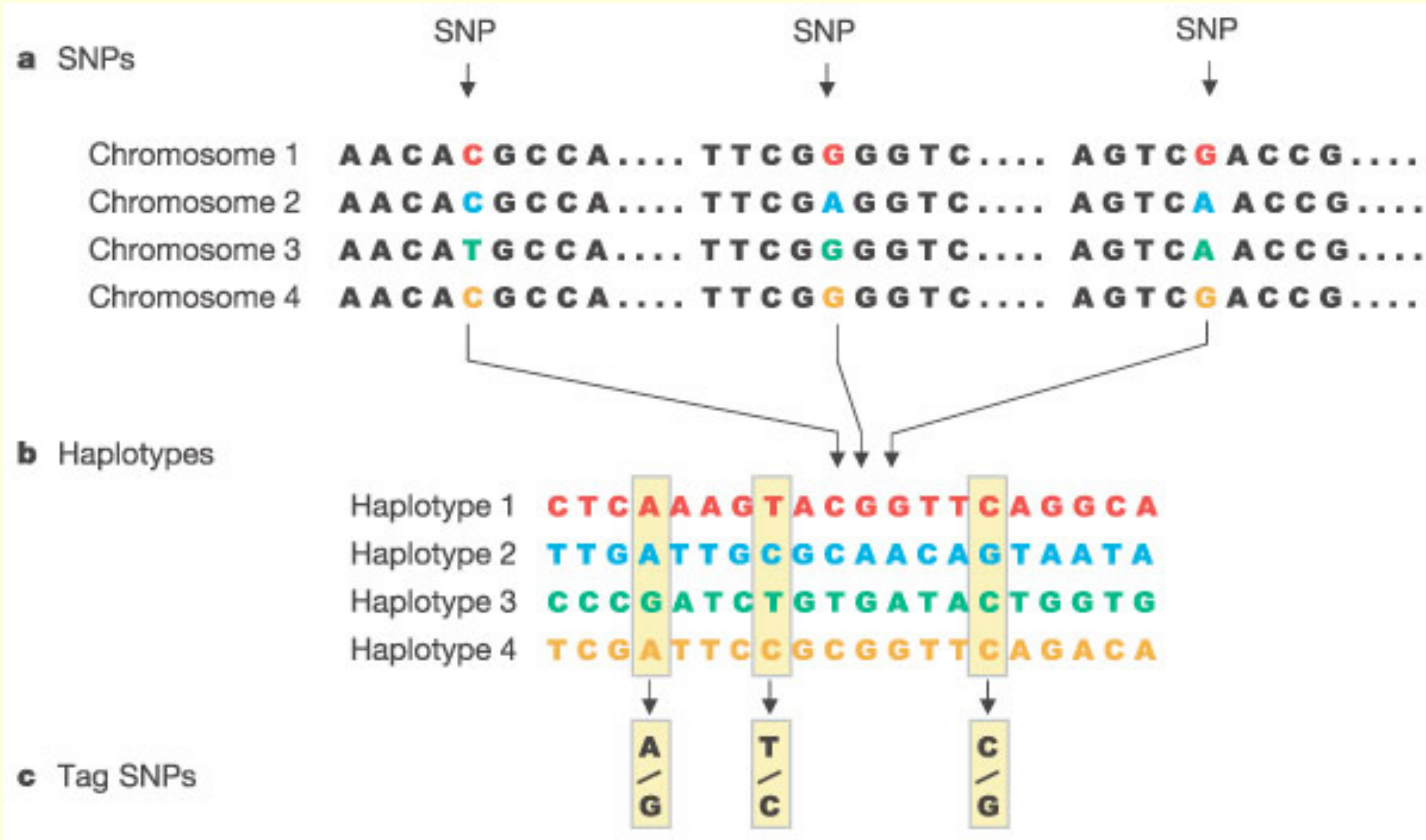


A 'single-nucleotide polymorphism' is a variation at a single nucleotide position in a gene that defines an 'allele'; about 90% of all variation.

It is believed that SNPs can help identify human disease - how do we identify them?

If we can rapidly collect a set of SNPs for individuals, it may be possible to actually to map variation to disease/function.

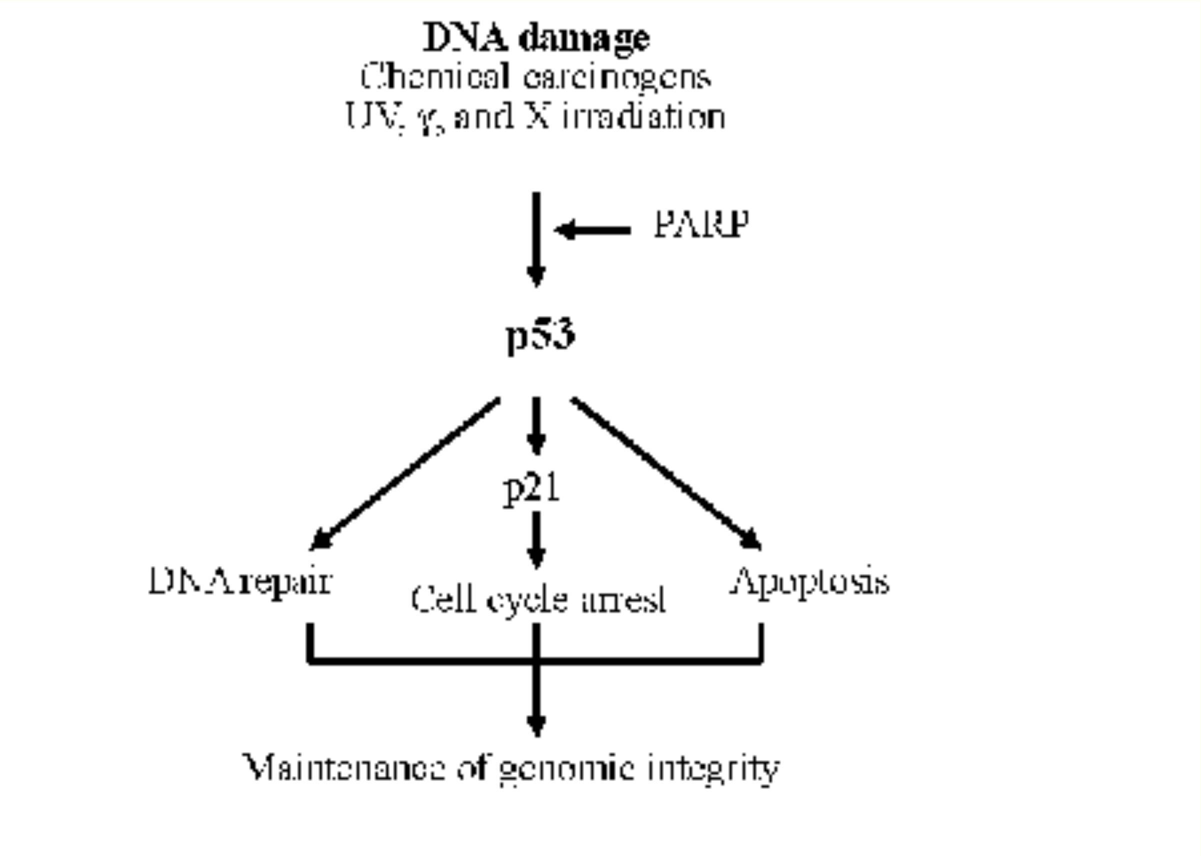
Haplotypes

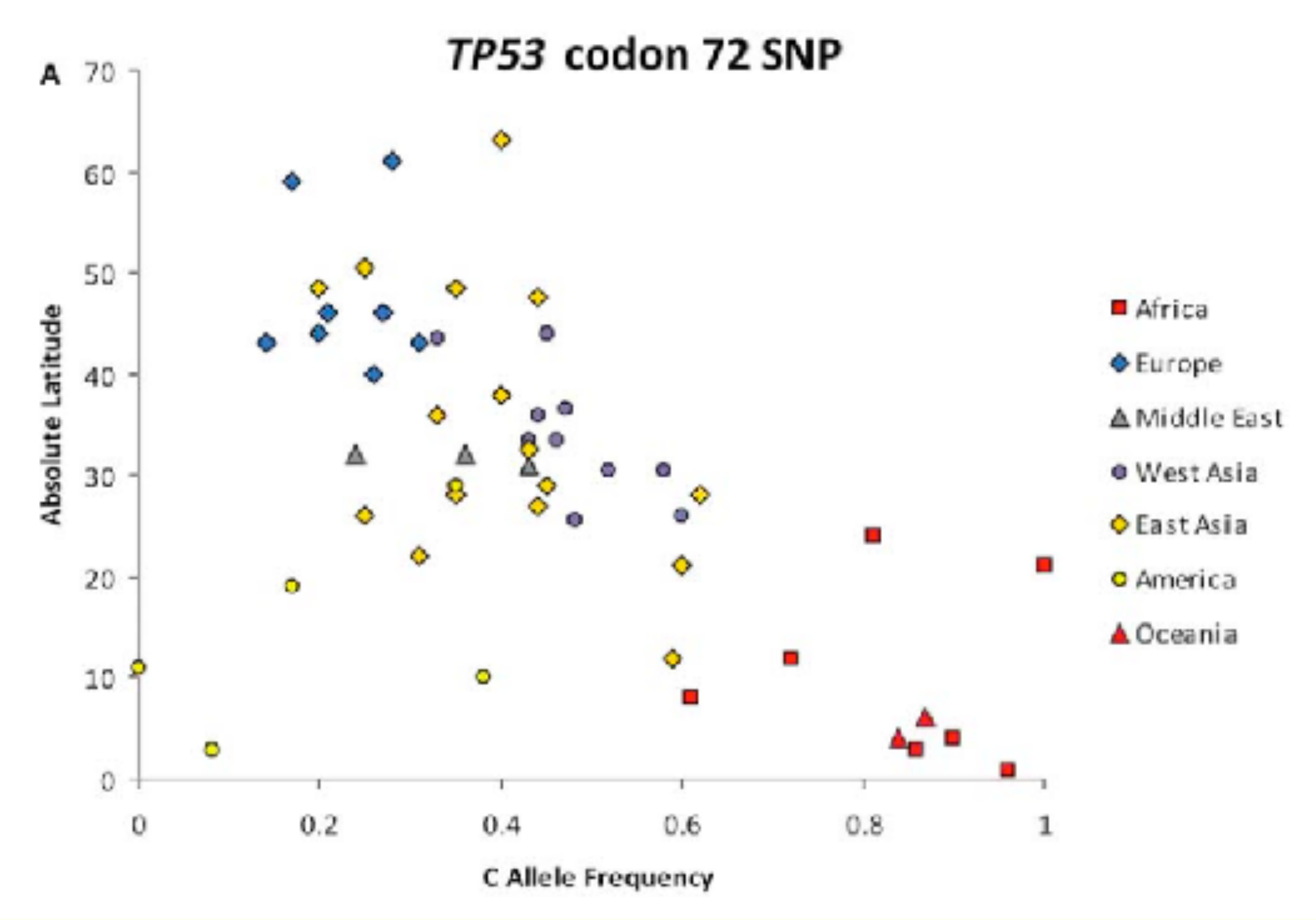


International HapMap Project

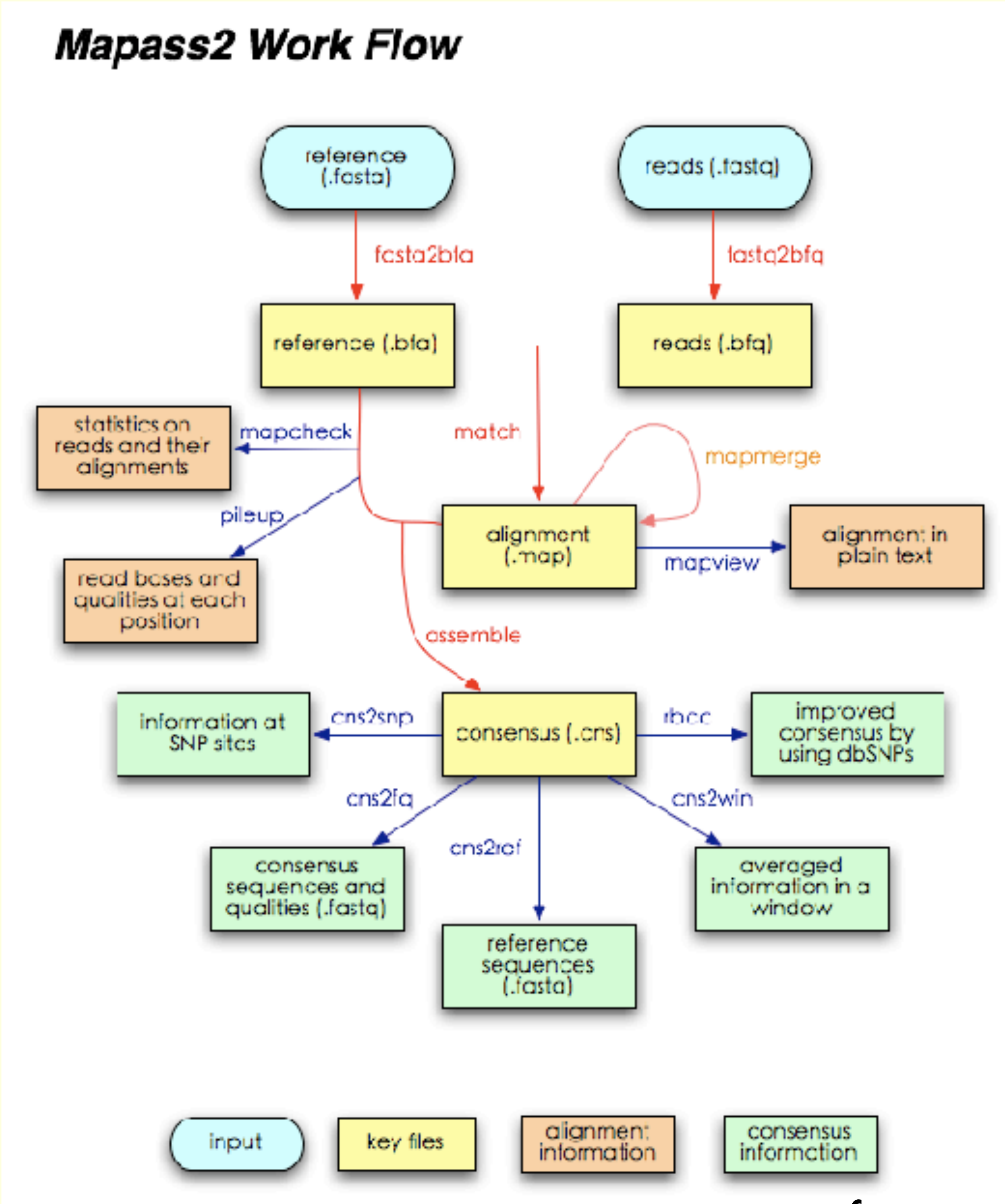
Wellcome Trust Case Control Consortium

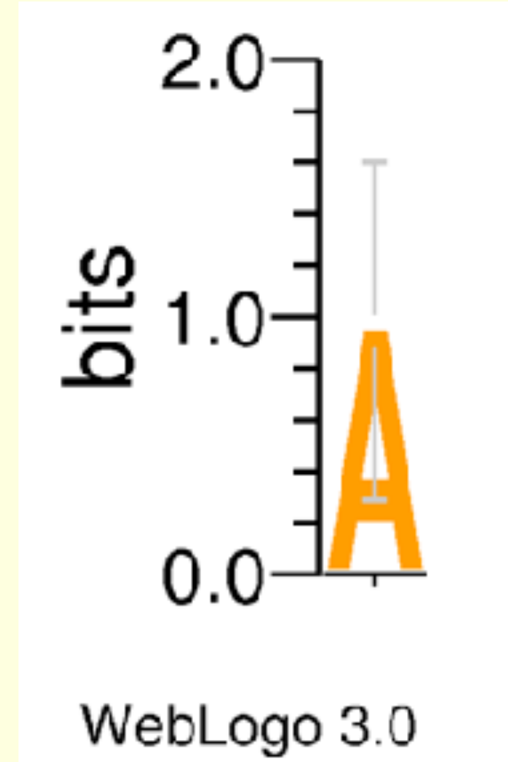
How does selective environmental pressure relate to genetic variation? Is there an observable relationship?

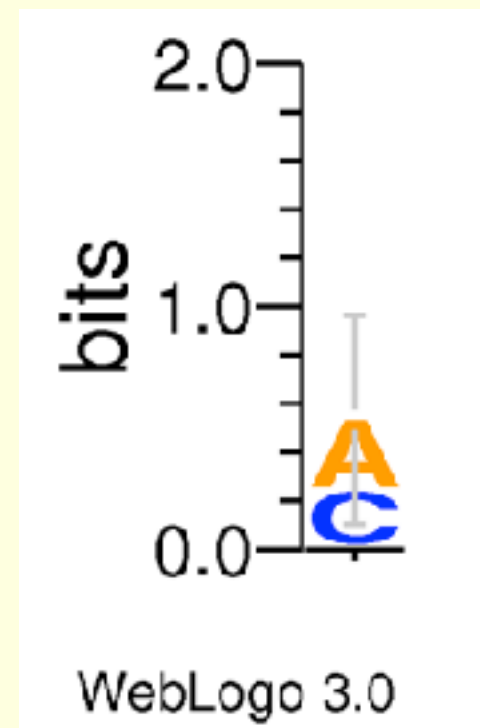




Recent work argues that the p53 pathway, which is key to managing DNA damage, has two SNPs with frequencies that can be used to map individuals to 52 unique populations [Sucheston et al, '11].







maq.sourceforge.net

The usual strength of high-throughput sequencing is to analyze variation in read mapping - SNPs can be found in this manner.

Personal Genomics

- SNPs can yield a very simple classification scheme - there is a cottage industry of 'personal genome analysis' based on this analysis.
- First, you must get your SNPs sequenced; then your risk for particular diseases can be calculated using previously collected statistics.
- deCODE, 23andme, Navigenics, DNA Tribes, SNPedia.
- deCODE hoped to collect haplotypes of the Icelandic population - this was ruled illegal in 2004 due to privacy concerns.

1000 Genomes Project

- The original consensus sequence of the human genome was constructed from 8 individuals.
- The goal of this project is to extend the current human genome sequence with information about variation.
- By sequencing a large set of genomes from diverse populations, they seek to identify variation that is present in more than 1% of each population.
- Variation between and within populations can be studied with relatively "light" read coverage (i.e., 4x).