

Bi430/530

# Theory of Recombinant DNA Techniques

Prerequisite: Molecular Biology (Bi 334)

First half:

DNA basics: isolation, detection, sequencing, gene cloning

Second half:

Manipulation of DNA and genomes

Stem cells and cloning

Transgenic animals and plants

**FRIDAY May 1:** Midterm exam

**WEDNESDAY JUNE 10:** Final exam

# Syllabus

**Bi 430 / 530**  
*(CRN 60373 / 60390)*

## Theory of Recombinant DNA Techniques

### Syllabus, Spring 2021

#### Lecture/discussion:

M-W-F, 10:15 – 11:20. In light of the continuing (but hopefully waning) COVID-19 pandemic, our class meetings will be held remotely at the scheduled times, using Zoom. Class meetings will generally cover a set of lecture materials, posted as a powerpoint file ahead of class. All Zoom class meetings will be recorded, and can be accessed later if you need to miss a class for any reason.

#### Instructor:

Dr. Michael Bartlett --- SRTC Room 458 --- 503-725-3858 --- [micb@pdx.edu](mailto:micb@pdx.edu)

#### Office hours:

Mondays 2 – 3:00 PM, Thursdays 11AM – noon, or by appointment. All office hours will be held using Zoom.

#### Required reading:

All course readings will be posted online.

# Syllabus--first half

## The basics of DNA manipulation

Class Schedule		Topics	Quiz/ assign	Questions to address
Week	Date			
1	M	The molecular revolution; DNA manipulation and biosafety	Q1	How are recombinant DNA risks defined and managed?
	W	Isolation of DNA and RNA		How is useful DNA and RNA isolated?
	F	Isolation of DNA and RNA II		
2	M	Visualization and detection of DNA, RNA, and protein	Q2	How are DNA, RNA and proteins detected and measured?
	W	Detection of <i>specific</i> DNA, RNA, and protein molecules		How can specific DNA, RNA and protein molecules be identified in a complex mixture?
	F	Enzymes for manipulation of nucleic acids		How can DNA be modified in the test tube?
3	M	DNA amplification by PCR	Q3 (A1 due)	Why is PCR such a versatile tool for nucleic acid studies?
	W	DNA sequencing		What DNA sequences exist in nature, and what are they for?
	F	The human genome and its implications		How is the human genome accessed and used?

# Syllabus--first half

## The basics of DNA manipulation

4 M Bioinformatics

How is biological sequence and functional information used?

W Bioinformatics II

Q5

F Genome-scale measurements: microarrays, RNAseq, chromatin immunoprecipitation, proteomes

A2

How can all of the genes in a genome be studied at once?

5 M Cloning genes I: plasmids and transformation

Q6

How is DNA moved into and between biological systems?

W Cloning genes II: special vectors and large DNA fragments

(A2 due)

F **Midterm exam**

# “Molecular Cloning” (2012), Green and Sambrook (4th ed.)

Chapter 1: Isolation and Quantification of DNA

Chapter 2: Analysis of DNA

Chapter 3: Cloning and Transformation with Plasmid Vectors

Chapter 4: Gateway Recombinational Cloning

Chapter 5: Working with Bacterial Artificial Chromosomes and Other High-Capacity Vectors

Chapter 6: Extraction, Purification, and Analysis of RNA from Eukaryotic Cells

Chapter 7: Polymerase Chain Reaction

Chapter 8: Bioinformatics

Chapter 9: Quantification of DNA and RNA by Real-Time Polymerase Chain Reaction

## Chapter 10: Nucleic Acid Platform Technologies

Chapter 11: DNA Sequencing

Chapter 12: Analysis of DNA Methylation in Mammalian Cells

Chapter 13: Preparation of Labeled DNA, RNA, and Oligonucleotide Probes

Chapter 14: Methods for In Vitro Mutagenesis

Chapter 15: Introducing Genes into Cultured Mammalian Cells

Chapter 16: Introducing Genes into Mammalian Cells: Viral Vectors

Chapter 17: Analysis of Gene Regulation Using Reporter Systems

Chapter 18: RNA Interference and Small RNA Analysis

Chapter 19: Expressing Cloned Genes for Protein Production, Purification, and Analysis

Chapter 20: Cross-Linking Technologies for Analysis of Chromatin Structure and Function

Chapter 21: Mapping of In Vivo RNA-Binding Sites by UV-Cross-Linking Immunoprecipitation (CLIP)

Chapter 22: Gateway-Compatible Yeast One-Hybrid and Two-Hybrid Assays

Readings from this manual will be posted

# Syllabus--second half

## Applications of rDNA

6	M	Cloning genes III: library construction and screening, recombination-based engineering, cloning in prokaryotes other than <i>E. coli</i>		How can a specific piece of DNA be identified and cloned?
	W	Protein expression I		How can cells be made to produce useful products?
	F	Protein expression II	Q7	
7	M	Mutagenesis, protein engineering, altering the genetic code		How can genes & organisms be altered for practical purposes?
	W	Applied mutagenesis: metabolic engineering, genome shuffling, synthetic genomes	Q8	
	F	Applied mutagenesis II	A3	
8	M	Cloning in <i>Saccharomyces cerevisiae</i> . Cloning in higher eukaryotic cells: cell culture, embryonic and induced pluripotent stem cells, organismal cloning	Q9	Why is yeast such a useful model system for eukaryotes? Why are stem cells so useful? How can an organism be cloned?
	W	Cloning in eukaryotic cells: transformation and viral transduction	(A3 due)	How is new DNA added to eukaryotic cells?
	F	Cloning in eukaryotic cells: selection strategies and genetic control	Q10	How are added genes controlled?

# Syllabus--second half

## Applications of rDNA

9 M Gene therapy and CRISPR-Cas9

How is gene therapy being done? How is Crispr-Cas9 being used

W CRISPR-Cas9 II

Q11

F Nucleic acid vaccines

A4 How can the immune system be programmed to prevent infectious diseases?

10 M *Memorial Day (no class)*

W Transgenic animals

Q12 How and why are transgenic animals and plants made?

F Genetic manipulation of plants

(A4 due)

Finals **Final exam**, Wed. June 9, 10:15 – 12:05

# Recombinant DNA Techniques during a viral pandemic?

## **Grading:**

<b>Grading:</b>	<u>Bi 430</u>	<u>Bi 530*</u>	
	40%	30%	Quizzes: lowest two quiz scores are dropped
	30%	20%	Homework: lowest homework score is dropped
	15%	15%	Midterm exam
	15%	15%	Final exam
	-----	20%	* In-class presentations, to be given during the second half of the class

Grading cut-offs will be as follows: 93% and up, A; 90 and up, A-; 88 and up, B+; 82 and up, B; 80 and up, B-; 77 and up, C+; 68 and up, C; 65 and up, C-; 62 and up, D+; 53 and up, D; 50 and up, D-; under 50, F.

**There are no makeup exams.** You must take both exams or you cannot earn a passing grade.

Academic dishonesty (cheating, plagiarism, etc.) will result in a zero for the assignment, and will be reported to student affairs, as described in the PSU Code of Conduct: <https://www.pdx.edu/dos/psu-student-code-conduct>

If you are a student with a documented disability and have registered with the Disability Resource Center, please contact me immediately to arrange academic accommodations.

# Readings

## **Readings to be posted online**

Information from: “Molecular Cloning, A Laboratory Manual”  
Sambrook and Russell (2012), Cold Spring Harbor  
Laboratory Press

Various papers (PDFs) through out the term

---

If you would like additional resources for specific topics, let me know

# Introduction to DNA manipulation

---

- 1) The simplicity of a DNA-based information system makes direct, deliberate genetic manipulation possible
  - 2) This represents an unprecedented power for human interaction with living systems
  - 3) Benefits and costs of technology require continuous assessment
-

# Guide to readings: Day 1

## Discovery of the genetic code

- *Nirenberg 1967.* Essay: “Will Society Be Prepared?”, predicts and ponders the effects of DNA manipulation.
- *NIH Nirenberg Papers.* Some historical context for the discovery of the genetic code.

## Basic lab safety and recombinant DNA

- *0.1 MC4 Safety:* short guide to lab safety from the Molecular Cloning manual.

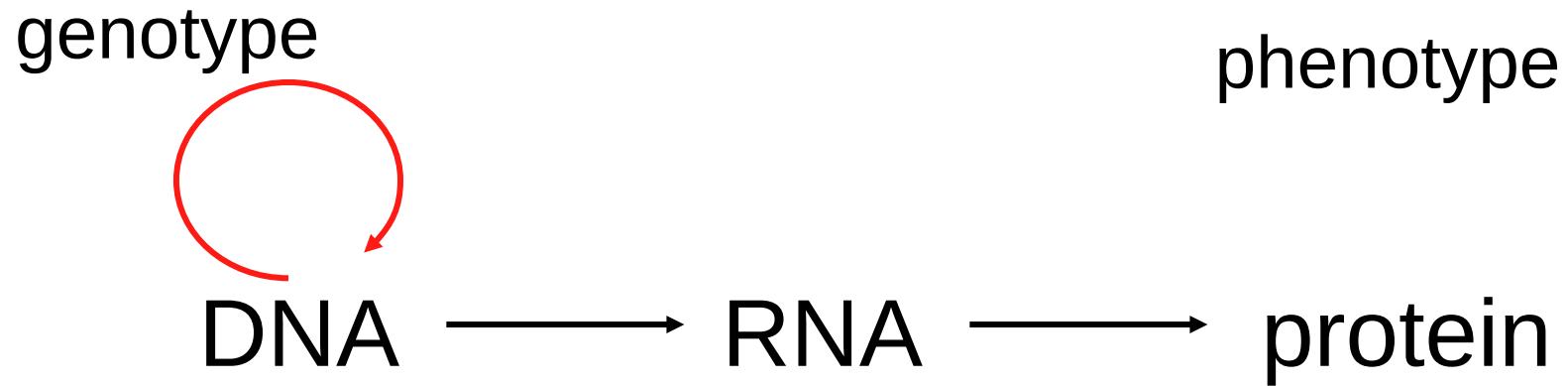
## Genetic modification in the 21<sup>st</sup> century

- *Reboot the debate...* by Jennifer Kuzma. Both product and process are important for genetic engineering (2016).

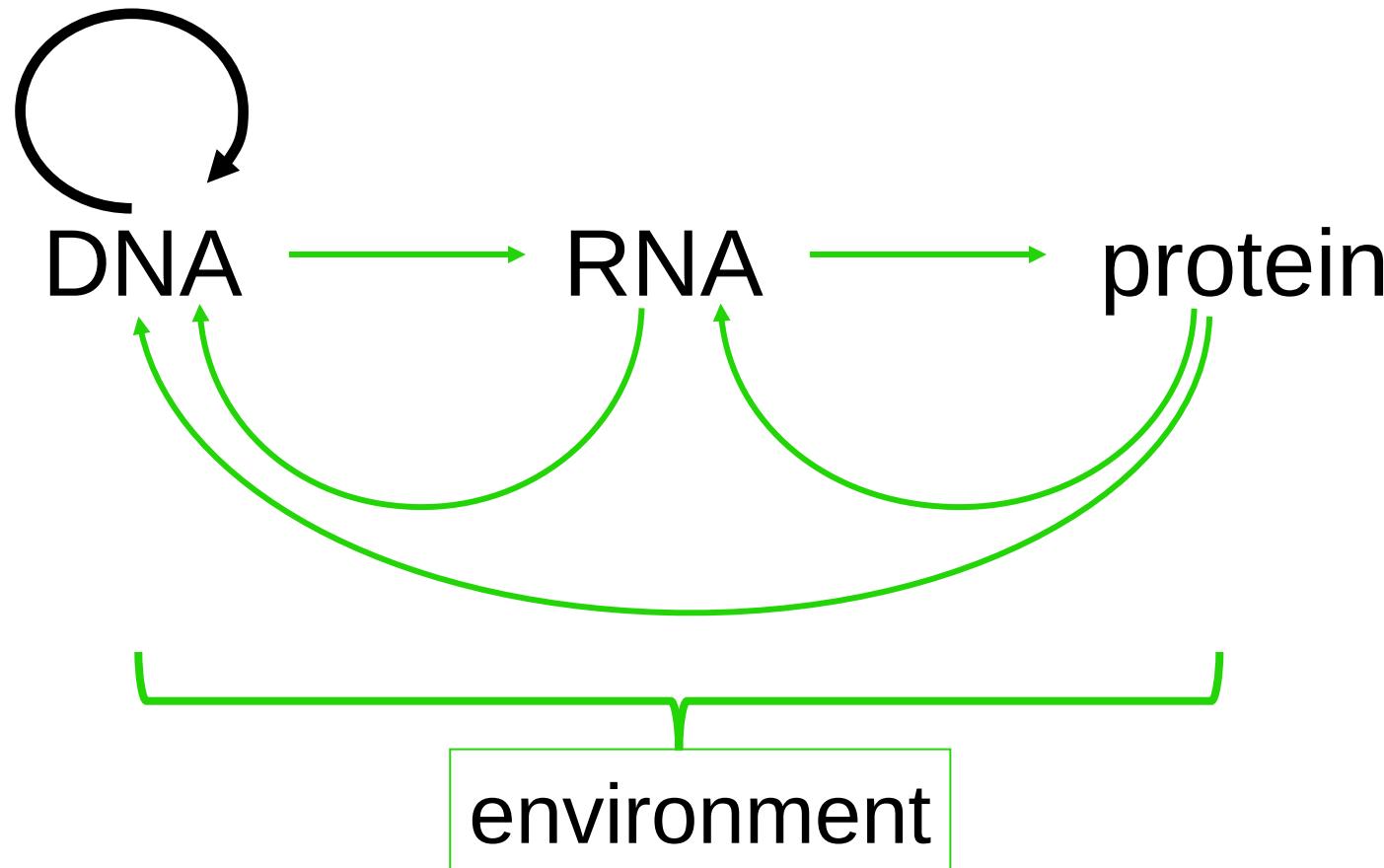
## Video: Paul Berg discussing the Asilomar meeting and contemporary analysis of the risks of rDNA

- <https://youtu.be/QSKe15I4vyM> (part I)
- [https://youtu.be/Eg2Sz\\_-l9UI](https://youtu.be/Eg2Sz_-l9UI) (part II)

# Information flow in the cell

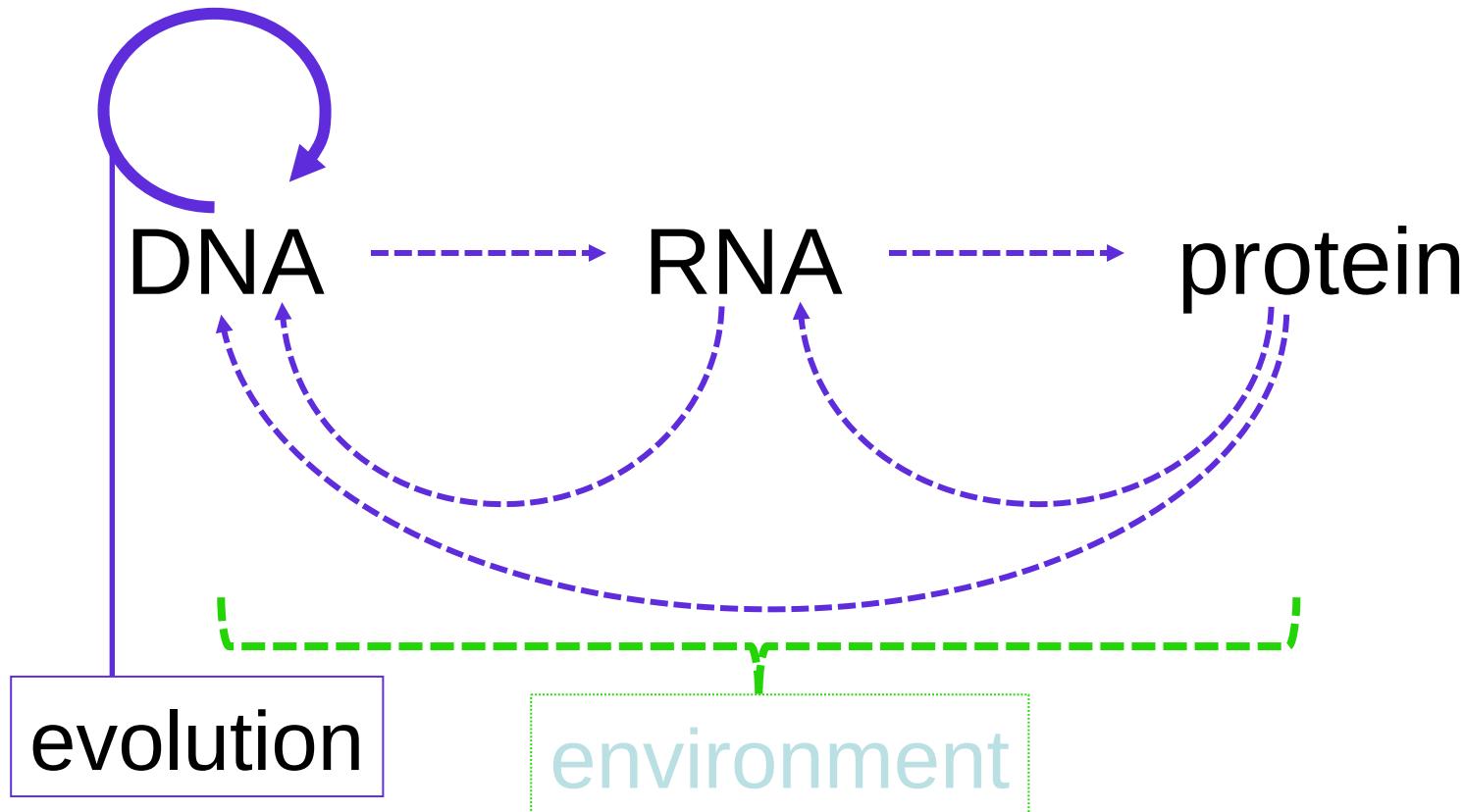


Organisms respond to their environment via information from sensory input, causing changes in gene expression



Dynamic but mostly transient modification of DNA program

Evolution: species respond to environment over long time frames via additions, deletions, rearrangements, and mutations in the DNA program



- Heritable modifications of genetic program
- Reflects adaptation to environment over long time scales

# DNA structure

Rosalind Franklin and Maurice Wilkins:

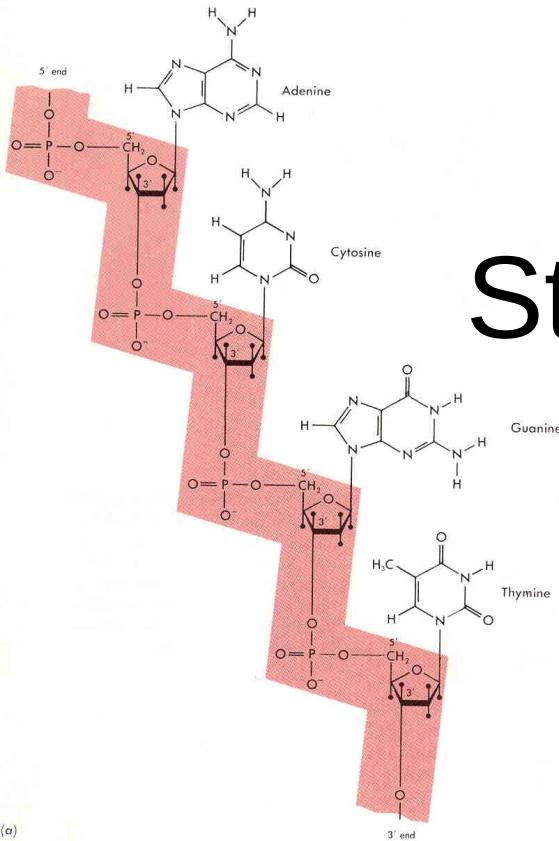
X-Ray fiber diffraction pattern of pure B-form DNA (1953)

James Watson and Francis Crick:

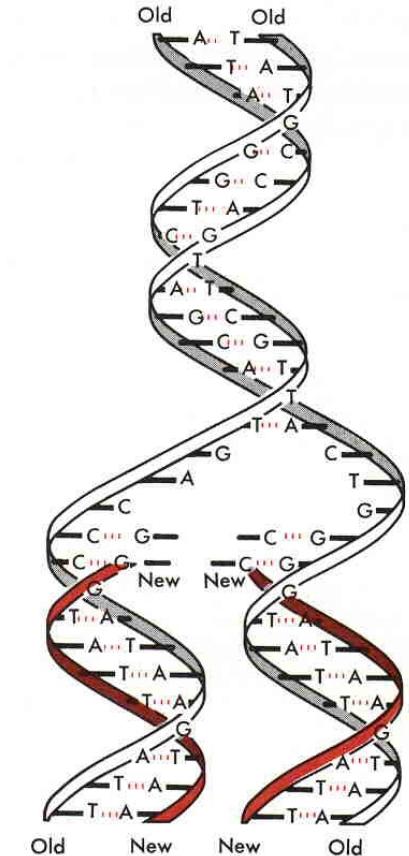
Proposed two antiparallel, helical strands forming a stable duplex with DNA bases on interior of the molecule, joined by hydrogen bonds (1953)

DNA had been known as the element of genetic transmission at least since 1947, when Avery showed that DNA could “transform” bacterial colony morphology

Why was the structure so important?



# Structure of DNA



DNA structure suggested:

- Mechanism for replication
- Stable information storage -- but accessing the information not difficult

The DNA structure provided a new template for hypotheses regarding biological phenomena

# DNA is very easy to work with...

- Easy to isolate -- plasmids, genomic DNA, PCR, etc.
- Stable -- not as chemically reactive as RNA (even archaeologically stable!)
- Easy to propagate and move from cell to cell
- Easy to make specific constructs
- Easy to make specific mutations
- Very easy to sequence
- Predictable behavior (the genetic code)
- Sequence lends itself to analysis (**genome projects**)

DNA is very easy to sequence: early completed genomes

<b>Genome sequenced</b>	<b>Year</b>	<b>Genome size</b>
Bacteriophage φX174	1977	5.38 kb
Plasmid pBR322	1979	4.3 kb
Bacteriophage λ	1982	48.5 kb
Epstein–Barr virus	1984	172 kb
Yeast chromosome III	1992	315 kb
<i>Haemophilus influenzae</i>	1995	1.8 Mb
<i>Saccharomyces cerevisiae</i>	1996	12 Mb
<i>Ceanorhabditis elegans</i>	1998	97 Mb
<i>Drosophila melanogaster</i>	2000	165 Mb
<i>Homo sapiens</i>	2000	3000 Mb
<i>Arabidopsis thaliana</i>	2000	125 Mb
Etc....		

# Genomes OnLine Database (GOLD)

<https://gold.jgi.doe.gov/>

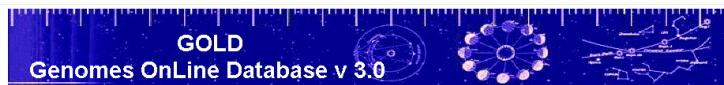


Home   Search   Distribution Graphs   Biogeog

Studies	<a href="#">49,580</a>
Biosamples	<a href="#">131,477</a>
Sequencing Projects	<a href="#">408,761</a>
Analysis Projects	<a href="#">320,257</a>
Organisms	<a href="#">410,765</a>

Welcome  
GOLD:  
metage

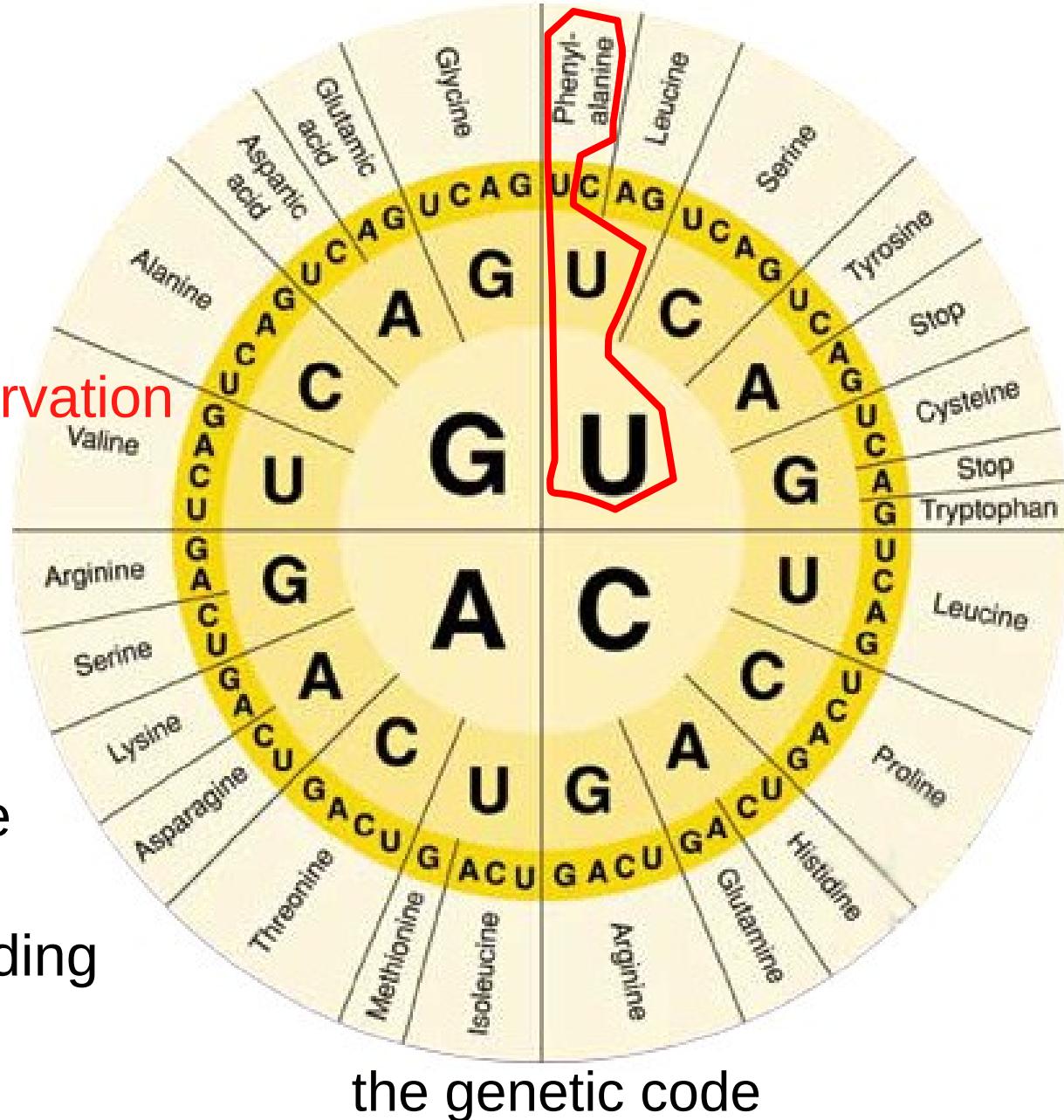
2021



Contact: <a href="#">Genomesonline</a>	Last Update: <a href="#">2011-03-25</a>	Location: <a href="#">www.genomesonline.org</a>
<b>1661</b> <small>Complete Published</small>	<b>Search GOLD: 10047 genome projects</b>	<b>307</b> <small>Metagenomes</small>
<b>210</b> <small>Archaeal Ongoing</small>	<b>5843</b> <small>Bacterial Ongoing</small>	<b>2003</b> <small>Eukaryal Ongoing</small>
<a href="#">GOLD RSS Feeds</a>		METAGENOME CLASSIFICATION
PROJECT TYPE DISTRIBUTION	SEQUENCING STATUS DISTRIBUTION	PHYLOGENETIC DISTRIBUTION

2011

# The behavior of DNA (genes) is predictable



Gene **sequence conservation** often indicates **functional similarity**

Non-protein coding information sequences can also sometimes be detected by homology (for example, DNA binding protein binding sites)

# The genetic code and the roots of biotechnology

1961

Marshall Nirenberg and Heinrich J. Matthaei:  
polyU mRNA encodes poly-phenylalanine

1966

Nirenberg and colleagues had deciphered the 61 codons (and 3 nonsense codons) for the 20 common amino acids

1968

Nobel prize for Nirenberg, Holley, and Khorana

1966

George and Muriel Beadle (geneticist & author) write:

“ The deciphering of the DNA code has revealed our possession of a language much older than hieroglyphics, **a language as old as life itself**, a language that is the most living language of all--even if its letters are invisible and its words are buried deep in the cells of our bodies.”

# The public reaction to the deciphering of the genetic code

---

## Wow

“...just as big a breakthrough in biology as [Newton's discovery of gravitation in the seventeenth century] was in physics.” --John Pfeiffer, journalist, 1961

## Optimism

“No stronger proof of the universality of all life has been developed since Charles Darwin's 'The Origin of Species' demonstrated that all life is descended from one beginning. In the far future, the hope is that the hereditary lineup will be so well known that science may deal with the aberrations of DNA arrangements that produce cancer, aging, and other weaknesses of the flesh.” Chicago Sun-Times, 1962

## Caution

...knowledge gained from the genetic code “might well lead in the foreseeable future to a means of directing mutations and changing genes at will.” 1961, A. G. Steinberg of Case Western Reserve University

...knowledge of the genetic code could “lead to methods of tampering with life, of creating new diseases, of controlling minds, of influencing heredity, even perhaps in certain desired directions.” 1961, Arne Wilhelm Kaurin Tiselius, 1948 Nobel Laureate in Chemistry

1967

"Will Society Be Prepared?" Marshall Nirenberg,  
editorial in *Science* (see letter online)

same language, with minor variations. Simple genetic messages now can be synthesized chemically. Genetic surgery, applied to microorganisms, is a reality. Genes can be prepared from one strain of bacteria and inserted into another, which is then changed genetically. Such changes are inheritable. Thus far, it has not been possible to program mammalian

What may be expected in the future? Short but meaningful genetic messages will be synthesized chemically. Since the instructions will be written in the language which cells understand, the messages will be used to program cells. Cells will carry out the instructions, and the program may even be inherited. I don't know how long it will take before it will be possible to program cells with chemically synthesized messages. Certainly the experimental obstacles are formidable. However, I have little doubt that the obstacles eventually will be overcome. The only question is when. My guess is that cells will be programmed with synthetic messages within 25 years. If efforts along those lines were intensified, bacteria might be programmed within 5 years.

Nirenberg, 1967

"When man becomes capable of instructing his own cells, he must refrain from doing so until he has sufficient wisdom to use this knowledge for the benefit of mankind....

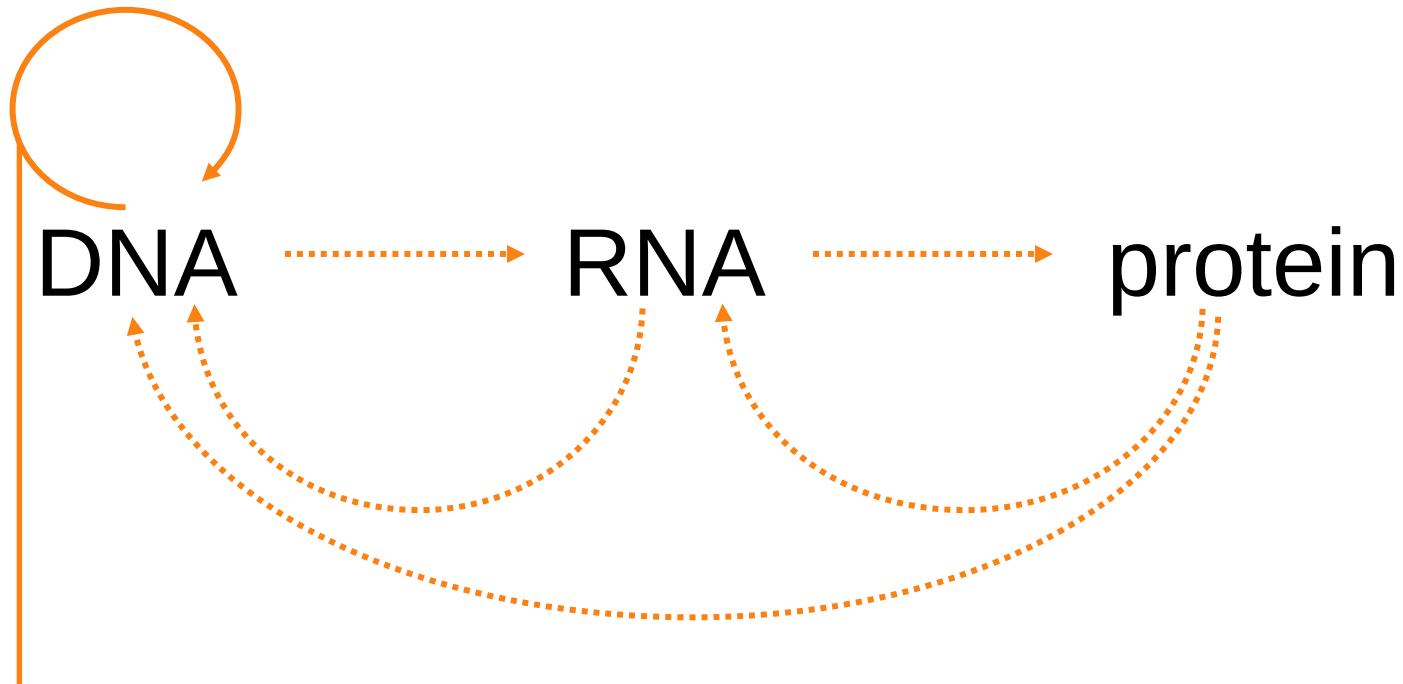
[D]ecisions concerning the application of this knowledge must ultimately be made by society, and only an informed society can make such decisions wisely."

Response from Joshua Lederberg, 1967 (see letter):

(paraphrased)

- We need to be particularly careful with manipulation of the germ cell lines (heritable changes).
- Considerations governing control of our biology are equally important to considerations governing control of our cultural institutions (given that culture is mutable and heritable)

Human activity: transient modifications of environment, permanent modifications of DNA program



Human intervention: genetics (indirect), rDNA (direct)

# The first recombinant DNA molecules: 1972

- Paul Berg and co-workers at Stanford Univ.
- SV40 (Simian virus 40) had the genes from the E. coli galactose operon inserted

# In 1974, a voluntary moratorium was declared on recombinant DNA research

---

- “...new technology created extraordinary novel avenues for genetics and could ultimately provide exceptional opportunities for medicine, agriculture and industry....  
...concerns that unfettered pursuit of this research might engender unforeseen and damaging consequences for human health and the Earth’s ecosystems.”
- 1975: The Asilomar Conference on Recombinant DNA

[http://nobelprize.org/nobel\\_prizes/chemistry/  
articles/berg/index.html](http://nobelprize.org/nobel_prizes/chemistry/articles/berg/index.html)

# 1975: The Asilomar Conference on Recombinant DNA

---

- In attendance: internationally prominent scientists, government officials, doctors, lawyers, members of the press
- Conclusion:  
“...recombinant DNA research should proceed, but under strict guidelines.”
- The moratorium was lifted, and “... guidelines were subsequently promulgated by the National Institutes of Health and by comparable bodies in other countries.”

[http://nobelprize.org/nobel\\_prizes/chemistry/  
articles/berg/index.html](http://nobelprize.org/nobel_prizes/chemistry/articles/berg/index.html)

## The Asilomar principles:

- 1) **containment** should be an essential consideration in the experimental design
- 2) the effectiveness of the containment should **match the estimated risk** as closely as possible.

### Additional suggestions:

Use **biological barriers** to limit the spread of recombinant DNA

- Fastidious bacterial hosts (able to grow only with specific nutrients) that are unable to survive in natural environments
- nontransmissible and equally fastidious vectors (plasmids, bacteriophages, or other viruses) that are able to grow in only specified hosts

# The Asilomar principles:

## Safety factors

- physical containment, exemplified by the use of hoods or where applicable, limited access or negative pressure laboratories
- strict adherence to good microbiological practices, which would limit the escape of organisms from the experimental situation
- education and training of all personnel involved in the experiments would be essential to effective containment measures.

# Regulation of biotechnology: US National Institutes of Health (NIH) Guidelines

- stipulations of **biosafety and containment** measures for recombinant DNA research
- delineations of critical **ethical principles** and safety reporting requirements for **human** gene transfer research

See <http://oba.od.nih.gov/rdna/rdna.html>

## “Reboot the debate” (essay by Jennifer Kuzma, posted)

---

Previous U.S. regulations of genetically modified organisms have focused mainly on the *product*, with little concern given to the *process* used to obtain the organism (following the Coordinated Framework for Reg. of Biotech [CFRB] of 1986)

Some processes are imprecise and introduce certain kinds of uncertainty, while others are much more precise

Some products are clearly innocuous, while others are potentially (or may be actually) dangerous

Both process and product need to be considered in debate on genetic modification

# The question of ‘synthetic biology’

- Synthetic biology: biological systems that are
  - Programmable
  - Self-referential\*
  - Modular
- The complexity of biological systems being created will likely lead to unexpected behaviors
- Rationale for this kind of work needs to be clearly stated. What is the utility of synthetic biological entities? How can public trust be retained?

\* refers to biological entities or systems that are composed of elements that can interact, respond to changes, self-modify, replicate or expand.

# Synthetic biology: four areas of ecological risk assessment

- 1) Specific physiology of the organism. Does it produce toxins, for example?
- 2) How will synthetic organism affect its environment? Will it affect biodiversity, for example?
- 3) Could the synthetic organism evolve quickly, adapting to new environments?
- 4) Can the synthetic organism transfer its genes to other organisms?

## What can we do with recombinant DNA technology?

- begin to learn how cells, tissues, organisms, communities work, interact, respond to the environment (gain **scientific knowledge**)
- improve **human health**
- **industrial production** of useful enzymes, metabolic products
- improve industrial process
- raise agricultural productivity
- investigate problems of genealogy, paternity, anthropology, archaeology
- investigate criminal cases
- etc....

# Recombinant DNA technology in medicine

---

- Understand molecular mechanisms of disease
- Predict and diagnose of disease
- Animal models for human diseases
- Therapies
  - nucleic acids: gene therapy
  - production of pharmacologically active proteins
  - small biomolecule synthesis and testing
- Antimicrobial strategies
  - Vaccines
  - Antibiotic development and production

## Summary:

---

- 1) The simplicity and predictability of a DNA-based information system makes genetic manipulation possible
  - 2) This represents an unprecedented level of interaction with living systems
  - 3) Benefits versus costs of recombinant DNA technology require continuous assessment
-

# DNA and RNA: isolation and purification

## I. DNA

- a. Genomic DNA
- b. Plasmid DNA
- c. Removal of contaminating proteins
- d. Concentrating dilute samples of nucleic acids
- e. Old/ancient DNA

## II. RNA

- a. What kind of RNA?
- b. Special problems with RNA
- c. Battling RNase

**Thought experiment:** devise separation strategy for these items, based on each item's unique properties:

- 18 basketballs
- 350 ping pong balls
- 15000 metal ball bearings
- 40 metal cannon balls
- 300 plastic balls (same size and density as wooden)
- 300 wooden balls
- 10 golf balls

## References:

### 1) MC4 DNA Purification

- Introduction to DNA purification, and kits (p. 2-5)
- Phenol extraction of proteins (p. 44-46)
- Ethanol precipitation of DNA (p. 21-25)
- Isopropanol precipitation of DNA (p. 26-27)
- Concentration of DNA (p. 28-30)

### 2) MC4 RNA Purification

- Overview and introduction to monophasic lysis reagents (p. 346-350)
- RNA quantification and storage (p. 365-371)
- Oligo dT beads for mRNA isolation (p. 377-380)
- Controlling RNases; DEPC (p. 450-453)

### 3) The Kit Generation: know how the kits work!

### 4) Friedrich Miescher: the first isolation of DNA

### 5) Ancient DNA 2001: challenges of old DNA analysis

## Pure DNA is essential

- Detect and clone genes
- Identify organisms/viruses
- Sequence DNA regions
- Create new DNA constructions (recombinant DNA)

## Pure RNA is also essential

- Identify transcribed genes, exons
- Determine transcription levels
- Clone transcribed genes
- control gene expression (RNAi)

# Basic principles of biomolecular separation

You break open cells to release molecules: but then what?  
how do you isolate the macromolecule you want?

- **Unique chemistry** of the biomolecule
  - hydrophobicity/hydrophilicity
  - surface charge
- **Size** of the biomolecule
- **Topological state** of the biomolecule
- Susceptibility/resistance to **enzyme treatment**
- **Interactions** with other biomolecules

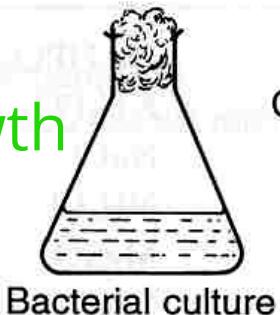
**Thought experiment:** devise separation strategy for these items, based on each item's unique properties:

- Protein
- RNA
- genomic DNA
- Small DNA molecules (like plasmids)
- Phospholipids
- Small molecules/ions (nucleotides, Mg<sup>++</sup> or Ca<sup>++</sup> ions, etc.)

# Isolating DNA: overview for bacterial cells

cell growth

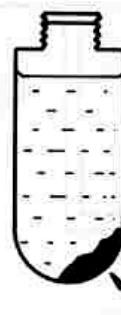
1 A culture of bacteria is grown and then harvested



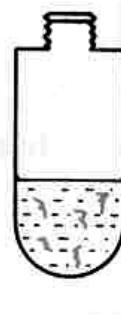
Centrifugation

for 10 minutes at 3000 rpm

at room temperature

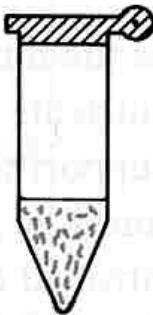


2 The cells are removed and broken to give a cell extract



cell harvest and lysis

DNA concentration



4 The DNA is concentrated

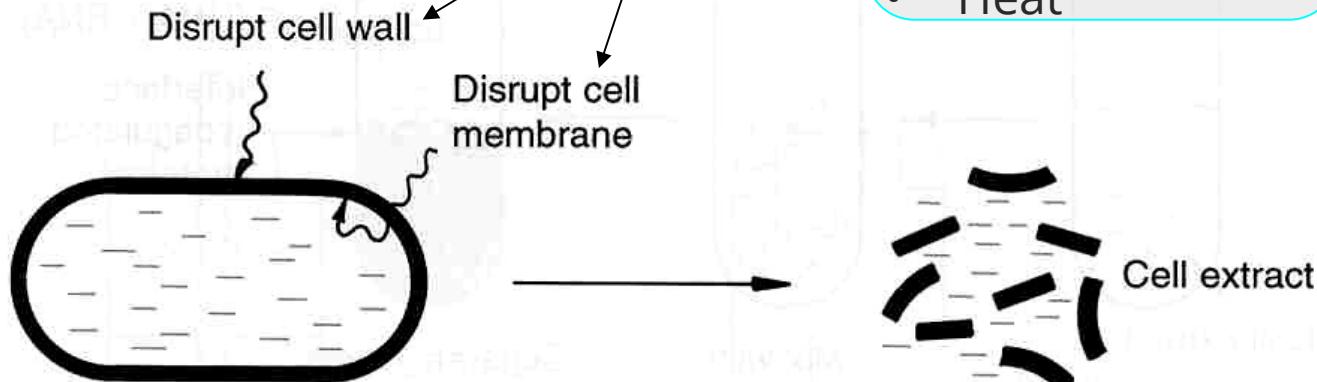
DNA purification



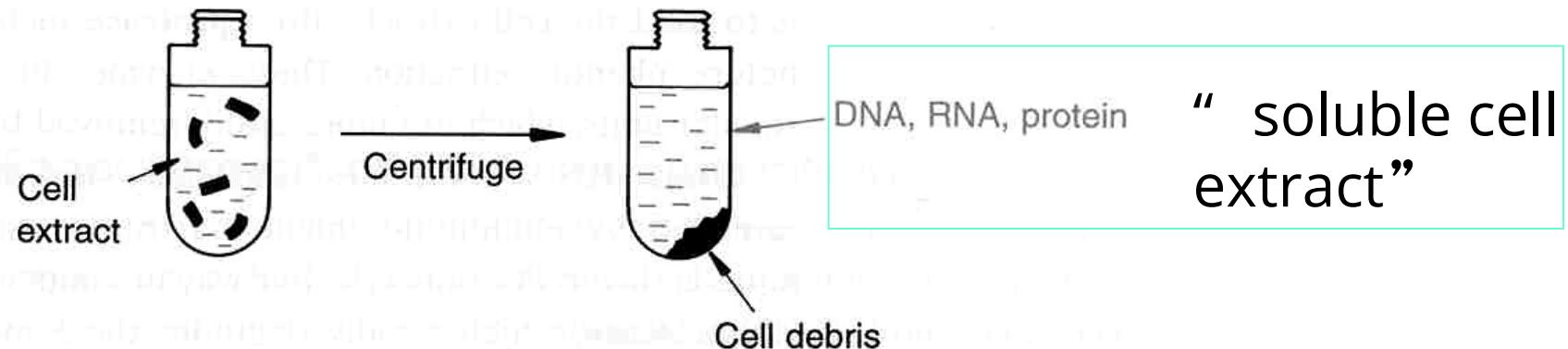
3 The DNA is purified from the cell extract

# Bacterial genomic DNA: cell extract

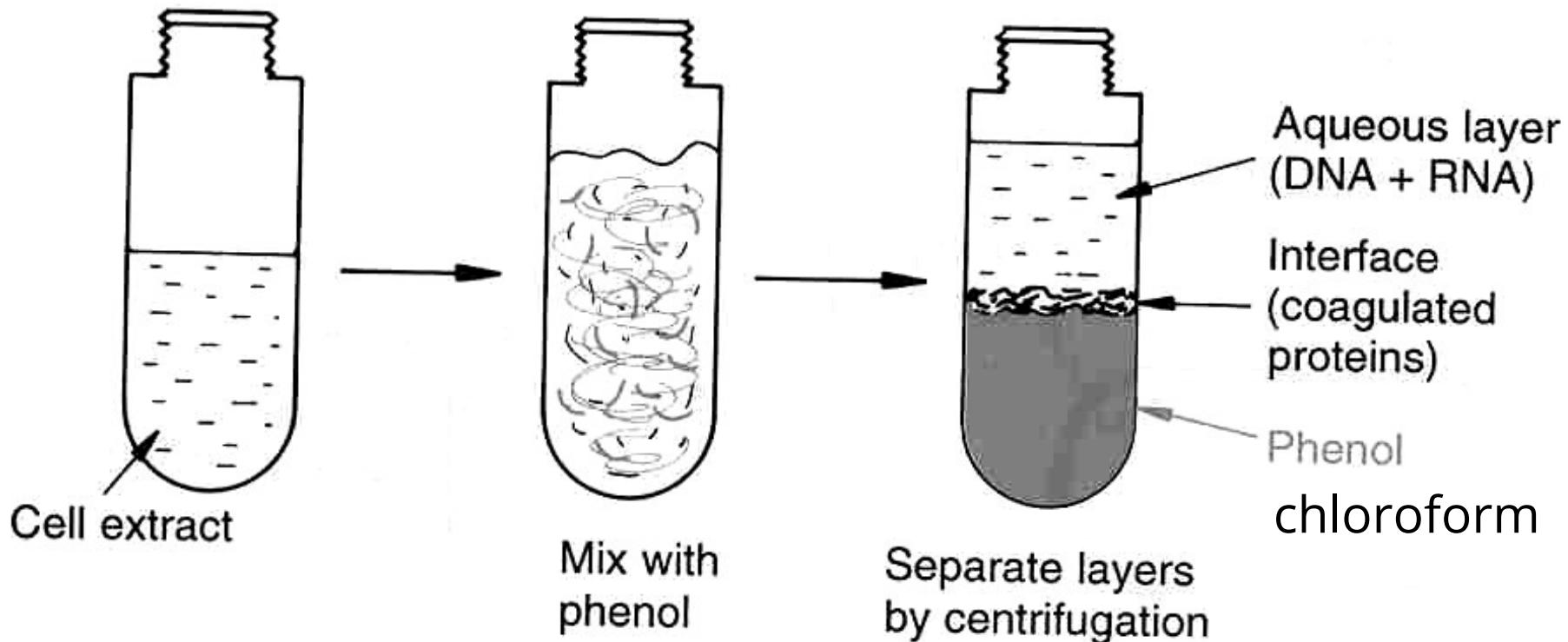
## (a) Cell lysis



## (b) Centrifugation to remove cell debris



# Genomic DNA: remove proteins and RNA

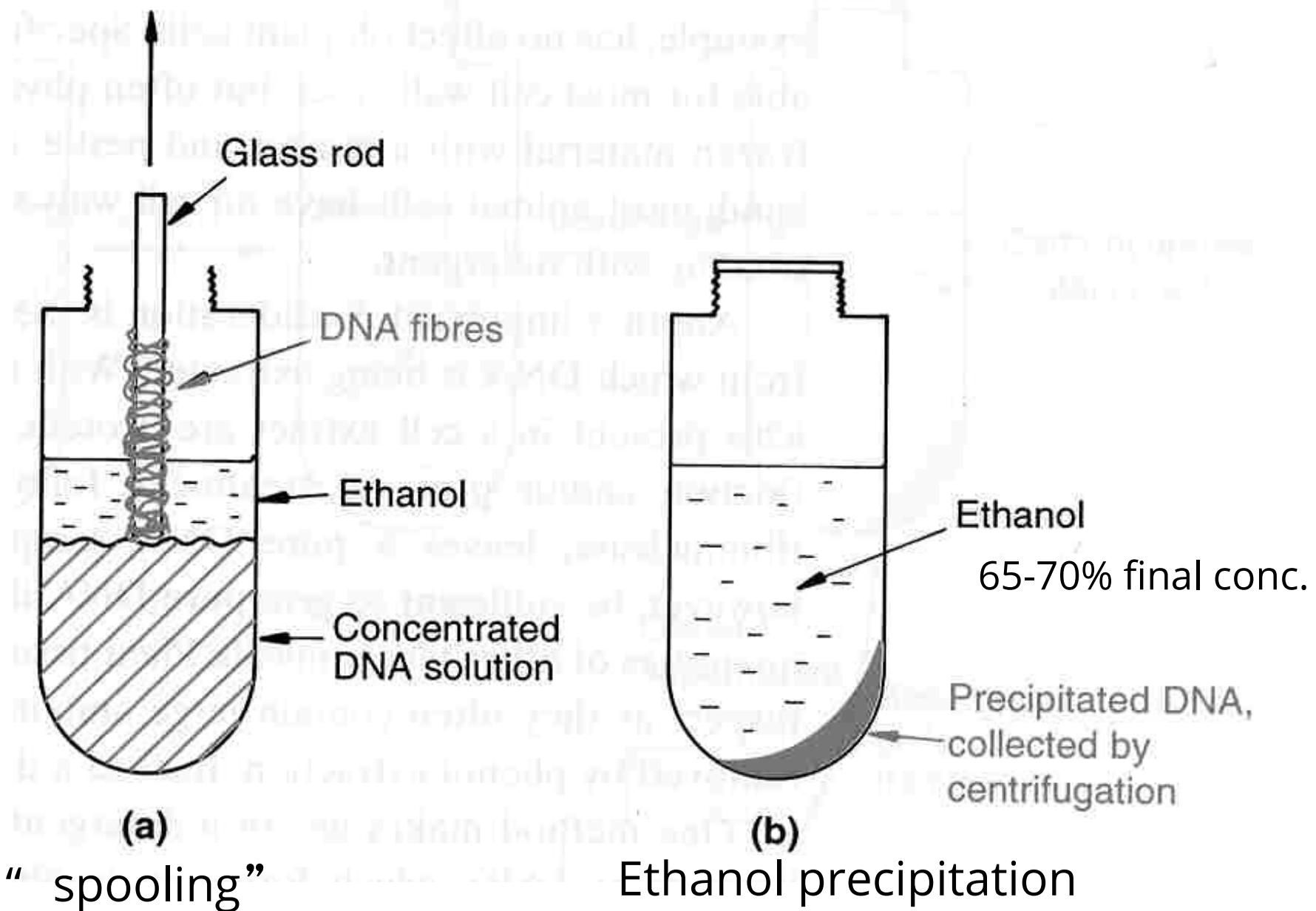


Mix gently (to avoid shearing breakage of the genomic DNA)

DNA and RNA are recovered in aqueous layer

Add the enzyme RNase to remove the RNA

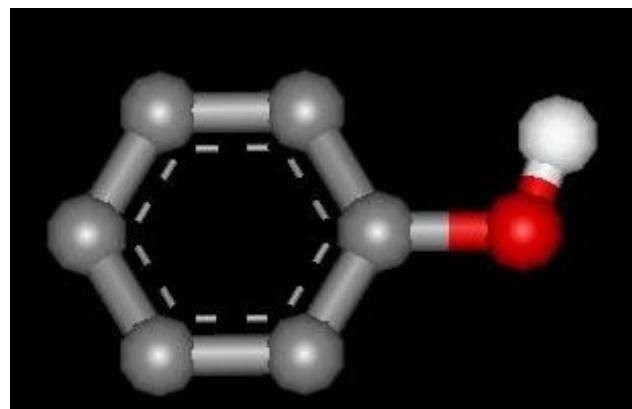
# Recover the genomic DNA as a solid precipitate



# Separating nucleic acids from protein: phenol extraction

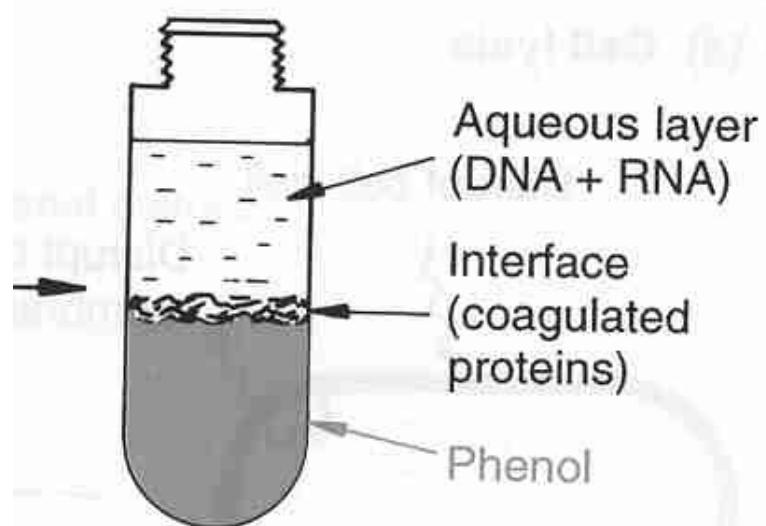
## Phenol

- Organic liquid, forms a separate phase from water
- Amphipathic molecule that unfolds proteins by disrupting hydrophobic core
- Insoluble proteins collect at the interface between water and phenol layers



# Phenol extraction to remove proteins

1. Aqueous volume (at least 200 microliters)
2. Add 2 volumes of phenol:chloroform, mix well
3. Spin in centrifuge, save aqueous phase in a new tube, avoiding the protein at the interface
4. Repeat steps 2 and 3 until there is no precipitate at phase interface
5. Extract aqueous layer with 2 volumes of chloroform to remove traces of phenol



# DNA concentration by ethanol precipitation

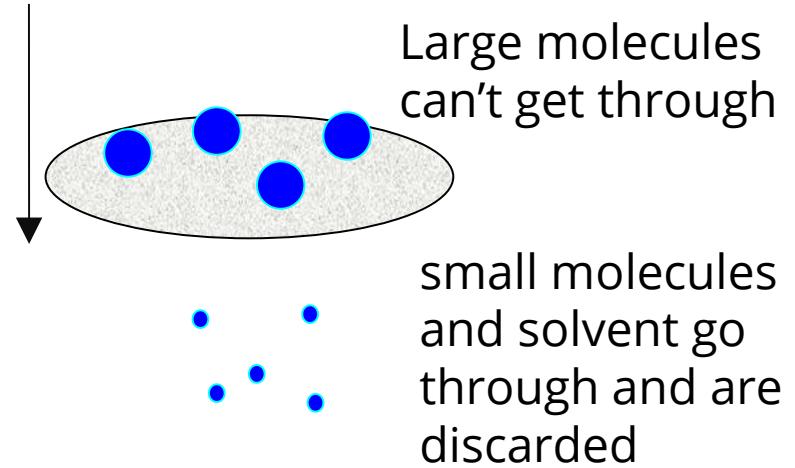
Ethanol depletes the hydration shell surrounding DNA...

- Allowing cations from added salt to interact with the DNA phosphates
  - Reducing repulsive forces between DNA strands
  - Causing aggregation and precipitation of DNA
- 
- Aqueous volume of dilute DNA: 180 microliters
    - add 20 microliters sodium acetate 3M pH 5.2
    - add 1 microliter of glycogen (gives a visible pellet)
    - add 2x volumes (400 microliters) 100% ethanol
    - mix well, centrifuge at high speed, decant liquid
    - wash DNA pellet (70% ethanol), dry, dissolve in small vol.  
H<sub>2</sub>O, determine DNA concentration)

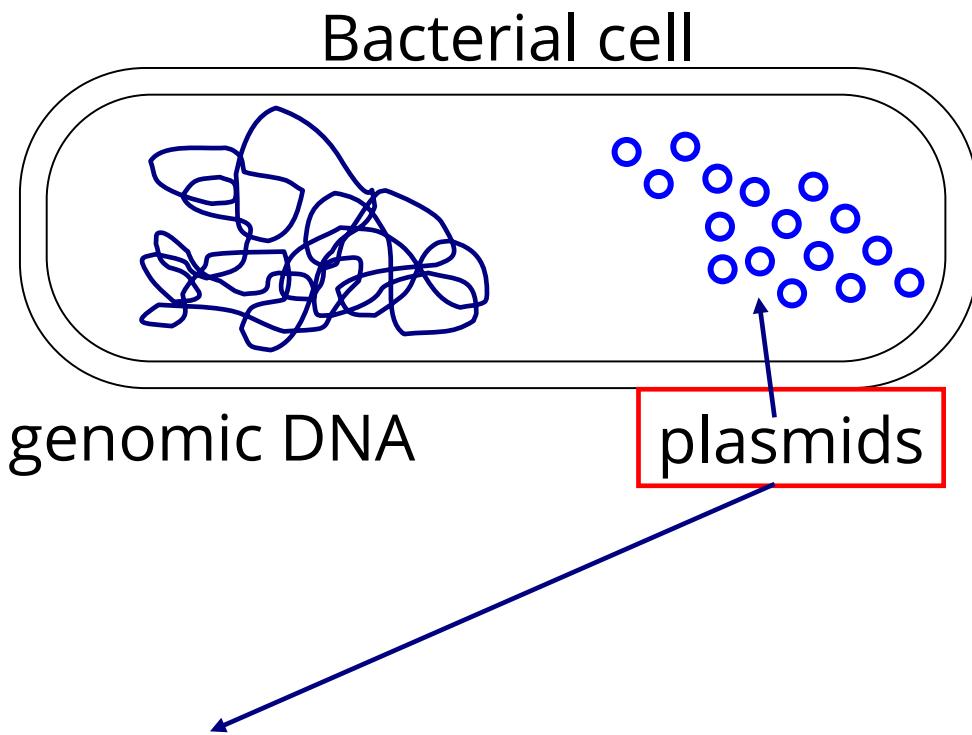
Another way to concentrate DNA solutions (or other large biomolecules)

## Molecular concentrators

- Filter with defined pore size (MWCO, molecular weight cutoff)
- Spin in centrifuge to increase rate of passage through the filter
- Water, salts and other small molecules pass through the filter
- DNA (and anything else larger than the cutoff) does not pass through filter



# Plasmids: essential for recombinant DNA work

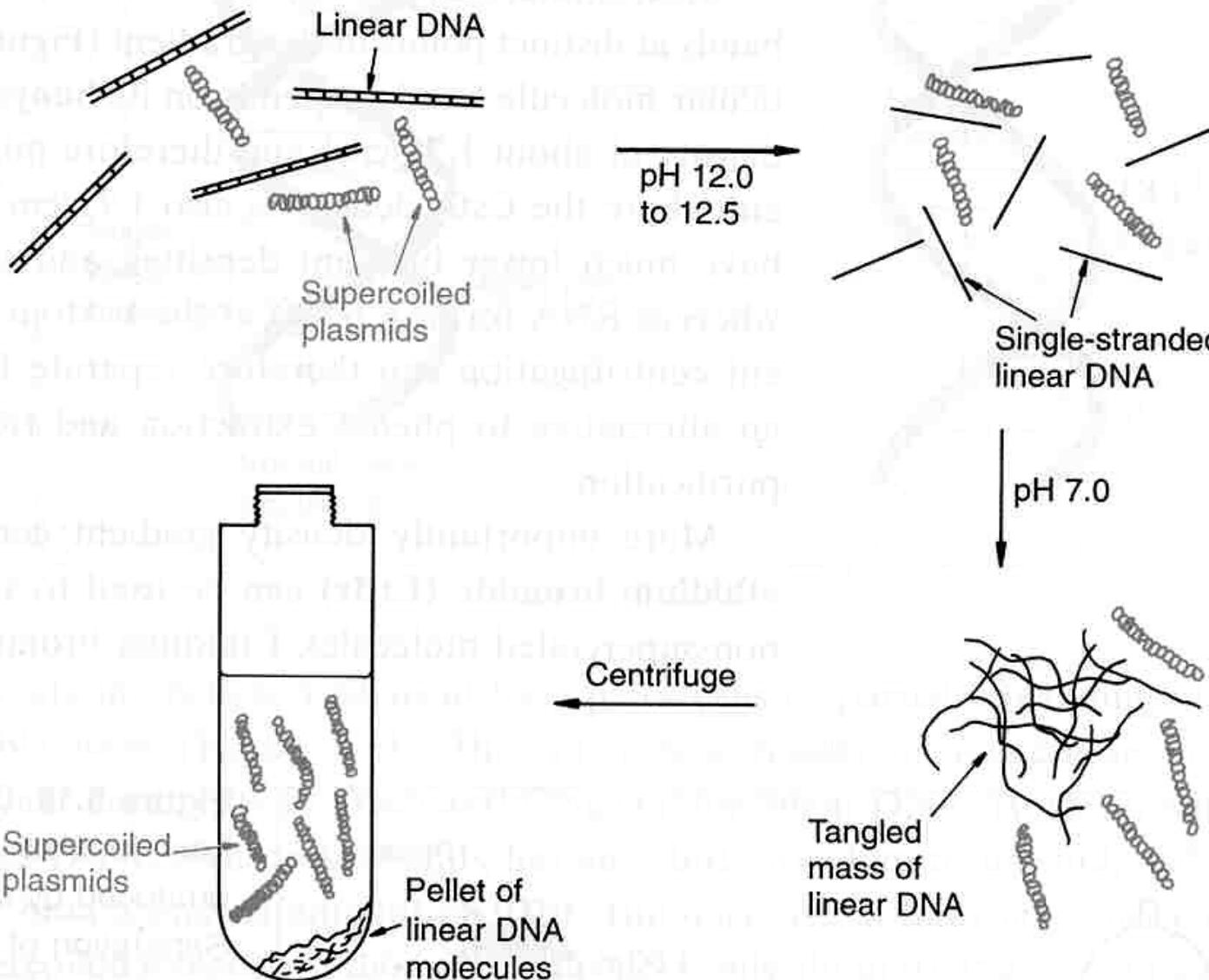


Small pieces of circular DNA that replicate independently of the chromosome

- Many copies per cell
- Easy to isolate and manipulate
- Easy to put back in cells

# Plasmid purification: “alkaline lysis”

Add RNase & break open cells with SDS



Alkaline conditions denature DNA

Neutralize:  
genomic DNA  
can't renature

(plasmids CAN because they never fully separate)

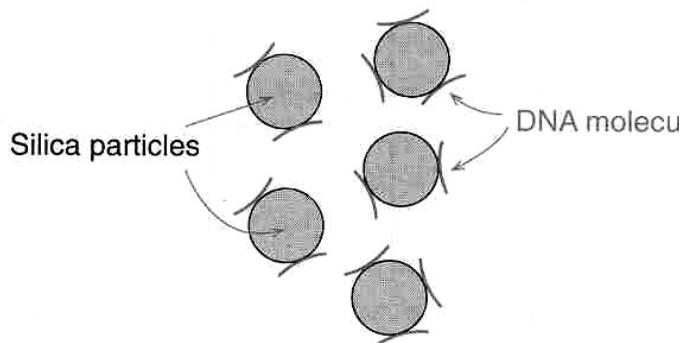
## **Kits for DNA purification:** high reproducibility, saves time

- Guanidinium salts disrupt hydration shell around nucleic acids
- Cations in solution form salt bridges between negative charges of DNA and silica or some other charged resin
- Ethanol (50%) washes away proteins and RNA, but leaves DNA
- Silica/resin can be in form of beads, column matrix, membranes, etc.

# DNA purification: silica binding

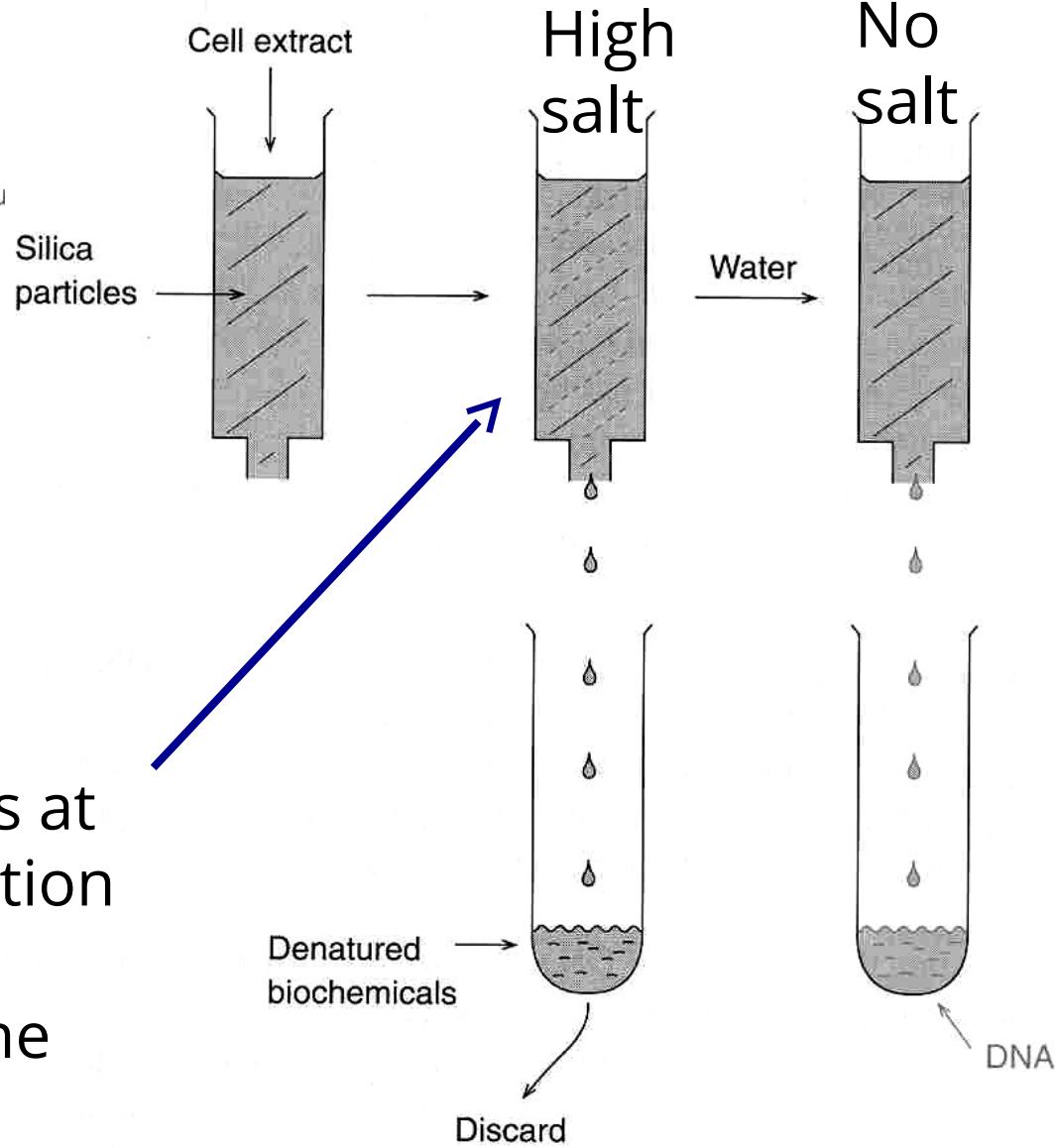
DNA purification by column chromatography

Attachment of DNA to silica particles

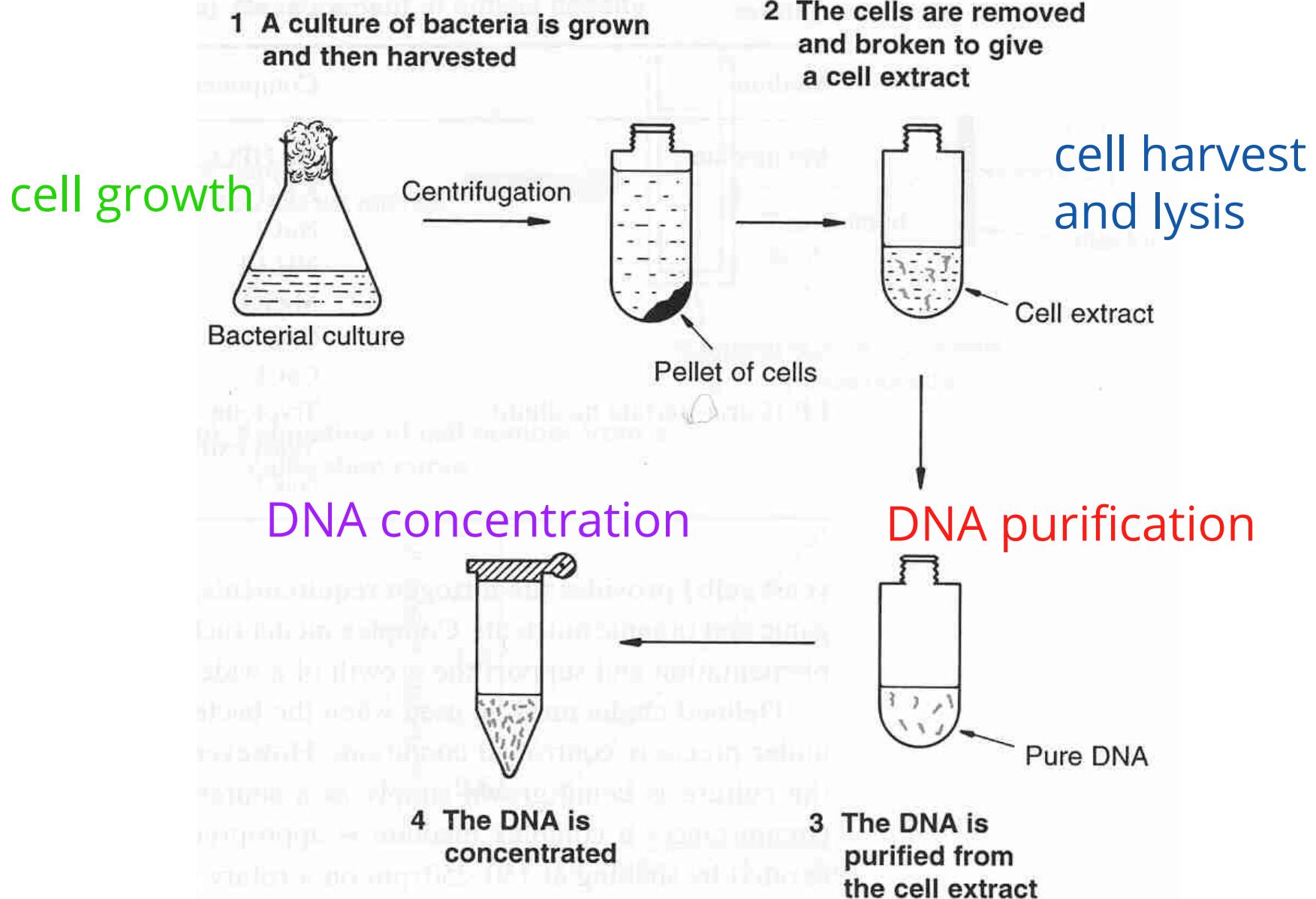


DNA binding occurs at high salt concentration

Low salt disrupts the binding



# DNA purification: overview



# What is aDNA, and what can it tell us?

- aDNA is isolated from archaeological, paleontological remains, museum specimens, etc.
- aDNA provides information for molecular evolution studies
  - Compare DNA sequences of modern organisms to ancestral organisms, trace speciation at the molecular level – example: human evolution
- aDNA can be used to define animal diets, which gives ecological and behavioral information
- aDNA can give information about ancient disease

# DNA sequences from extinct animals: snapshots of genetic information from the past

quagga, marsupial wolf, sabre-toothed cat, moa, mammoth, cave bear, blue antelope, giant ground sloth, Aurochs, mastodon, New Zealand coot, South Island piopio, Steller's sea cow, Neanderthal, Aptornis defossor, Shasta ground sloth, pig-footed bandicoot, moa-nalo and *Myotragus balearicus*



1985



1990



1995



2000

# What happens to nucleic acids following death?

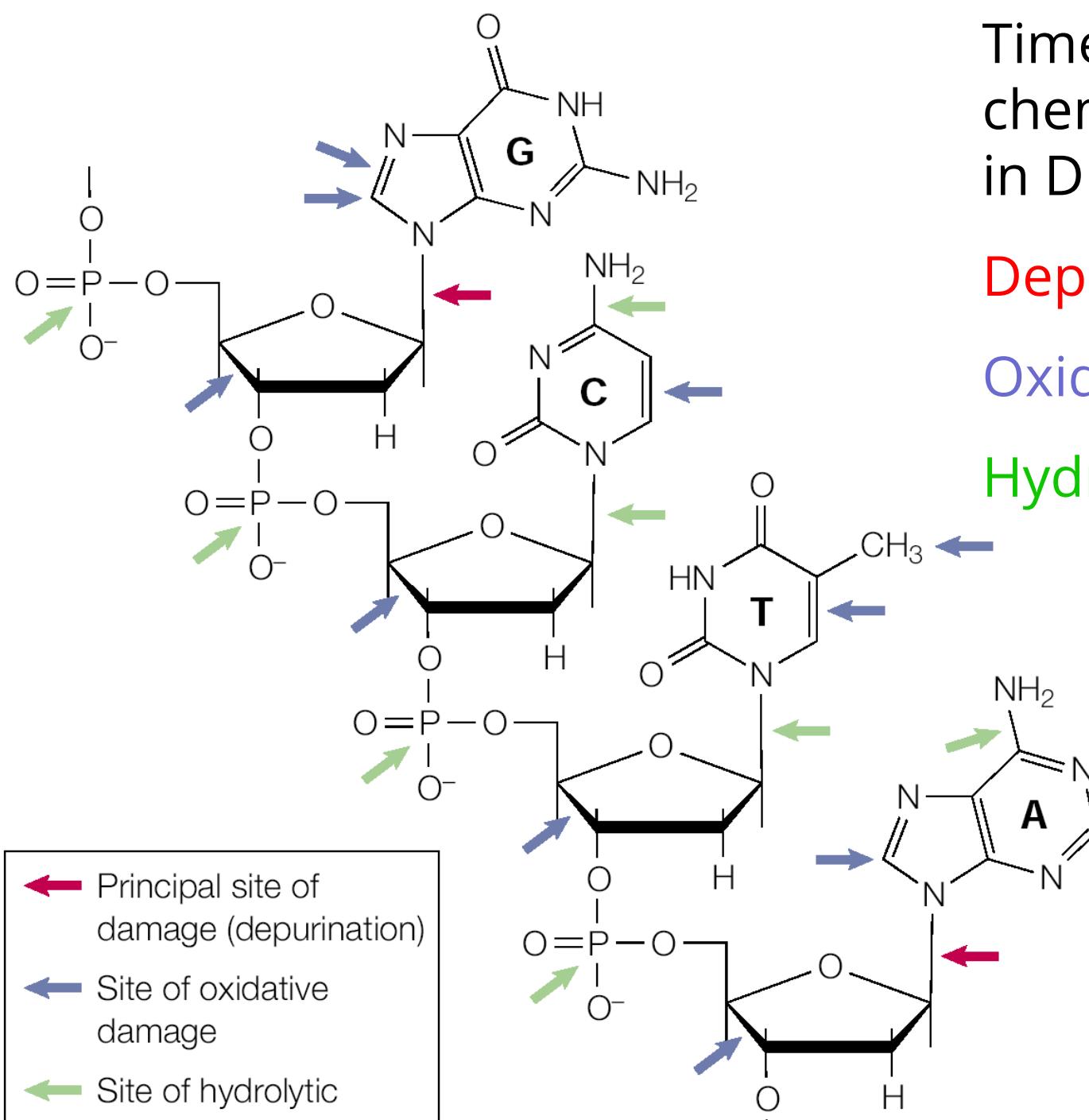
- Rapid decay from the action of nucleases, microbial decomposition
- Occasionally DNA is spared this fate:
  - Rapid dessication
  - Low temperatures
  - High salinity
- However, spontaneous, slow decay is inevitable
  - Depurination (loss of A and G bases)
  - Oxidative damage
  - Hydrolytic damage

Time-dependent  
chemical changes  
in DNA:

Depurination

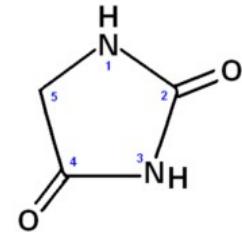
Oxidative damage

Hydrolytic damage



# Effects of DNA damage

- Backbone breakage -- fragmentation
- C and T residues oxidize to make hydantoins, blocking DNA polymerases (PCR)
- Deamination of C causes wrong base to be added during PCR--false mutations
- Increasing time, increasing degradation, decreasing utility
- **100,000 to 1,000,000** years is the approximate age limit for DNA to yield useful sequences



# PCR is good, but not perfect for ancient DNA isolation

- Need very little template DNA
- sequence PCR products directly (no need for cloning)
- Specific genes or DNA regions can be targeted
  - Mitochondrial DNA is typical target in aDNA PCR isolations
  - Copy number of mitochondria is high relative to nuclear DNA

However

- Generally only short pieces of ancient DNA can be amplified, because of damage
- PCR artifacts may cause sequence misreads

# First retrievals of old DNA

- Quagga (extinct relative of the zebra) DNA isolated from museum specimen (Higuchi et al. 1984)
- 2430 year-old Mummy DNA cloned (Paabo 1985)
  - 1) Isolated DNA (20 micrograms/gram mummy tissue)
  - 2) Treated with Klenow enzyme (DNA polymerase) to make DNA fragments blunt ended
  - 3) Cloned into alkaline phosphatase treated pUC8 (pMUM plasmids)

\*\*\*Cloning presents problems, eg. repair of mutagenized DNA following transformation, which gives false sequence

DNA -----> mRNA -----> protein

### Information from mRNA:

When is a gene expressed?

How is the mRNA spliced?

What is the timing of gene expression?

What is the level of gene expression?

Isolation of RNA – Molecular Cloning reading

# RNA in a typical eukaryotic cell:

---

$10^5$  micrograms RNA

80-85% is ribosomal RNA

15-20% is small RNA (tRNA, small nuclear RNAs)

About 1-5% is mRNA

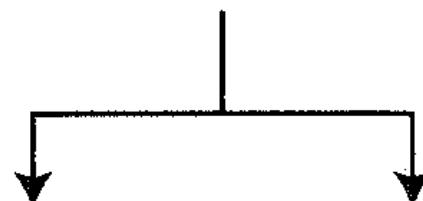
- variable in size
- usually contains 3' polyadenylation

# Making and using mRNA (1)

Isolation + purification

Analysis of RNA  
(Protocols 12–18)

Mammalian cells + tissues  
and lower eukaryotes

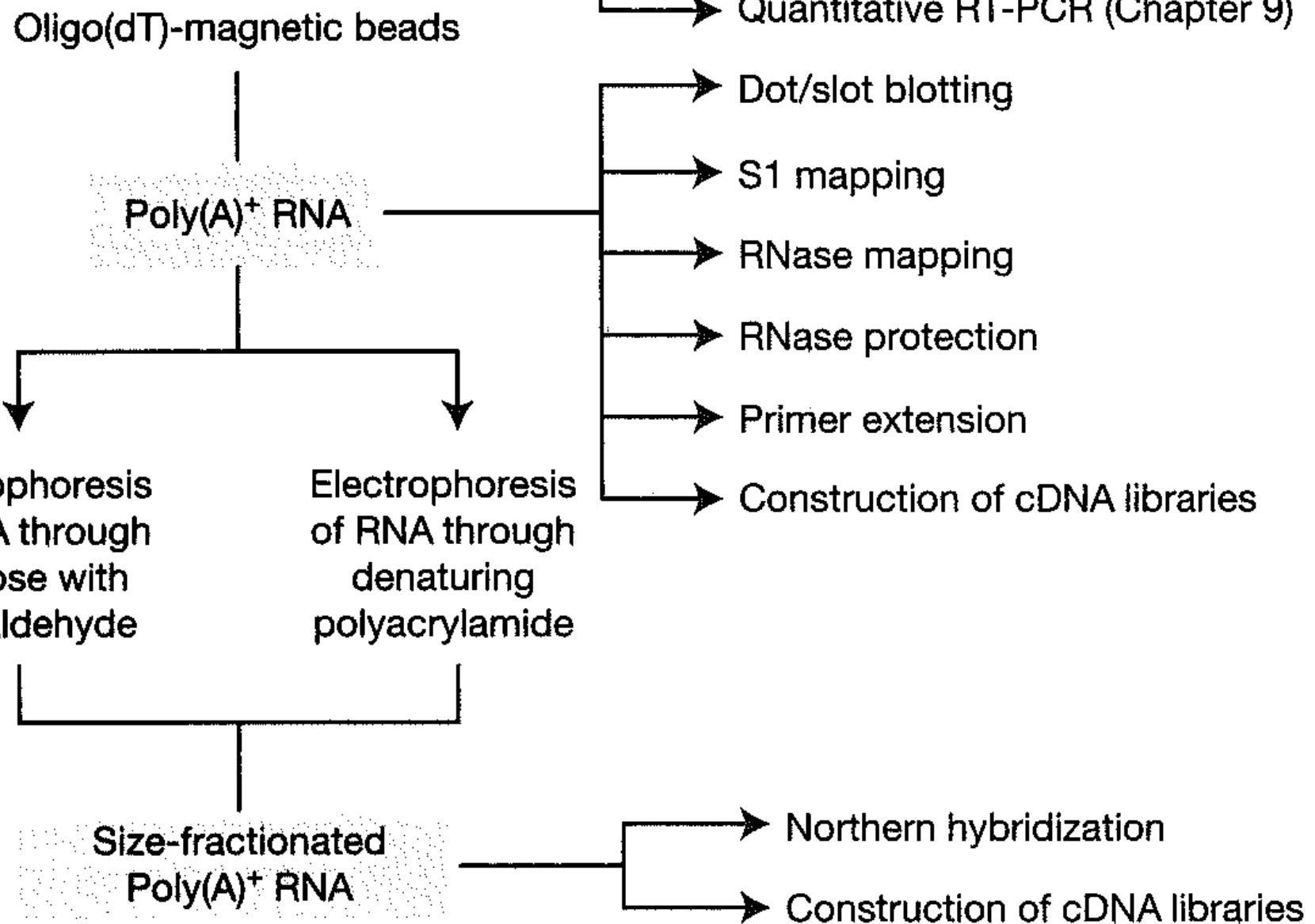


Monophasic lysis      Hot acid-phenol extraction

- Northern hybridization
- Dot/slot blotting
- RNase protection
- Construction of cDNA libraries
- Isolation of small noncoding RNAs (Chapter 18)
- Quantitative RT-PCR (Chapter 9)

Oligo(dT)-magnetic beads

# Making and using mRNA (2)



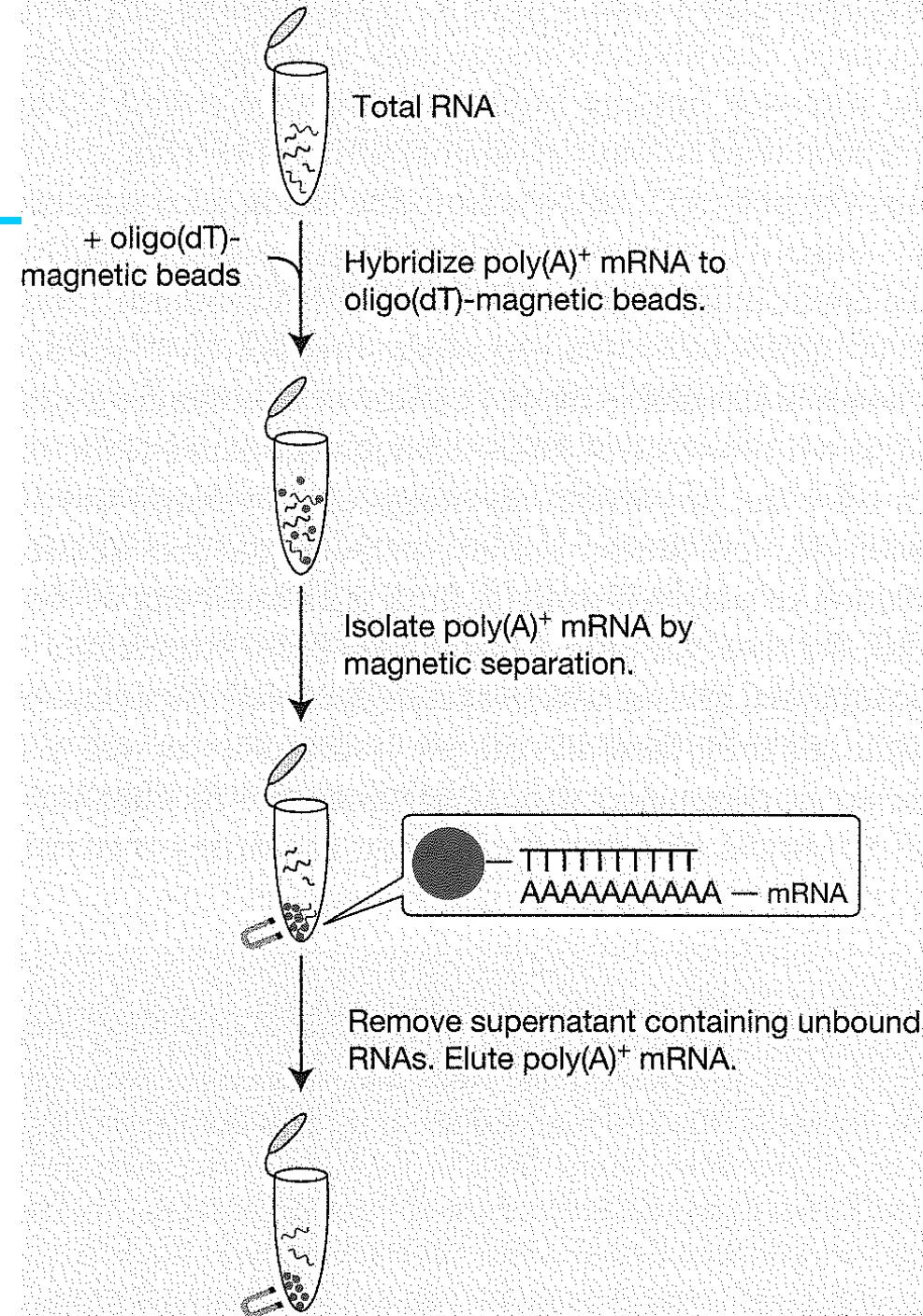
# Selective capture of mRNA: oligo dT-cellulose

Oligo dT is linked to magnetic beads

RNA is mixed with beads at high salt concentration

Non-polyadenylated RNAs do not hybridize to beads

polyA RNA can be removed under low salt conditions



# The problem(s) with RNA:

---

## RNA is chemically unstable

spontaneous cleavage of phosphodiester backbone via intramolecular transesterification

## RNA is susceptible to nearly ubiquitous RNA-degrading enzymes (RNases)

RNases are released upon cell lysis

RNases are present on the skin

RNases are very difficult to inactivate

- Very stable (disulfide bridges)

- Divalent cations not needed for activity

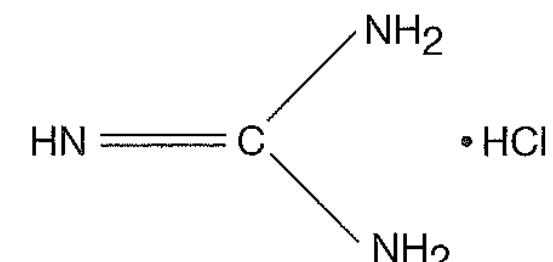
# Purifying RNA: the need for speed

Break the cells/solubilize components/inactivate RNases by the addition of guanidinium salts (very powerful protein denaturant)

Extract RNA using phenol/chloroform (at low pH)

OR use a monophasic lysis reagent (trizol):

- guanidinium salt
- acidified phenol
- phenol solubilizer (e.g. glycerol)



Guanidinium  
hydrochloride

Precipitate the RNA using ethanol/LiCl

Store RNA:

- in DEPC-treated  $\text{H}_2\text{O}$  (-80°C)
- in formamide (deionized) at -20°C

# Storage of RNA: need to prevent degradation

In solution:

- pH 7 to 7.6
- SDS (sodium dodecyl sulfate), a detergent that inhibits RNases
- EDTA (Ethylene Diamine Tetra Acetate) captures (chelates) metal ions (e.g.  $Mg^{+2}$ ) that can degrade RNA
- Store at very low temperatures (-80°C)

Long term: store in ethanol precipitation conditions at -80°C. The low temp and high alcohol concentration reduces RNase activity

# Inhibitors of Rnase

---

## DEPC: diethylpyrocarbonate

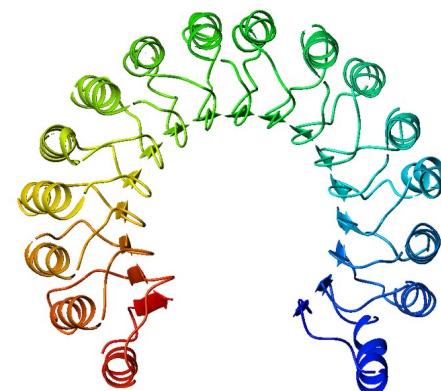
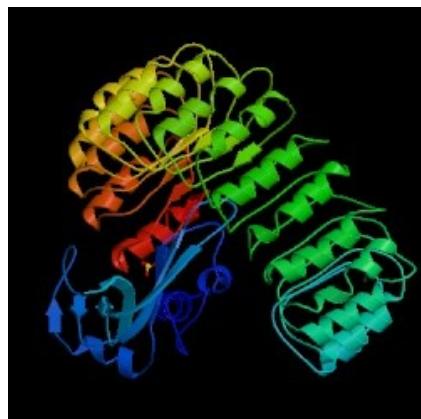
- alkylating agent, modifies and inactivates enzymes, including RNases

## Modified ribonucleoside complexes

- competitive inhibitors that bind to RNase enzyme active site

## Protein inhibitors of RNase

- horseshoe-shaped, leucine rich protein, found in cytoplasm of most mammalian tissues



# 10 common sources of RNase contamination

- 1) Ungloved hands
- 2) Tips and tubes
- 3) Water and buffers
- 4) Lab surfaces
- 5) Endogenous cellular RNases
- 6) RNA samples
- 7) Plasmid preps
- 8) RNA storage (slow action of small amounts of RNase)
- 9) Chemical nuclease action ( $Mg^{+2}$ ,  $Ca^{+2}$  at 80°C for 5' +)
- 10) Impure enzyme preparations

# Common sources of RNase and how to avoid them

---

## Contaminated solutions/buffers

Use good sterile technique

Treat solutions with DEPC (when possible)

Make small batches of solutions, and don't reuse

## Contaminated equipment

Use "RNA only" pipets, glassware, etc

Bake glassware, 300°C, 4 hours, to 'kill' RNases

USE " RNase-free" pipet tips

Treat equipment with DEPC

# DNA and RNA isolation and purification

---

- I. How are biomolecular separations accomplished in general?
- II. How is genomic DNA prepared?
- II. How is plasmid DNA prepared?
- III. How can DNA be separated from other cell components?
- IV. How can RNA be isolated successfully?

# Visualizing DNA (and RNA, protein): non-specific (bulk) detection methods

- I. Quantitation of nucleic acids (chemical properties: bases, dye binding)
- II. Electrophoresis ( $\text{PO}_4^-$  groups, size)
- III. Visualizing macromolecules (e.g. dye binding)

Note:

Many protocols can be found at <http://openwetware.org>

# Guide to readings: DNA & protein visualization (non-specific)

1) 3 MC4 DNA quantitation. Discussion of UV spectroscopy, and stain-based methods for DNA analysis.

## 2) Electrophoresis to separate DNA and proteins

- 4 MC4 Agarose electrophoresis. Details of agarose gels for DNA analysis.
- 5 MC4 Polyacrylamide gel electrophoresis (PAGE). Protocol for DNA PAGE gels.
- 6 MC4 SDS-PAGE for proteins. Protocol for separation of proteins on polyacrylamide gels.

## 3) Staining to reveal biomolecules

- 7 MC4 vis DNA. Stains for DNA visualization. Also, biotinylation, and the use of magnetic beads for DNA
- 8 MC4 vis protein. Stains for protein visualization.

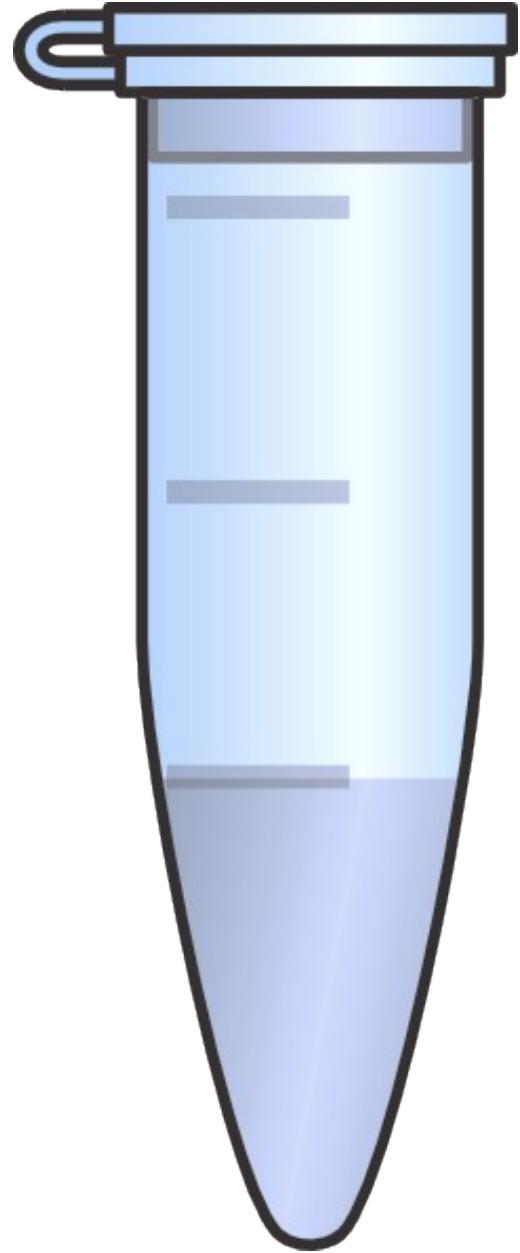
You just purified some genomic DNA

It's dissolved in water.

How much DNA is in there?

Do you have enough to proceed with the next step in the protocol?

Do you have the right DNA?



DNA absorbs short wavelength light (260 nm). The more DNA in a solution, the less light gets through

The Beer-Lambert law:

$$I = I_o 10^{-\varepsilon dc}$$

How much light gets through a solution depends on what's in it and how much there is

$I$  = intensity of transmitted light

$I_o$  = intensity of incident light

$\varepsilon$  = molar extinction coefficient

$d$  = optical path length

$c$  = concentration of absorbing material

DNA and RNA have specific  $\varepsilon$ 's

## The Beer-Lambert law

$$I = I_o 10^{-\varepsilon d c}$$

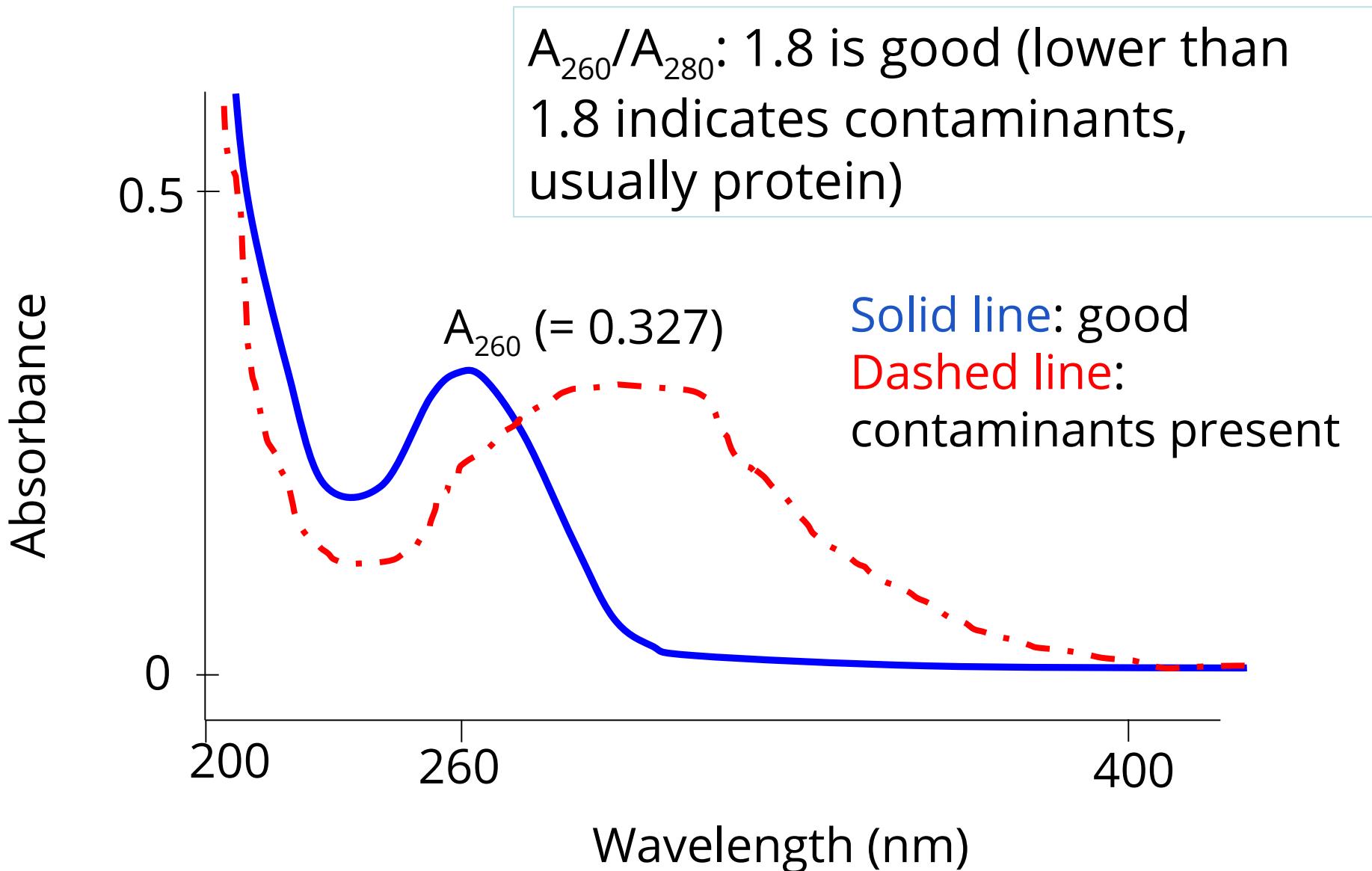
Absorbance: A measured by a spectrophotometer is  
 $\log I/I_o$

A is called optical density (OD) when the path length *d* is 1 cm

If you know the  $\varepsilon$  of the substance, the absorbance of a solution will tell you its concentration:

$$\text{OD}_\lambda = \varepsilon c$$

# A " scan" (multiple wavelength absorbance measurement) of a DNA sample



# Quantitation of DNA: UV absorbance at 260 nm

- Nucleic acids: the aromatic bases have a characteristic absorbance maximum at around 260 nanometers
- Sample must be pure for accurate measurements (RNA, EDTA, phenol, etc. absorb at 260 nm)
- For a reading of 1.0 A<sub>260</sub> (1 cm light path)
  - DNA (double stranded) is 50 micrograms/milliliter
  - DNA (single stranded) 33 micrograms/milliliter
  - RNA is 40 micrograms/milliliter
  - <http://nebiocalculator.neb.com/#!/od260>

# SI Unit prefixes and symbols

Factor	Prefix	Symbol	Example
$1,000,000,000 = 10^9$	giga	G	1 gigameter (Gm) = $10^9$ m
$1,000,000 = 10^6$	mega	M	1 megameter (Mm) = $10^6$ m
$1,000 = 10^3$	kilo	k	1 kilogram (kg) = $10^3$ g
$100 = 10^2$	hecto	h	1 hectogram (hg) = 100 g
$10 = 10^1$	deka	da	1 dekagram (dag) = 10 g
$0.1 = 10^{-1}$	deci	d	1 decimeter (dm) = 0.1 m
$0.01 = 10^{-2}$	centi	c	1 centimeter (cm) = 0.01 m
$0.001 = 10^{-3}$	milli	m	1 milligram (mg) = 0.001 g
$*0.000\,001 = 10^{-6}$	micro	$\mu$	1 micrometer ( $\mu$ m) = $10^{-6}$ m
$*0.000\,000\,001 = 10^{-9}$	nano	n	1 nanosecond (ns) = $10^{-9}$ s
$*0.000\,000\,000\,001 = 10^{-12}$	pico	p	1 picosecond (ps) = $10^{-12}$ s
$10^{-15}$	femto	f	1 femtomole (fmol)= $10^{-15}$ mole

\* For very small numbers, it is becoming common in scientific work to leave a thin space every three digits to the right of the decimal point.

$10^0$  = the unit of measurement: -mole, -meter, -gram, etc.

Convert  $A_{260}$ :

- to micrograms/ml
- to molar concentration

Example:

sample of 250 bp fragment of double stranded DNA

$$A_{260} = 0.327$$

What is the DNA concentration?

( $1.0 A_{260} = 50$  micrograms/ml double stranded DNA)

$$\text{DNA conc.} = 0.327 \times 50 = 16.35 \text{ micrograms/ml}$$

# Molar concentration of a DNA solution

Average molecular weight (MW) per base pair = 650

250 base pair DNA MW =  $1.6 \times 10^5$ , so

So  $1.6 \times 10^5$  grams of this DNA fragment per mole

Solve for molarity (moles/liter):  $1.02 \times 10^{-7}$  M

Convert to a less unwieldy notation: 102 nanomolar (nM)

Important to know how to do this calculation and conversion

Or go to: <http://nebiocalculator.neb.com/#!/dsdnaamt>

What is the molarity of a 16.35 microgram/ml solution of a 250 base pair DNA fragment?

16.35 micrograms	1000 ml	1 gram	1 mole
1 ml	1 L	$10^6$ micrograms	$1.6 \times 10^5$ grams

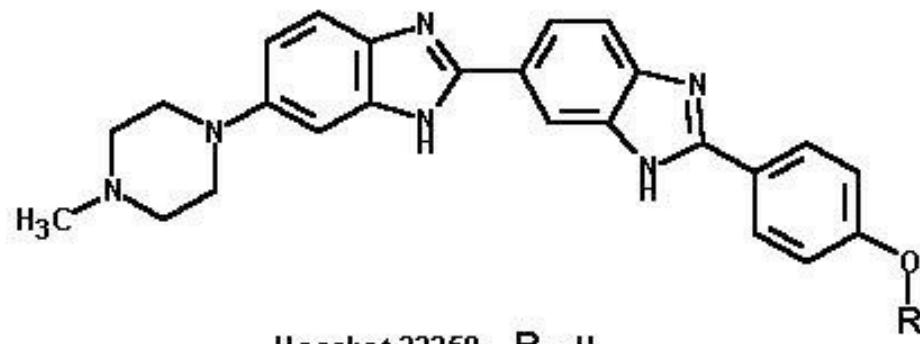
$1.02 \times 10^{-7}$  molar

$0.102 \times 10^{-6}$  molar [0.1 micromolar ( $\mu\text{M}$ )]

$102 \times 10^{-9}$  molar [102 nanomolar (nM)]

# Fluorometry: another method for quantitation of DNA

- Hoechst 33258 (a fluorescent dye) binds to DNA in the minor groove (without intercalation)
- Fluorescence increases after DNA binding
- Good for quantitation of low concentrations of DNA (10-250 ng/ml [pg/ $\mu$ l])
- rRNA and protein do not interfere
- Requires a fluorometer

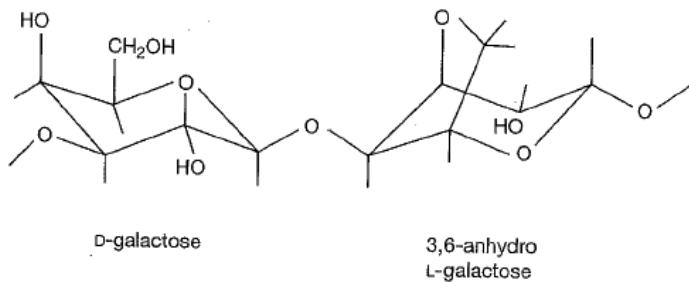


# Visualizing biomolecules: electrophoresis

- Separate biomolecules based on their size
- The separation matrix, or gel (agarose or polyacrylamide), is saturated with an electrically conductive buffer. Samples are loaded, an electric field is applied, and negatively charged biomolecules in the sample travel toward the cathode
- The choice of matrix depends mainly on the size of DNA, being analyzed
- Larger molecules travel slower through the gel matrix
- Dyes allow visual estimate of travel through the gel

# Agarose gels

Agarose: a polysaccharide polymer of alternating D- and L-galactose monomers, isolated from seaweed



Agarose structure

- Pore size is defined by the **agarose concentration** (higher concentration, slower DNA migration overall)
- The **conformation of the DNA** (supercoiled, nicked circles, linear) affects the mobility of the DNA in gels
- Rate of DNA migration is affected by **voltage** (5 to 8 Volts/cm is considered optimal)
- Many kinds of agarose are available (variable melting temperatures, generated by differential hydroxyethylation of the agarose)

# More about gels

**There has to be a buffer/salt** (for carrying current and maintaining pH)

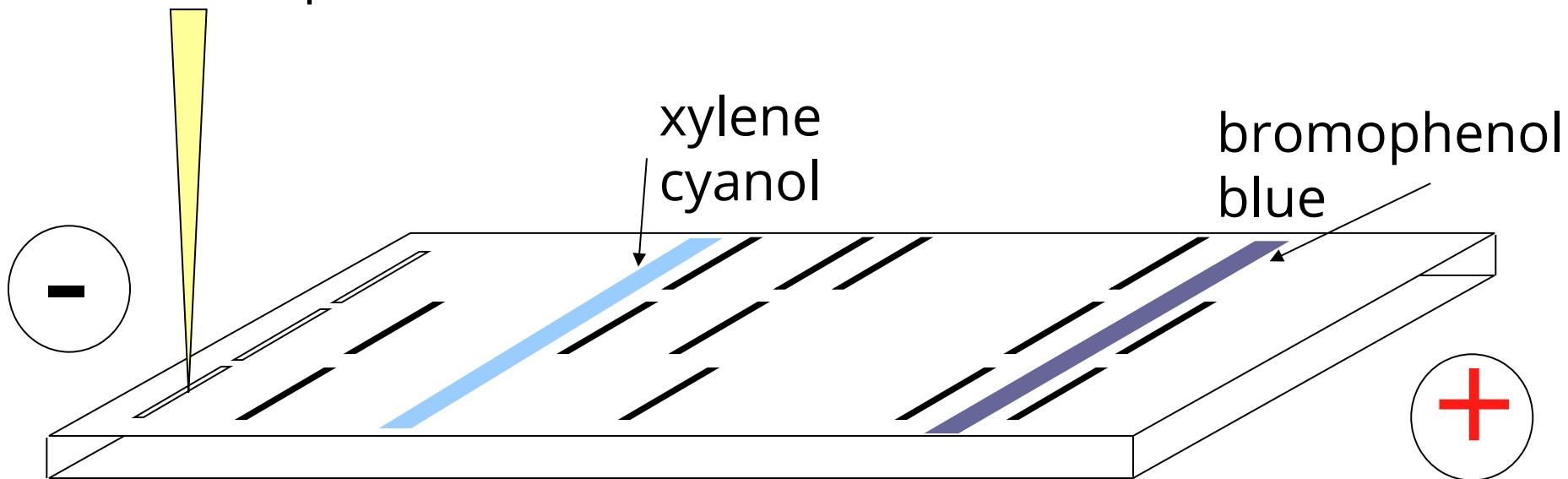
- TAE (Tris-acetate-EDTA): good resolution of DNA, but buffering capacity is quickly depleted
- TBE (Tris-borate-EDTA): High buffering capacity, resolution is pretty good

Use gel loading “ buffers” (relatively simple)

- Dense material to carry sample to bottom of wells (sucrose, glycerol, or ficoll)
- Dyes for tracking progress of electrophoresis
  - Bromophenol blue: fast migration
  - Xylene cyanol: slow migration
- Occasionally denaturant is present (formamide) for denaturing gels (e.g. sequencing gels)

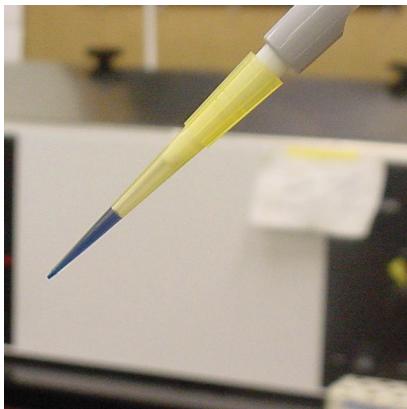
# Typical agarose gel

Load samples in wells

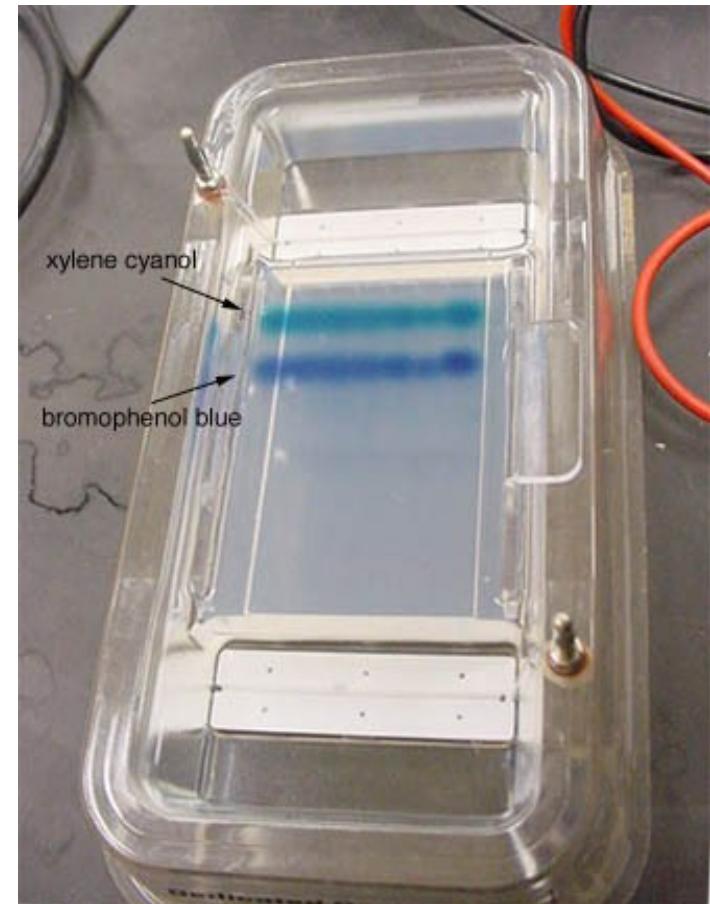
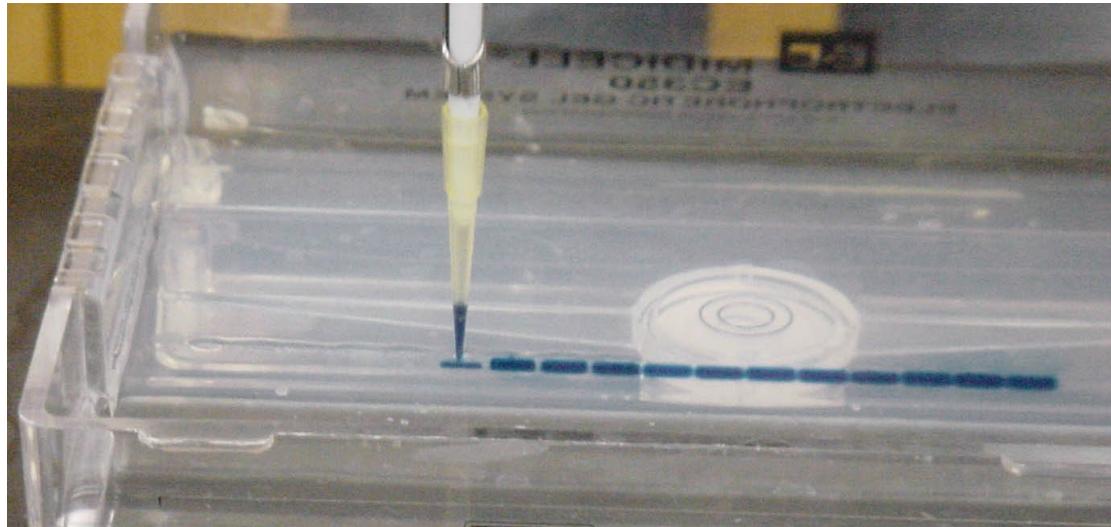


time of electrophoresis  
(progress monitored by marker dyes)

(the DNA fragments  
are not visible  
without some sort  
of staining)



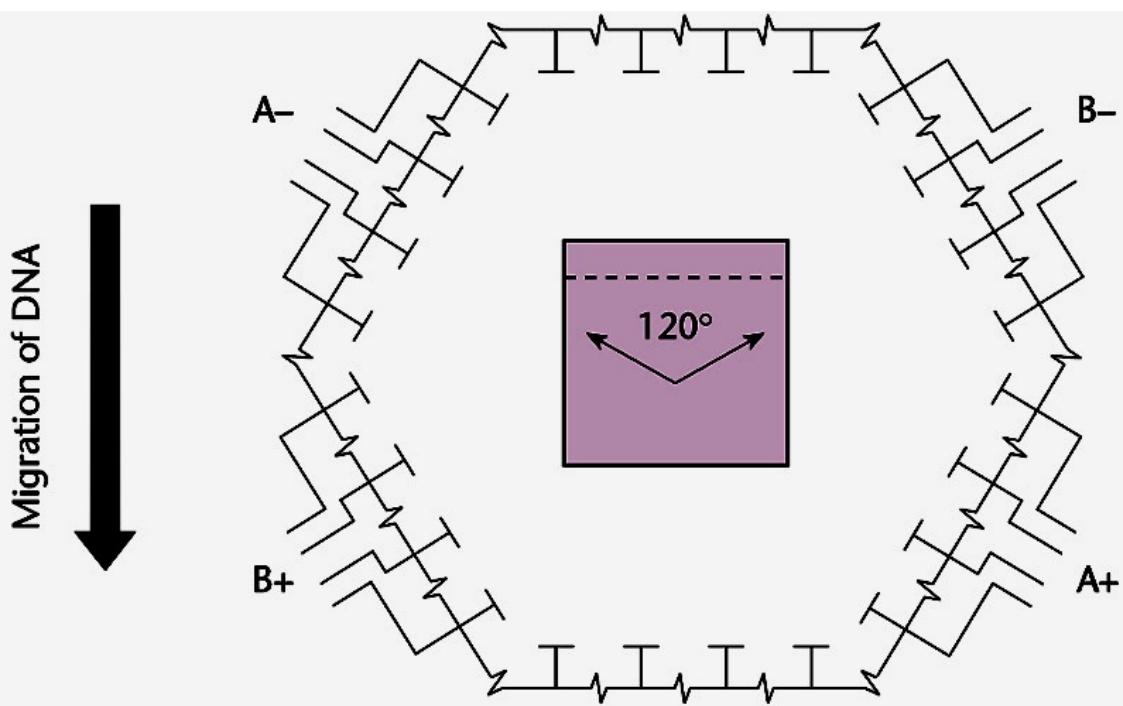
## Agarose gels



# Agarose gels

Standard gels can separate DNA fragments from 100 bp to about 20,000 bp

Pulsed-field gels separate very large DNA fragments (up to 10,000,000 bp, or 10 Mb)



This apparatus induces periodic shifts in the direction of DNA migration: 120° refers to the reorientation angle (difference between orientation of electric fields A and B

# Polyacrylamide gel electrophoresis

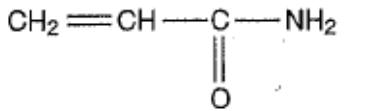
---

- Acrylamide monomers (toxic!) polymerized to form gel matrix
  - The gel structure is held together by the cross-linker-- usually N, N'-methylenebisacrylamide ("bis" for short)
  - Pore size defined by concentration of gel (total percentage) and concentration of the crosslinker (bis) relative to acrylamide monomer
  - Very high resolution (better than agarose)
- 
- Works well for smaller nucleic acids (from 6 to 1000 base pairs in length, RNA or DNA)

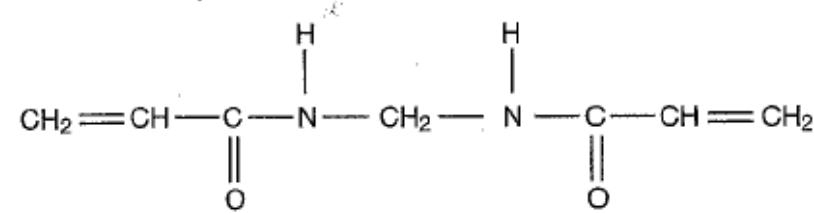
# Polyacrylamide synthesis

Monomers → polymer

acrylamide (**toxic!**)

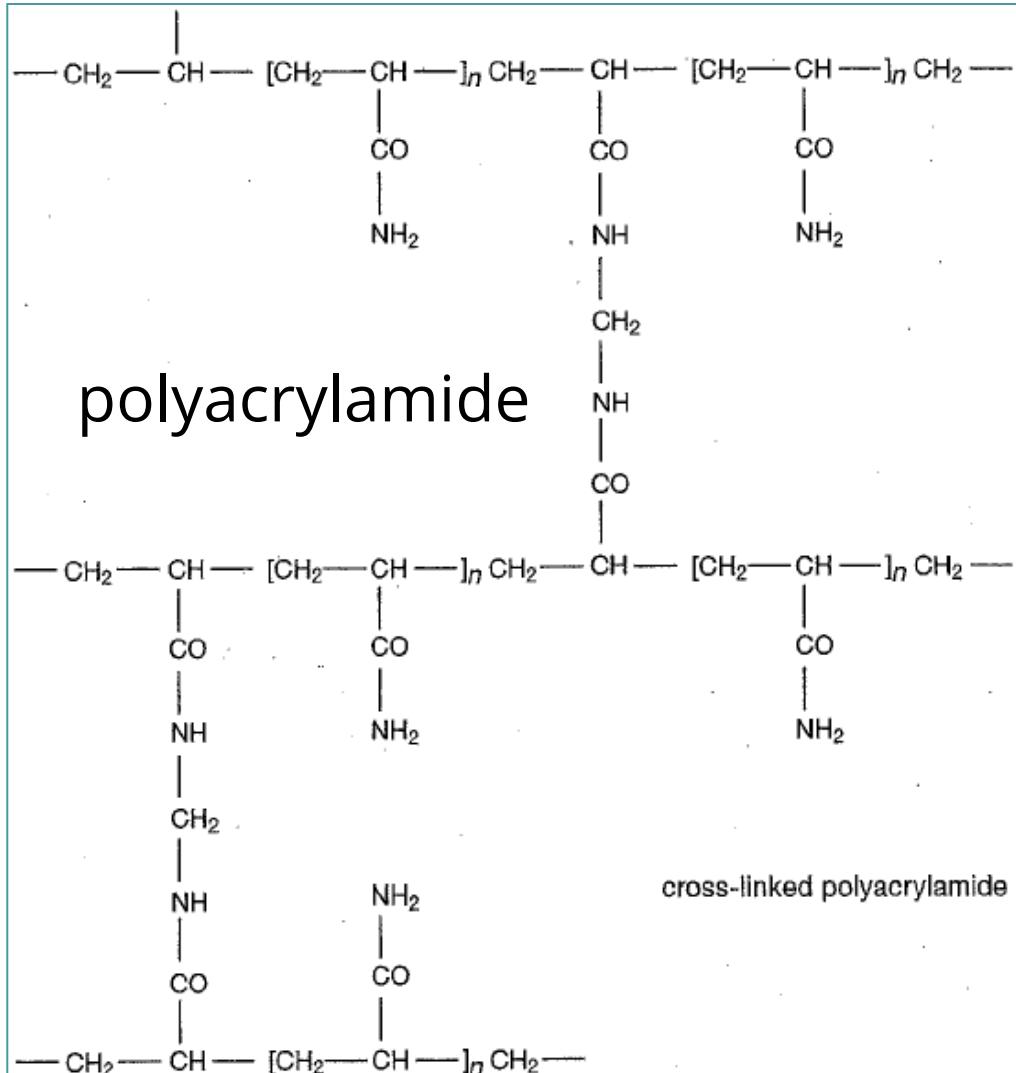


acrylamide



*N,N'*-methylenebisacrylamide

bisacrylamide



# Recipe for a polyacrylamide gel

- Acrylamide (anywhere from 4 to 20 %, depending size of nucleic acids or proteins in the gel)
- Bis-acrylamide (the ratio of Bis to regular acrylamide is important)
- Water
- Buffer

To initiate polymerization, add

## APS: Ammonium persulfate

-- generates free radicals needed for polymerization

## TEMED: N,N,N' ,N' - tetramethylethylenediamine

-- accelerates free radical generation by APS

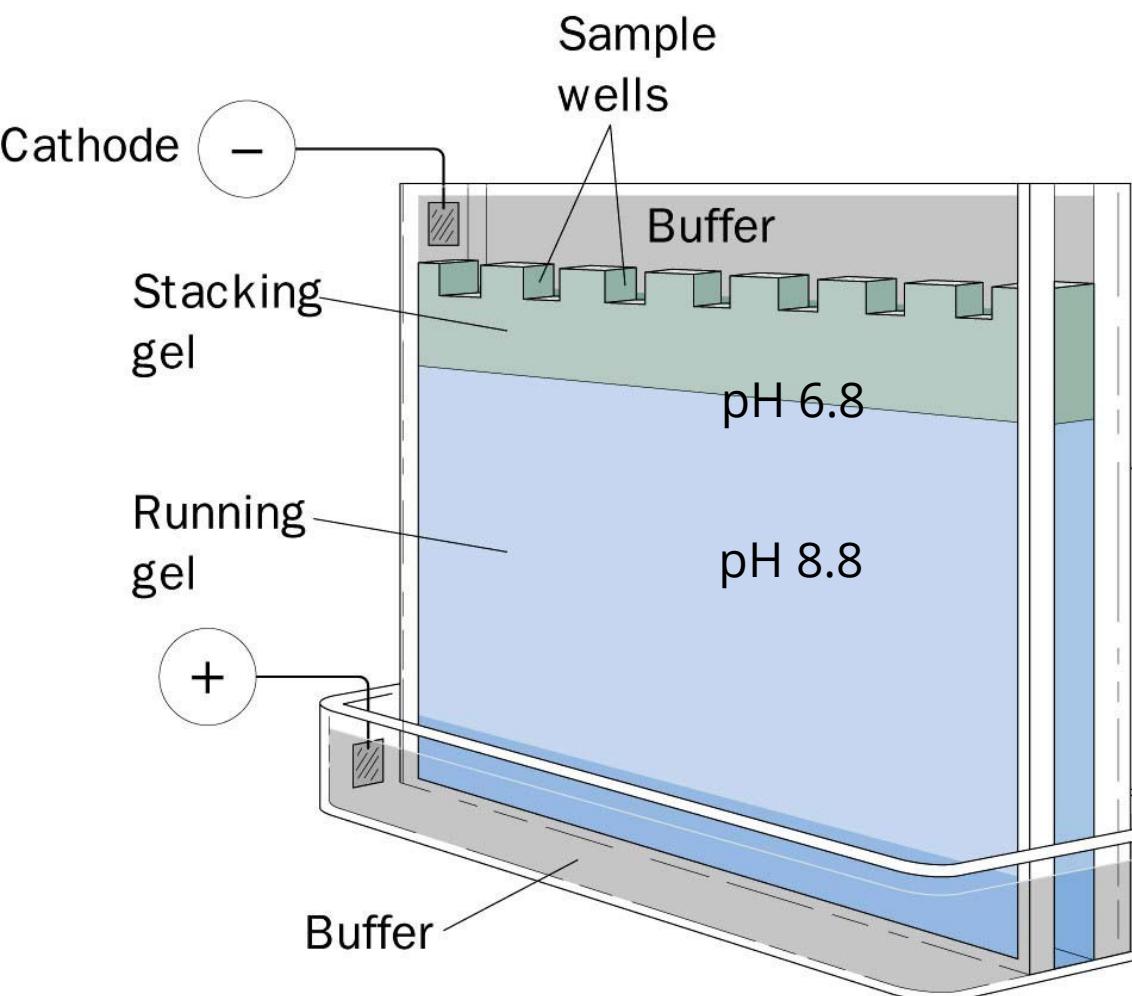
# Polyacrylamide gel electrophoresis

- Native gel (DNA stays double-stranded), or:
- Denaturing gel -- run in the presence of high concentrations of denaturant (usually urea. DNA runs in single stranded form (sequencing gels)

# Protein electrophoresis

- Polyacrylamide gel
- The anionic detergent SDS (sodium dodecyl sulfate) is used to denature the proteins, giving each protein a “ uniform” negative charge
- Protein separation occurs as a function of size
- Discontinuous Tris-Cl/glycine buffer system:
  - o Stacking gel: pH 6.8, low polyacrylamide concentration, focuses proteins into thin layer (gives higher resolution upon separation)
  - o Separating gel: pH 8.8, separates proteins on the basis of size

# Protein gel “SDS-PAGE”



## Stacking gel

At *low pH*, glycine tends to be protonated (no negative charge),  $\text{Cl}^-$  ions form the leading edge, glycine trails, steep voltage gradient in between,

proteins get “ focused” into a thin band (isotachophoresis)

## Separating gel

At *high pH*, glycine deprotonates, runs with the  $\text{Cl}^-$  at the leading edge, and the proteins separate based on size

# UV shadowing:

Detection of nucleic acids directly (no dye required)

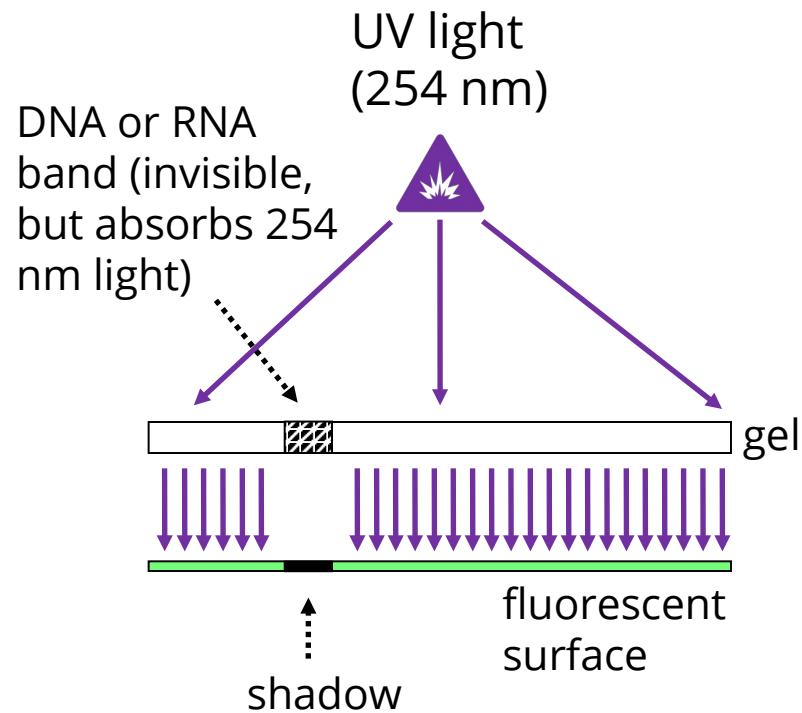
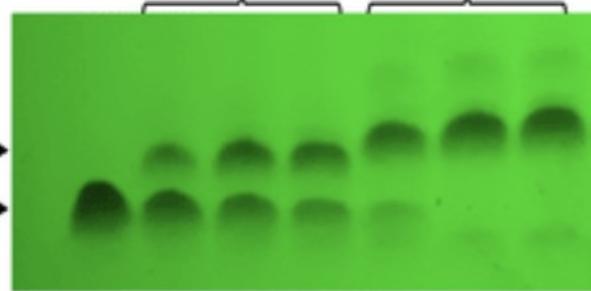
A

lane  
time (h)  
 $\text{Pd(OAc)}_2(\text{L1})_2$  (equiv.)  
boronic ester (equiv.)  
boronic ester

1    2    3    4    5    6    7  
3    6    9    3    6    9

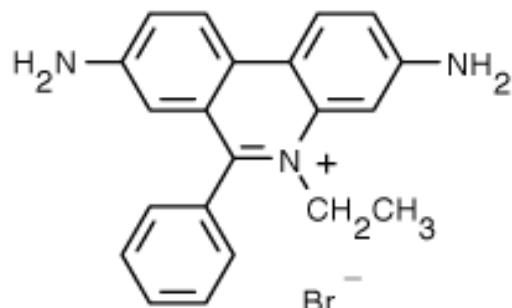
1    100    1  
100    9    100    10

cross-coupled  
RNA product →  
transcript 4 →

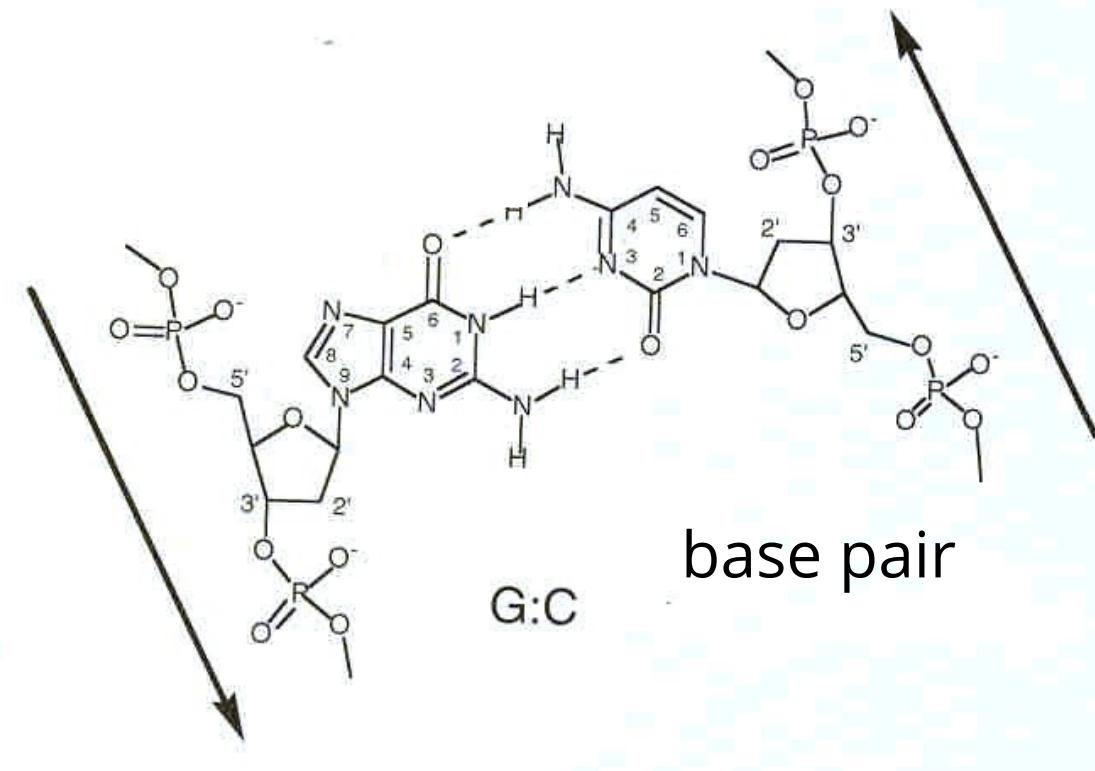


[https://www.researchgate.net/figure/A-Suzuki-reaction-on-iodo-labeled-RNA-ON-4-using-1-equivalent-of-Pd-catalyst-and-100-fig6\\_323830929](https://www.researchgate.net/figure/A-Suzuki-reaction-on-iodo-labeled-RNA-ON-4-using-1-equivalent-of-Pd-catalyst-and-100-fig6_323830929)

# Making nucleic acids visible: stains



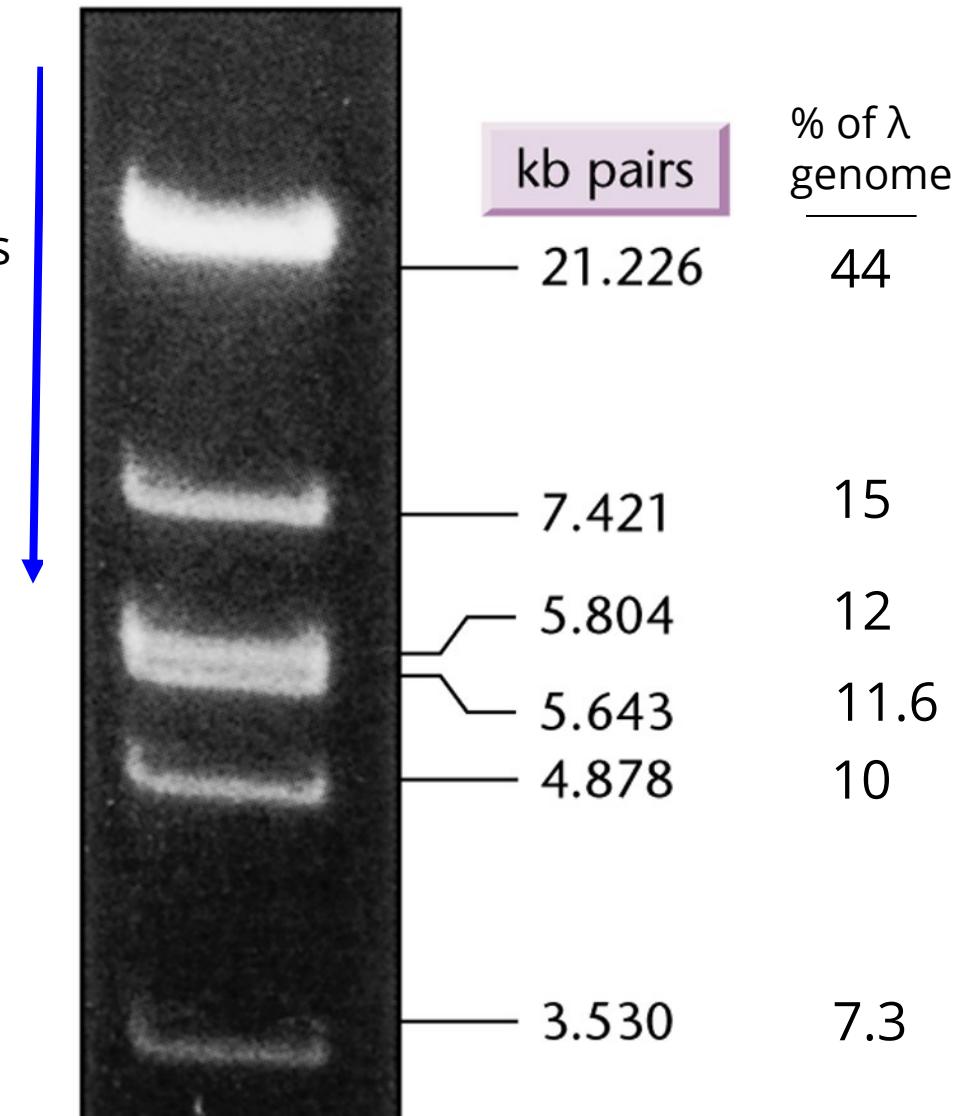
Ethidium bromide



- Ethidium bromide (EtBr) is fluorescent
- EtBr 'intercalates' into stacked base pairs
- Fluorescence increases upon DNA binding
- UV illumination reveals where the DNA is

# Agarose gel stained with ethidium bromide

Direction of  
electrophoresis

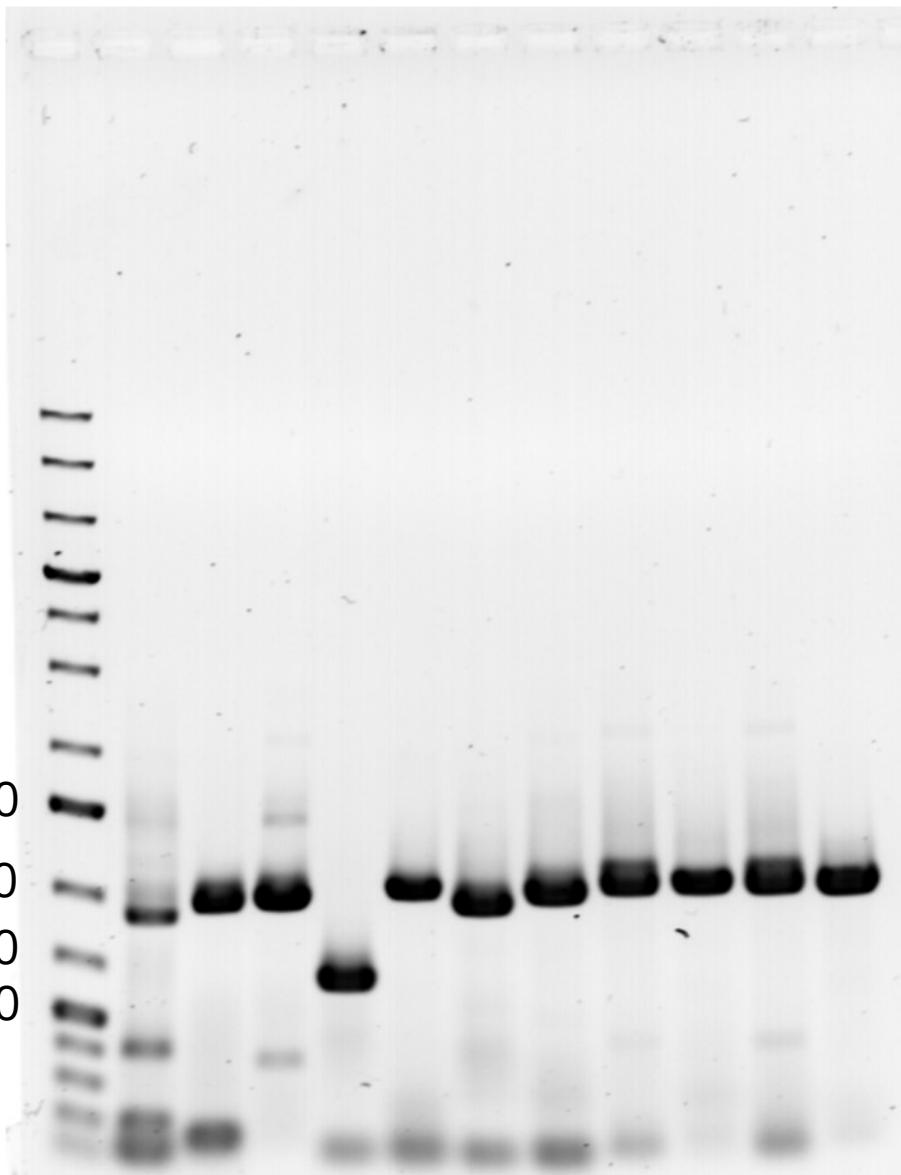


phage  $\lambda$   
genomic DNA  
(48 kb) cut with  
the restriction  
enzyme Hind III

The fragments  
are equimolar--  
why is the band  
intensity  
different?

# Marker DNA gives an estimate of size for samples

M      DNA samples



The marker lane (M) gives size standards for comparison with the sample lanes

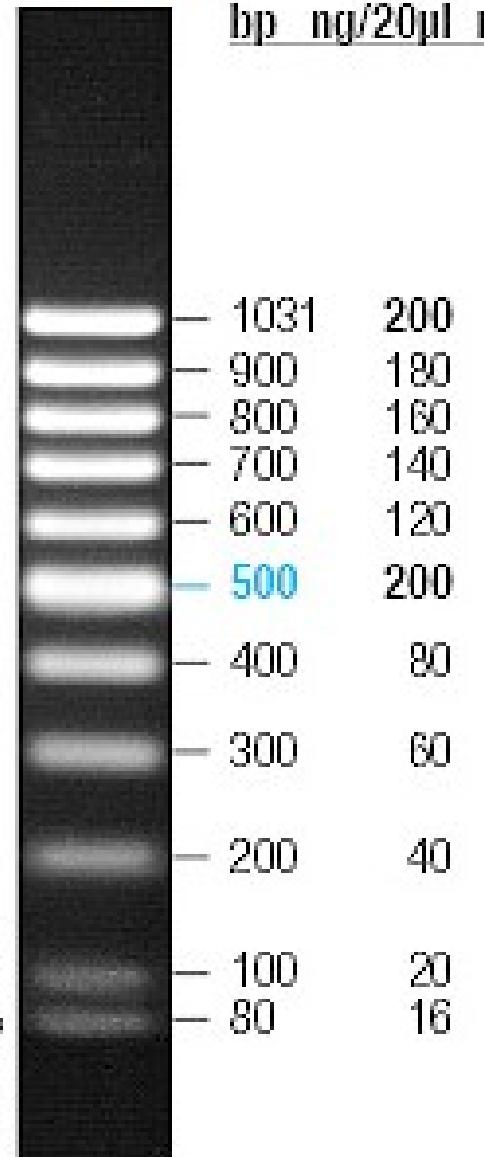
The image was inverted to give black bands on a white background

## Another way to quantify DNA:

Ethidium bromide (fluorescent dye) binding

- Compare sample DNA fluorescence to standards of known concentration (dilution series)
- In solution \*or\* using gel electrophoresis

A commercially available quantitative DNA standard



1.7% agarose  
20µl/lane,  
8cm length gel,  
1X TBE, 5W/cm, 1.5hrs

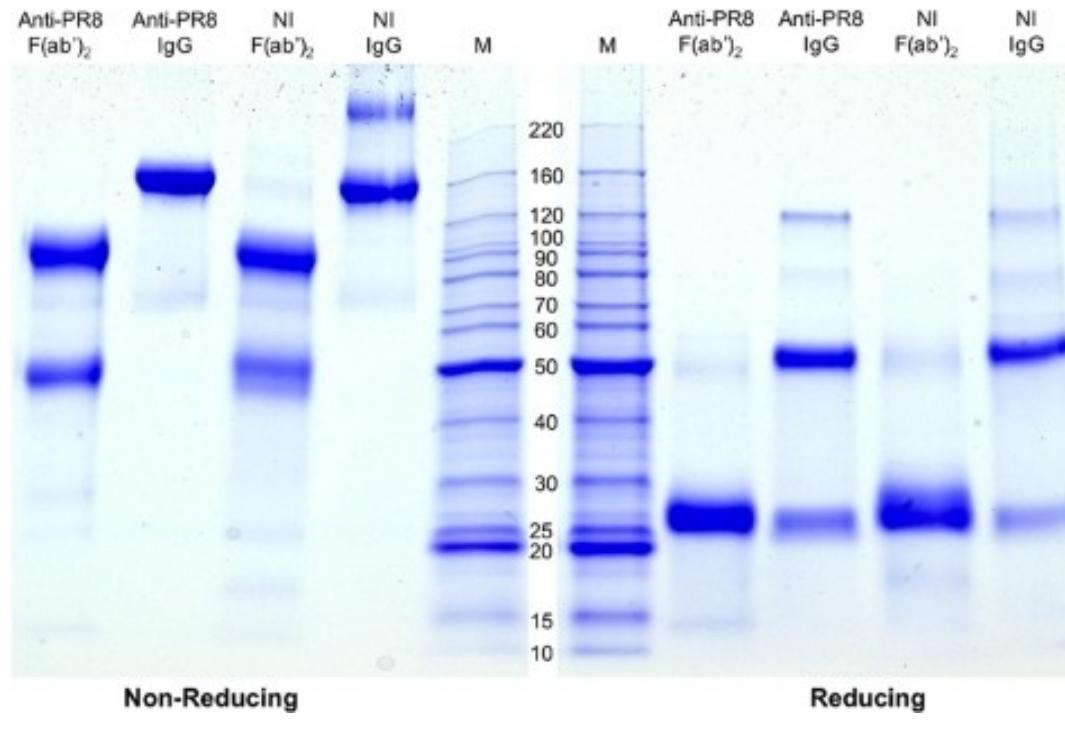
## Other DNA staining options

- methylene blue: staining protocol is time consuming, sensitivity is lower
- SYBR gold and other commercial options: can be more sensitive than ethidium bromide for detecting DNA, but costly

# Protein detection in gels

Coomassie Brilliant Blue R-250: dye from the textile industry that has a high affinity for proteins

- Proteins in gels are “ fixed” (rendered insoluble) with acetic acid/methanol
- Dye probably interacts with NH<sub>3</sub>- groups of the proteins, as well as via van der Waals forces



# Protein detection in gels

## Silver stain:

- 100 to 1000-fold more sensitive than Coomassie (requires far less sample)
- Silver in solution interacts with amino acid side chains and is selectively reduced (similar to early photographic process)
- There is protein-to-protein variability of staining



good



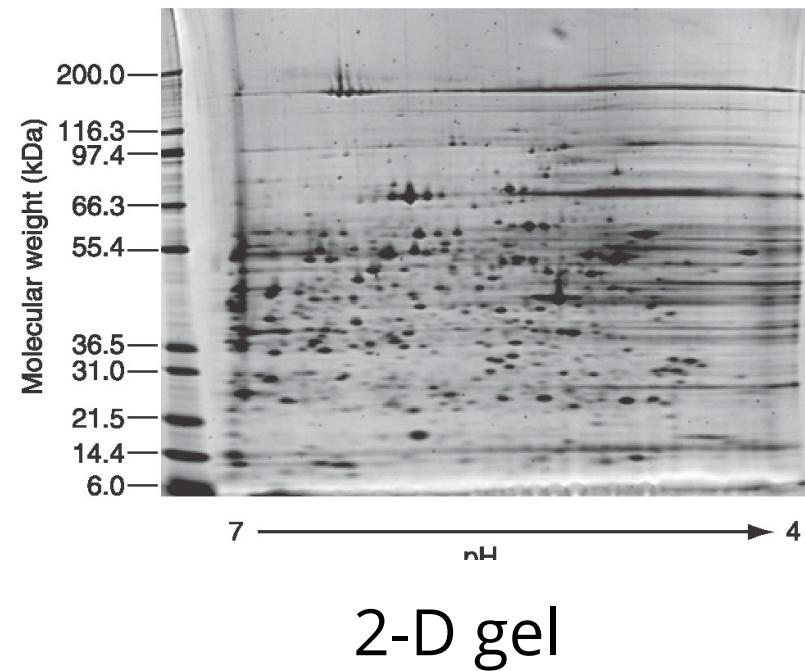
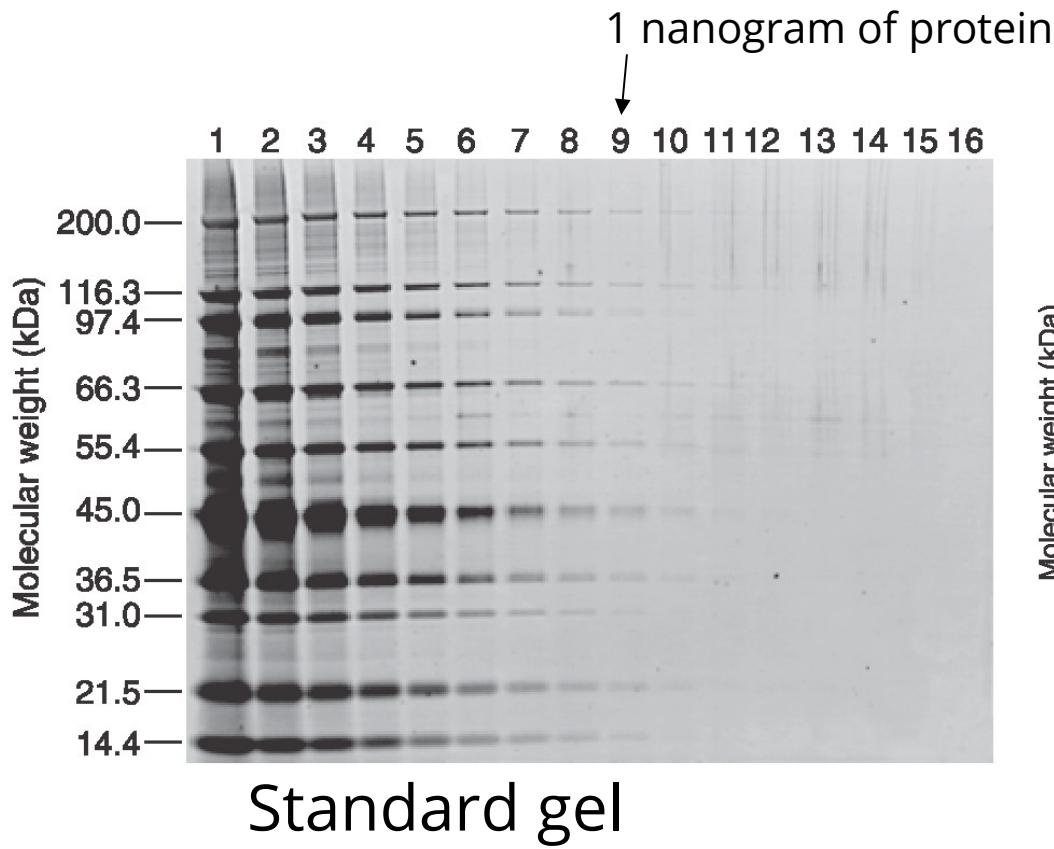
bad (overstained)



ugly (keratin)

# Protein detection in gels

- Sypro Ruby (Molecular Probes inc, proprietary compound)
  - As sensitive as silver staining, less variability
  - Fast protocol
  - \$



# Visualizing DNA (and RNA, protein): non-specific detection methods

- I. Quantitation of nucleic acids (chemical properties: bases, dye binding)
- II. Electrophoresis ( $\text{PO}_4^-$  groups, size)
- III. Visualizing macromolecules (dye binding)

Note:

Many protocols can be found at <http://openwetware.org>

# Detection of specific biomolecules following immobilization: interaction and report

- 1) Nucleic acid hybridization (base pairing)
  - a) Southern blots: DNA-DNA hybridization(Methods for labeling “probe” DNA)
  - b) Northern blots: DNA-RNA hybridization
- 2) Antibody-antigen interactions
  - a) Western blots (detection of proteins with specific antibodies)

## Guide to readings: Specific Biomolecule Detection

- 1) 9 MC4 *Southern blots*. Technique for detecting specific DNA fragments by nucleic acid base hybridization
- 2) 10 MC4 *Northern blots*. Technique for detecting specific RNAs by nucleic acid hybridization
- 3) 11 MC4 *Western blots*. Technique for detecting specific proteins by antibody recognition
- 4) 12 MC4 *Specific detection*. Detailed discussion of the various reagents available for probe detection.

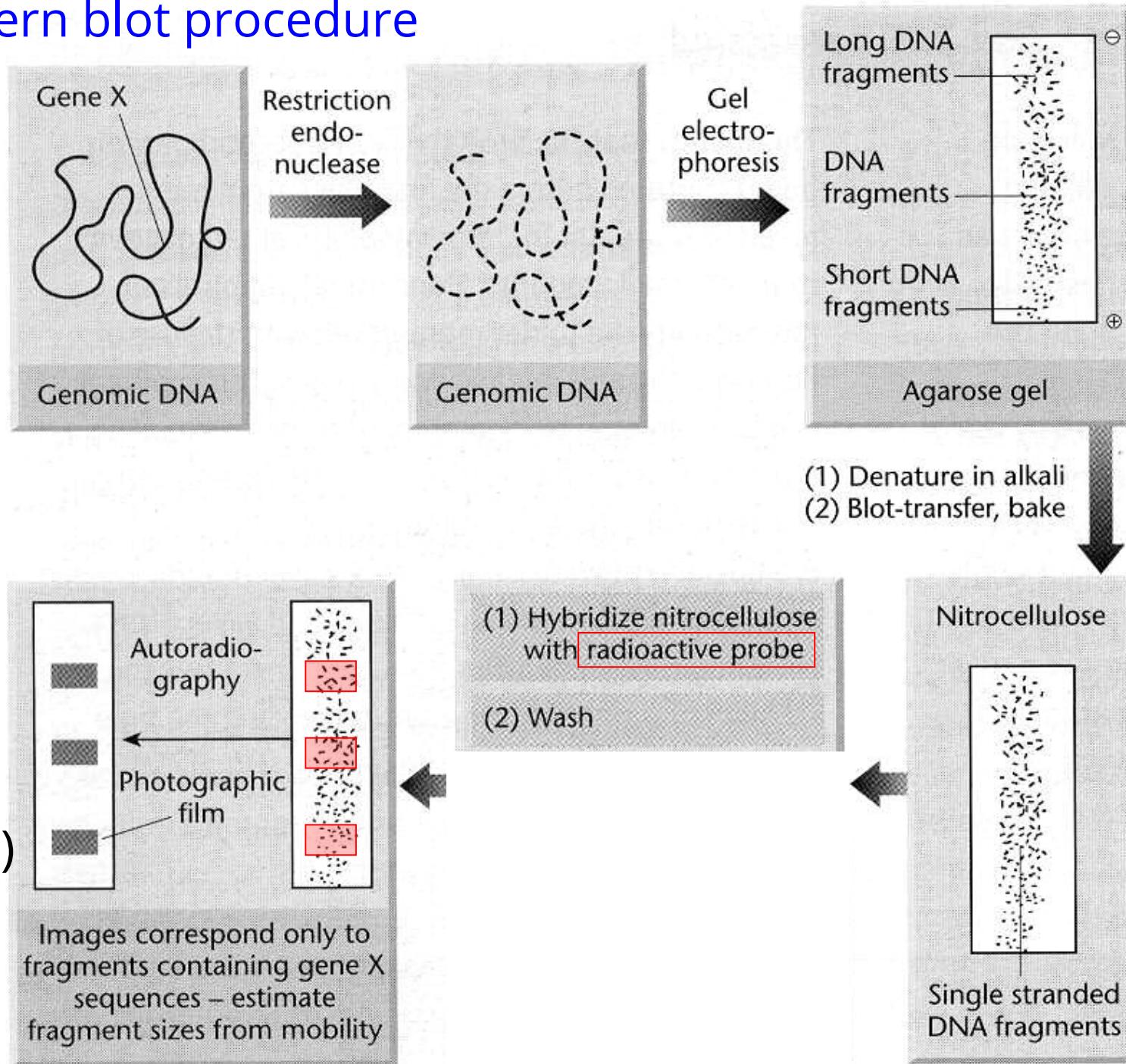
# Visualizing DNA, RNA and Protein: detecting specific sequences or proteins

- Detect **specific** DNA, RNA, or protein in a large, mixed population: cell extracts, genomic DNA preparations, etc.
- For DNA and RNA:
  - specific sequence detection
  - based on DNA and RNA complementarity/base-pairing/hybridization
- For proteins
  - Specific shape/chemistry of the protein
  - Antibodies recognize the protein of interest
  - a specific assay for activity of the protein

## Southern blot: the original method for detecting presence of a specific DNA sequence

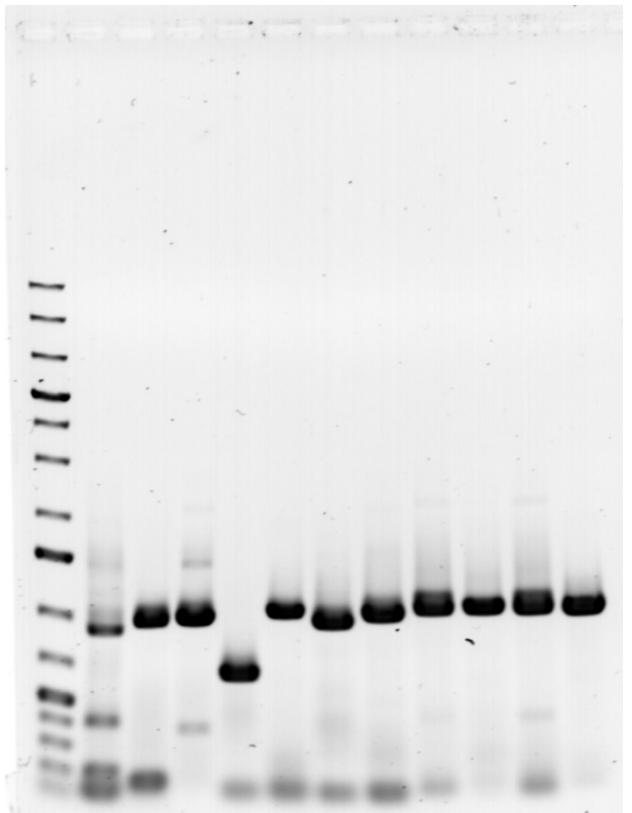
- 1) Prepare genomic DNA
- 2) Digest sufficient amount of DNA to completion with restriction enzyme
- 3) Run gel to separate DNA fragments according to size
- 4) Transfer, fix DNA to a membrane
  
- 5) Prepare probe DNA
- 6) Wash membrane with probe DNA
- 7) Visualize probe on membrane (appearing as bands where probe binds)

# Southern blot procedure



# DNA transfer: making a print from a gel

Run the gel



Make a print of the  
gel on a sheet



Detect DNA  
sequence



# **Probe to detect sequence of interest: *base-pairing* (hybridization)**

- Probe DNA
  - synthetic oligonucleotide
  - or
  - cloned gene (single stranded)
- The probe has to be easy to detect
  - Radioactivity
  - Fluorescence
  - Enzyme dependent color change
  - Enzyme dependent luminescence

# Hybridize (base pair) probes to target DNA

- blocking agents (e.g. milk, SDS) prevent non-specific interactions between probes and membrane
- Volume exclusion agents (eg. dextran sulfate) increase rate and level of hybridization
- Wash blot with increasing stringency...
  - Low stringency: high salt, low temperature, probe base pairs with sequences with mismatches
  - High stringency: low salt, higher temp., probe will base pair only to fully complementary sequences

# How to make a nucleic acid probe: order online

(Example: IDT dna.com)

Paste DNA sequence of the oligonucleotide in online order form

Define a modification:

- Fluorescence
- Attachment chemistry, e.g. biotin, digoxigenin
- Modified bases
- Randomized bases

Place order (company synthesizes oligo by automated phosphoramidite chemistry, sends oligo to you) (

<https://www.sigmaldrich.com/technical-documents/articles/biology/dna-oligonucleotide-synthesis.html> )

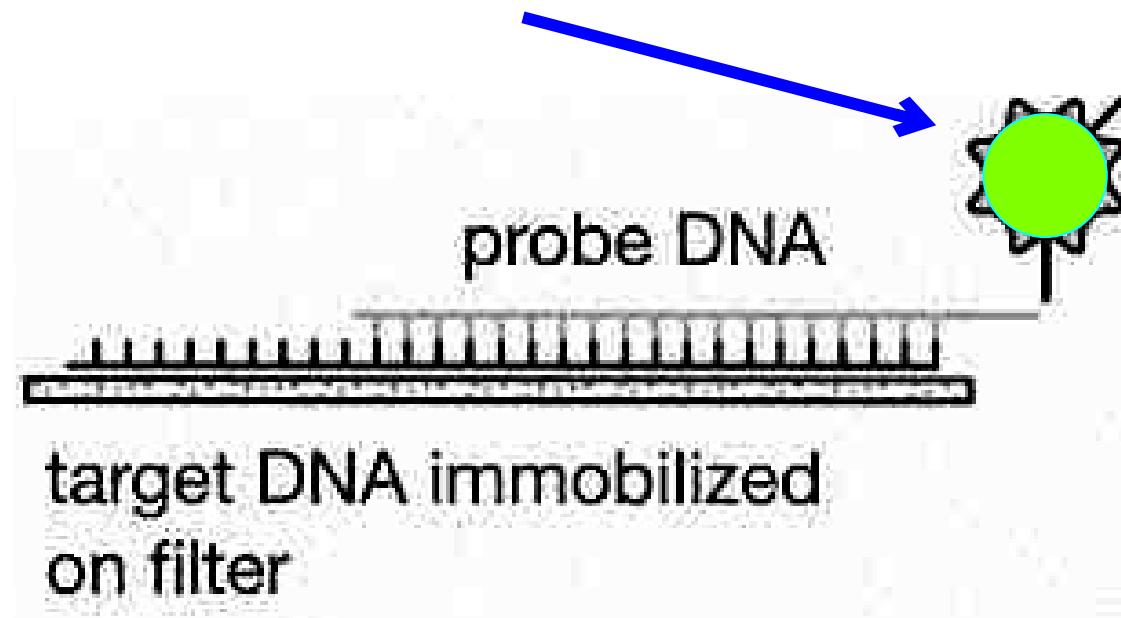
Do experiment

# Radioactive probes

- Example:  $^{32}\text{P}$  label
  - Add  $^{32}\text{P}$  ATP to the 5' end of probe DNA by kinase reaction
  - Probe DNA base pairs with target DNA
  - $^{32}\text{P}$  radioactive decay produces detectable signal
  - Detect radiolabel with
    - autoradiography: X ray film
    - phosphorimager: phosphor coated plates store the energy of the radioactive decay

# Fluorescence for detection

Fluorophores: Cy3, Cy5, etc.



Induction and detection of fluorescence (example: Cy3):  
excitation wavelength: 547 nm  
emission wavelength: 563 nm  
(<http://www.bdbiosciences.com/us/s/spectrumviewer> )

# How to amplify the DNA probe signal: add enzyme

Peroxidase, alkaline phosphatase enzyme activity leads to easily detected color change or emitted light

1.1) Covalently attach the **enzyme** to the DNA

Or

1.2) Attach a tag to the DNA:

- Digoxigenin (DIG)
- Biotin

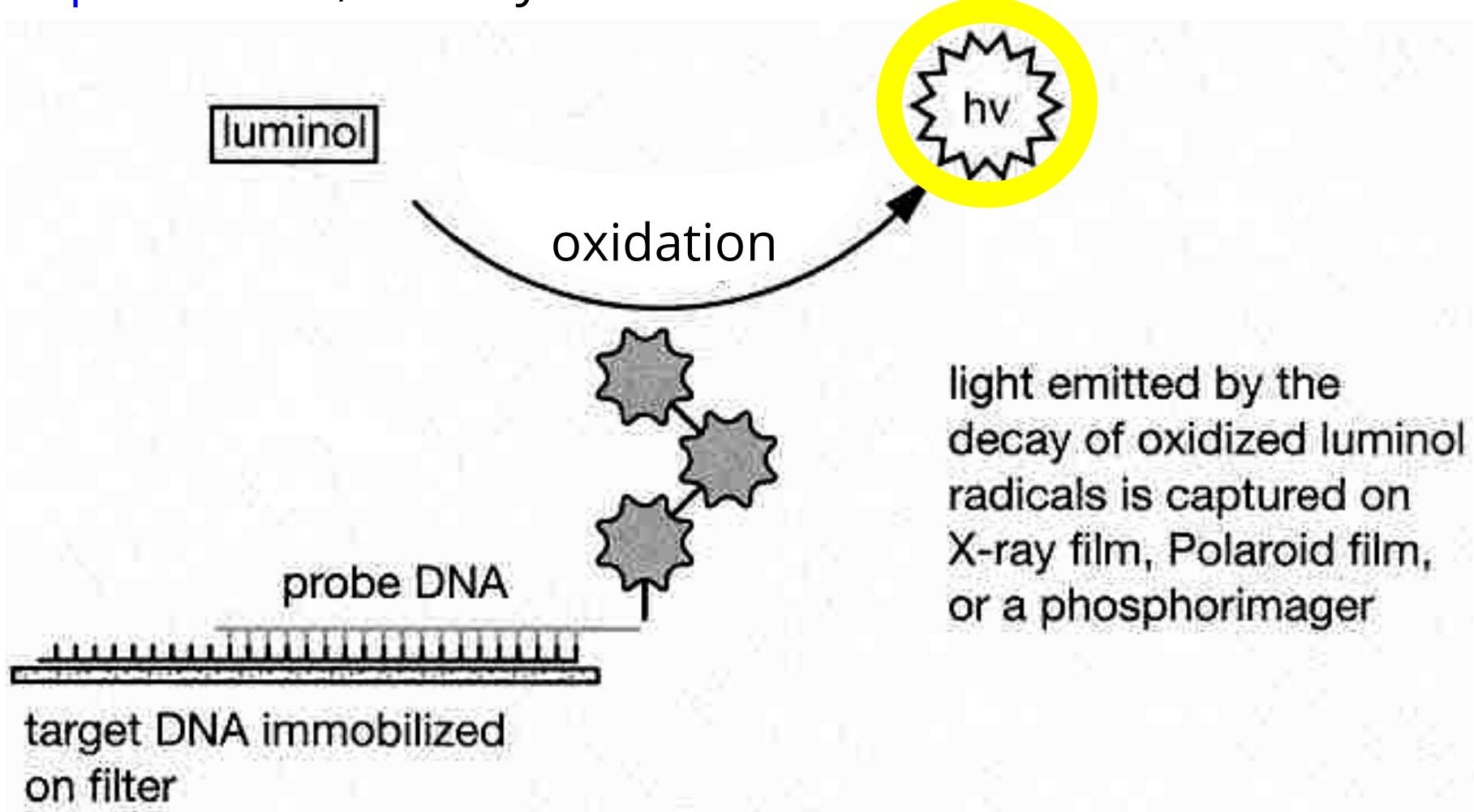
Then bring **enzyme** to the DNA tag:

- Conjugate it to antibody that recognizes DIG
- Conjugate it to streptavidin that binds to biotin

2) Detect **enzyme** through its activity

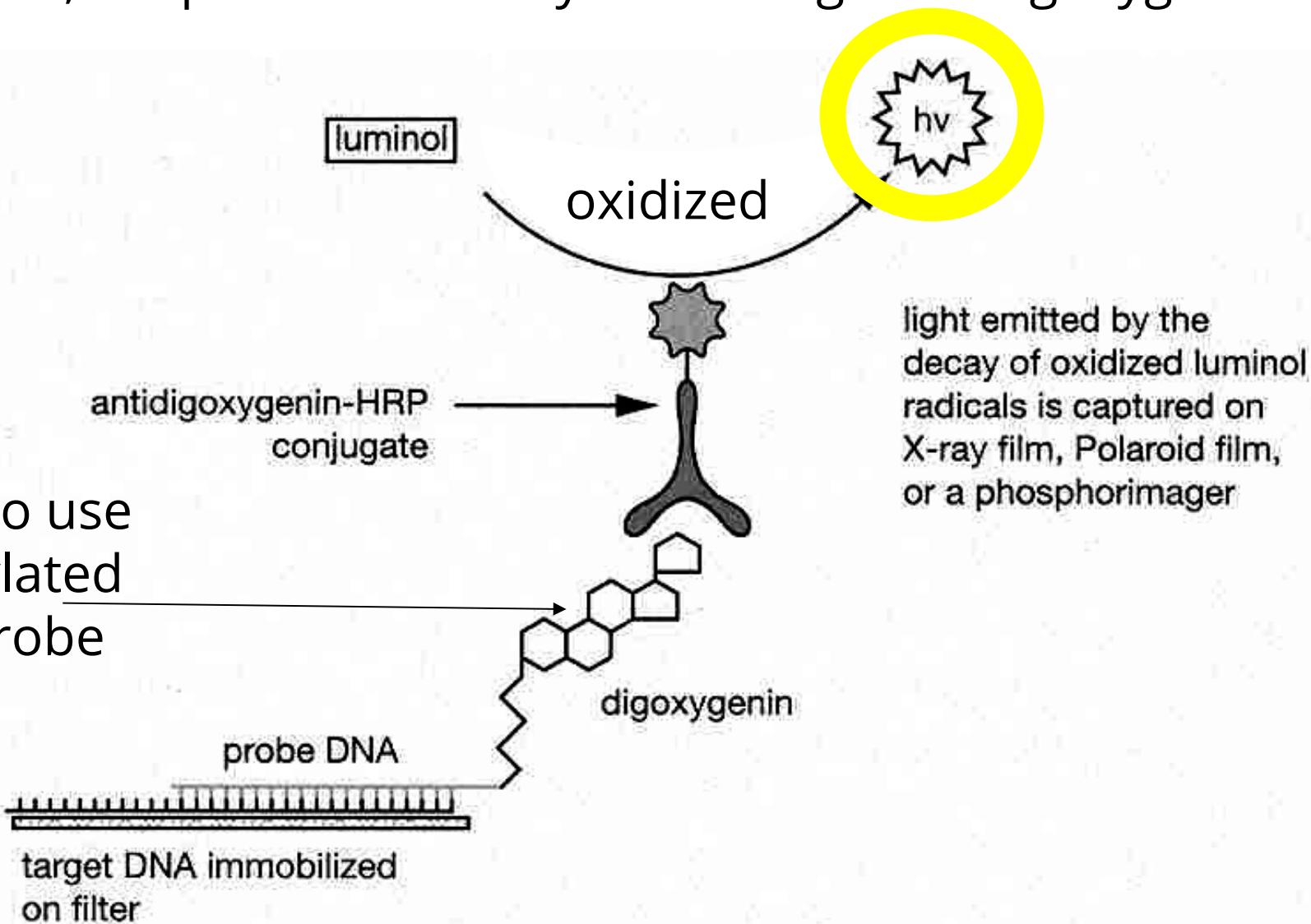
# Enzyme-linked probes: covalent linkage

peroxidase, directly attached



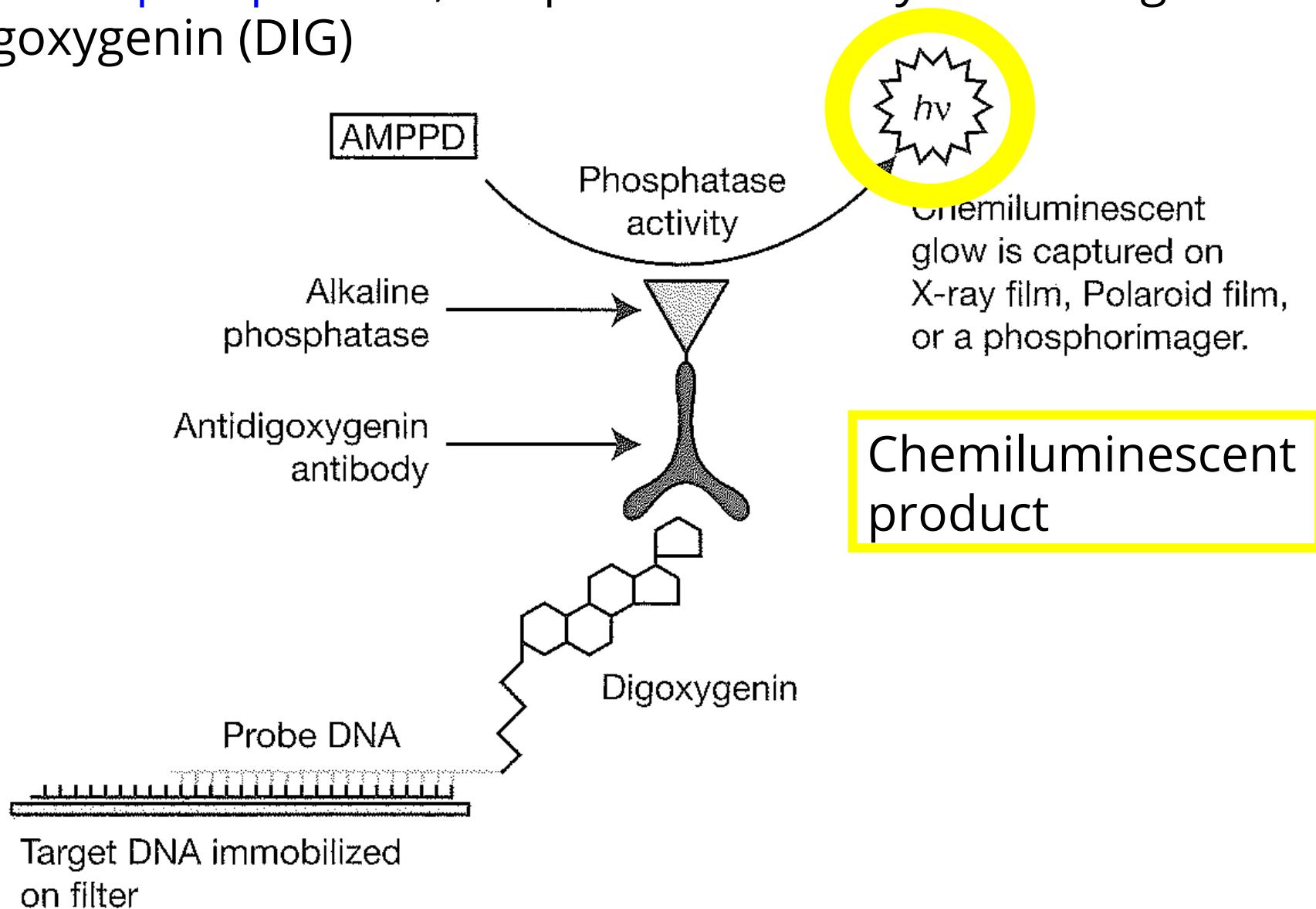
# Enzyme-linked probes: epitope recognition

peroxidase, coupled to antibody that recognizes digoxigenin (DIG)



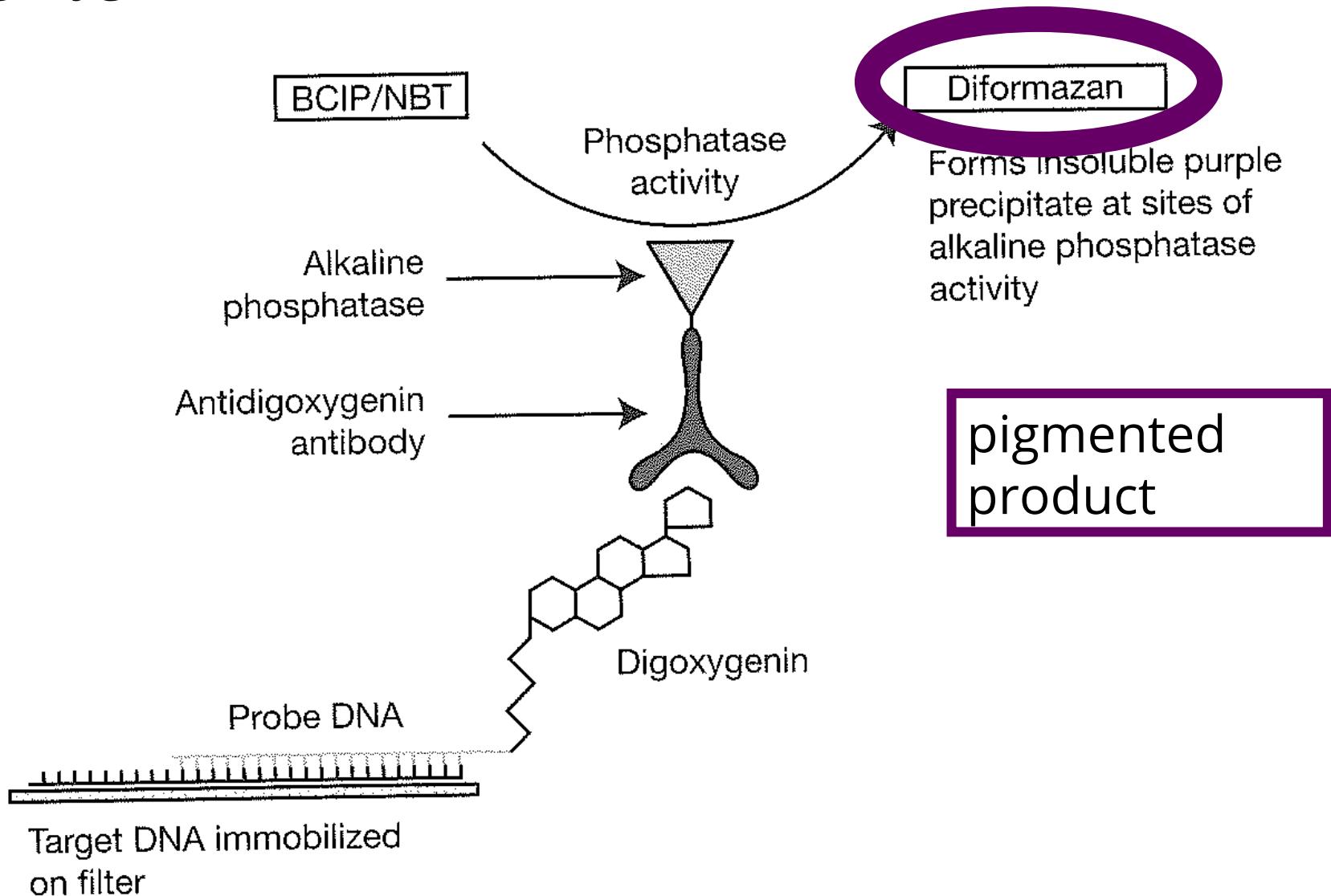
# Enzyme-linked probes: epitope recognition

alkaline phosphatase, coupled to antibody that recognizes digoxigenin (DIG)

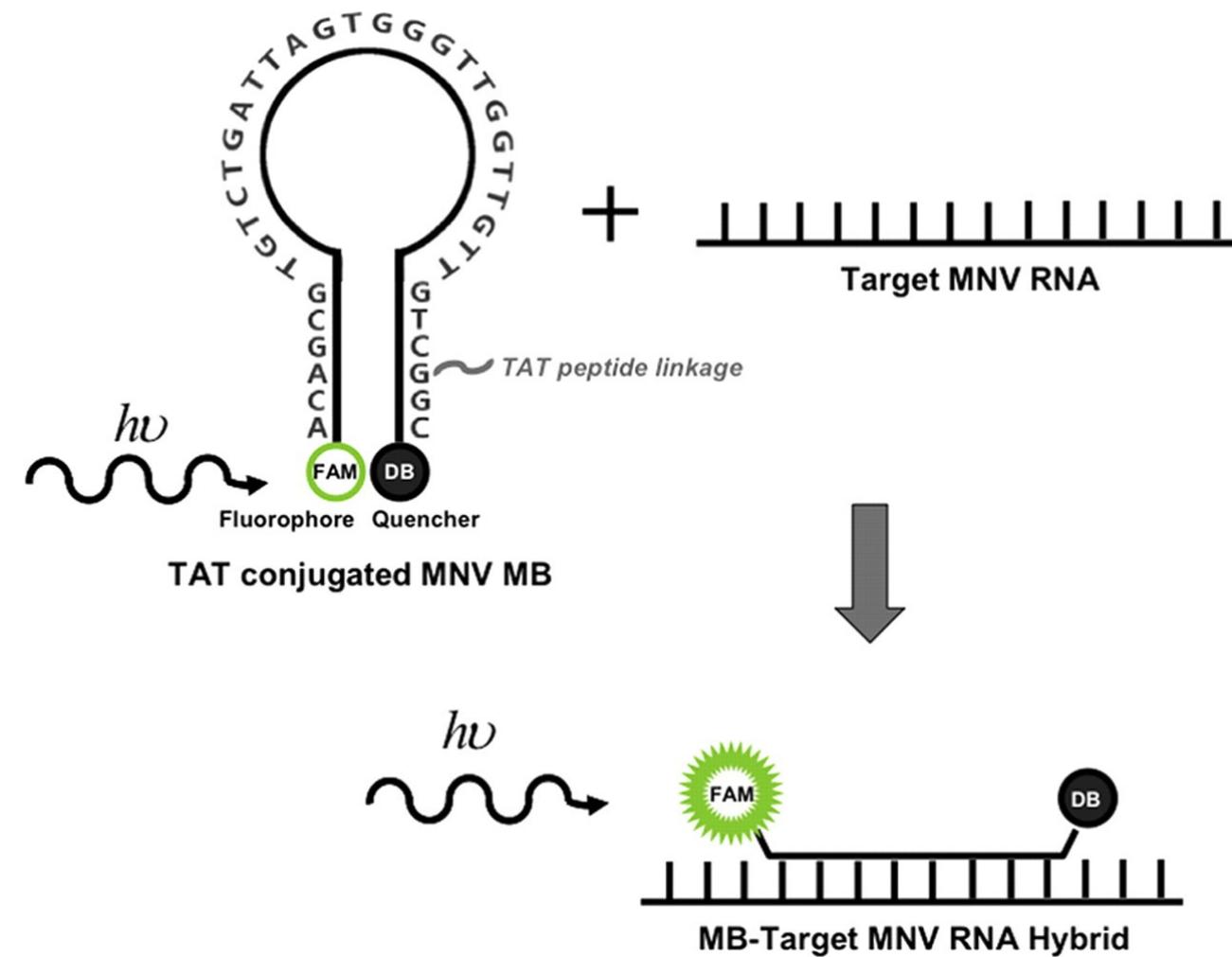


# Enzyme-linked probes: epitope recognition

alkaline phosphatase, coupled to antibody that recognizes digoxigenin (DIG)



# Detection of specific nucleic acids: in solution



The 'quencher' suppresses fluorescence in the hairpin structure



Target recognition disrupts hairpin, allows fluorescence

Probe DNA fluorescent only after target is detected

## Northern blots: RNA

Same basic technique as Southern blots, but **RNA** is run on the initial gel and is transferred to the membrane.

This method was used to measure levels of gene transcription *in vivo* (detecting changes in the levels of RNA transcript under differing conditions)

Microarrays for measuring mRNA abundance are based on this principle, but many probes are immobilized in a regular array -- reverse transcribed (and fluorescently labelled) RNA “lights up” the probes on the microarray

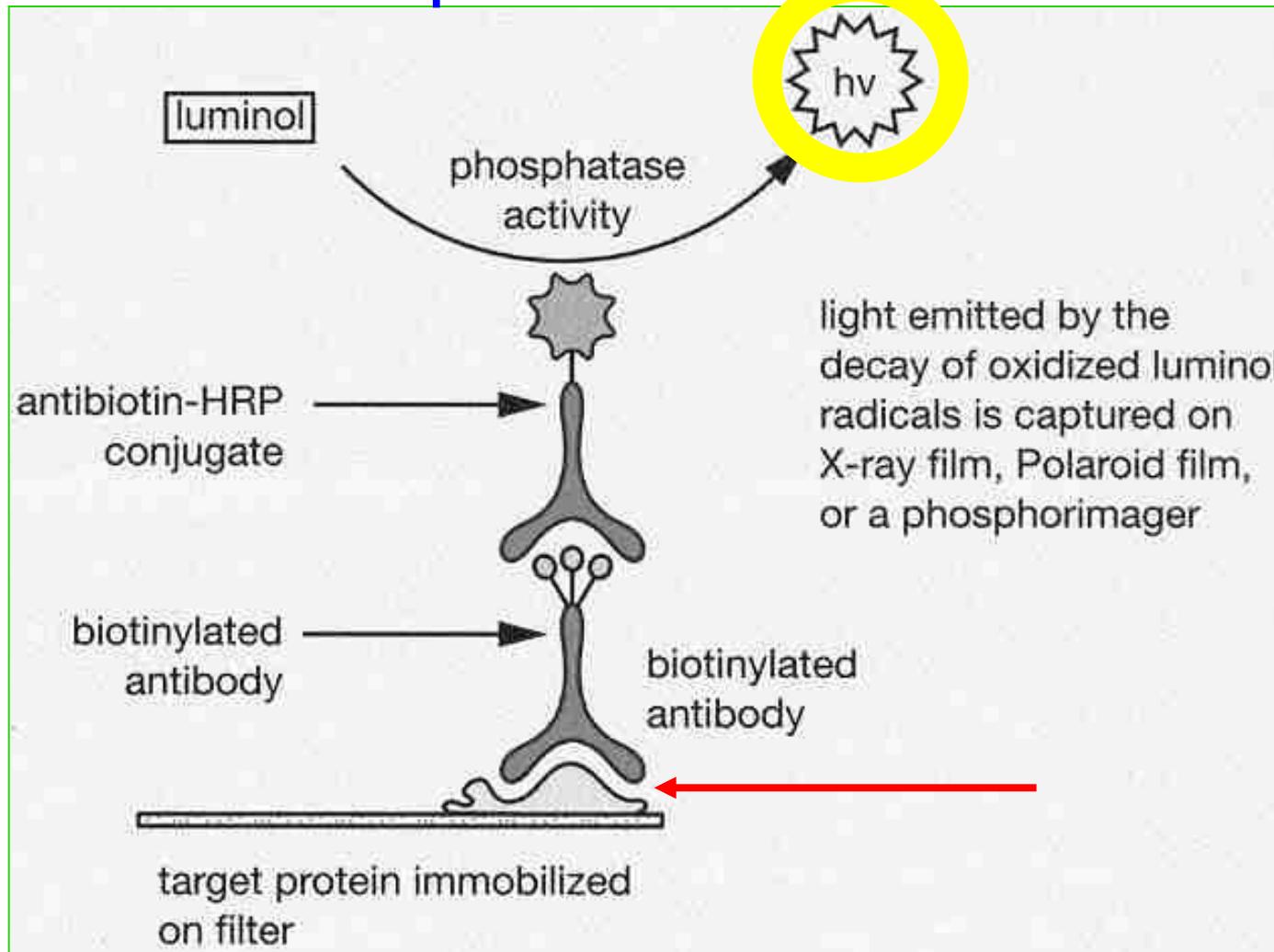
Current alternative for single genes is RT-PCR (reverse transcriptase to convert RNA to DNA, then PCR

# Protein detection in samples

Is a specific protein being made in a cell? How much is there?  
When is the protein present?

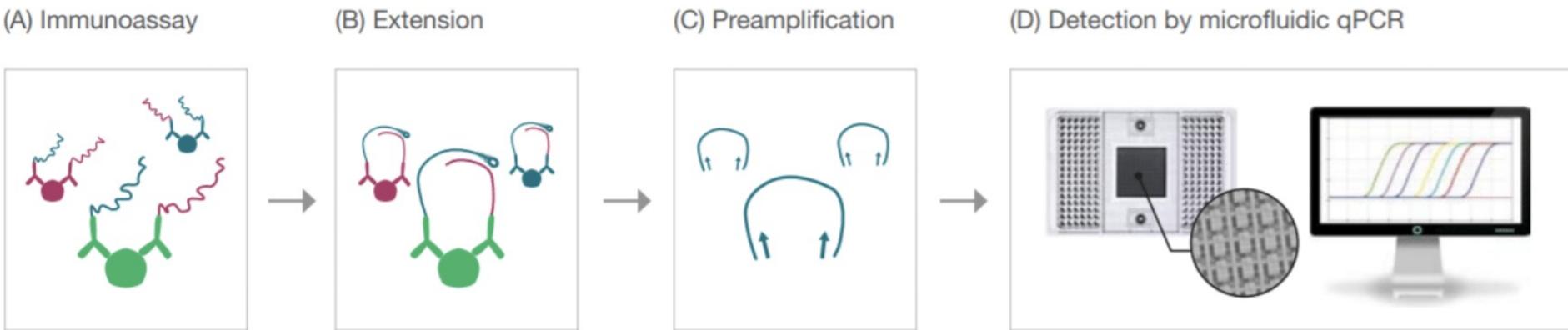
- Many proteins are made by cells. You need a way to detect a specific protein. Hybridization won't work!
  - Purify the protein
  - Raise antibodies to the protein (rabbits, goats, chickens, llamas)
  - Isolate the antibodies from animal blood
  - Test the antibodies for specificity
- Proteins separated by SDS PAGE transferred to membranes using the same principle as Southern blots
- Specific proteins detected by probing blot with antibodies to protein of interest

# Western blots: proteins



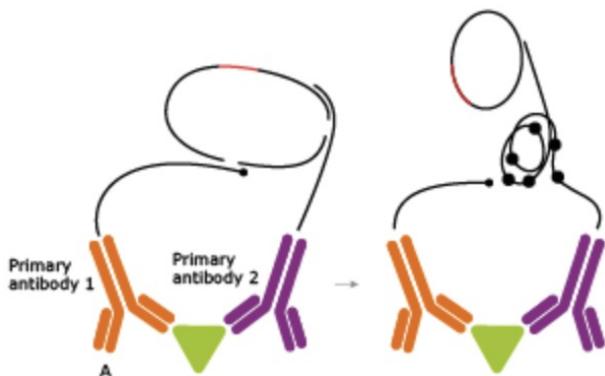
First antibody binding is detected by 'secondary' antibody that has enzyme (horseradish peroxidase, alkaline phosphatase) or radioactivity ( $^{125}\text{I}$ ) conjugated to it

Using antibodies to detect proteins in cells is not always straightforward – there can be ‘cross-reactivity’  
One solution is the “Proximity Extension Assay” to increase specificity



a) PCR detects primer extension product

b) Primer extension creates DNA hybridization sites



# Methods for detecting **specific** biomolecules

- 1) (If necessary, separate DNA, RNA, or proteins on the basis of size, by gel electrophoresis)
- 2) Immobilize (blot) the DNA, RNA, or protein
- 3) “Probe” the blot with something that will specifically interact with a target
  - a) DNA and RNA: interacts with a complementary nucleic acid
  - b) Protein: interacts with an antibody that specifically recognizes the protein

Types of blots:    Southern, DNA (named for E.M. Southern)

                  Northern, RNA

                  Western, protein

# Proteins (& nucleic acids) for manipulating DNA

- 1) Enzymes and other proteins require appropriate buffers and solution conditions for proper function
- 2) Specific tools and their uses
  - a) *Nucleic acid polymerases*: make and repair DNA
  - b) *Nucleases*: break down DNA or RNA
  - c) *Restriction endonuclease*: cut the DNA backbone at a specific site
  - d) *Ligase*: fix gaps in DNA backbone
  - e) The importance of hybridization/base pairing in putting DNA together/finding targets
  - f) *CRISPR-Cas9*: a genetic homing device that uses an RNA to find specific DNA targets

## Guide to readings:

- 1) 13 MC4 *Buffers and Reagents*. Tris, Good, and phosphate buffers, buffer recipes for various enzymes/protocols.
- 2) 14 MC4 *Enzymes*. Activity and uses for DNA polymerases, single subunit RNA polymerase, alkaline phosphatase
- 3) 14.5 MC4 Cut and paste. Restriction enzymes, ligases, and other information about this cloning technique
- 4) Cas9: The new frontier (2014)
- 5) Berg First rDNA (1972)

# Enzyme “ reaction buffers” : typical components

- **Buffer:** Tris or other buffer, maintain constant pH
- **Salt:** NaCl, KCl, PO<sub>4</sub><sup>-</sup>, etc. – maintains protein structure, and facilitates protein-DNA interactions
- **Divalent metal ions:** Mg<sup>2+</sup>, Ca<sup>2+</sup>, Zn<sup>2+</sup>, etc. – protein structure, enzyme activity
- **Glycerol:** (for storage) – stabilizes protein structure

# Enzyme “ reaction buffers” : typical components

- **EDTA**: chelates (removes) divalent cations – important especially for storage, if your enzyme is especially sensitive to metal ion-dependent proteases
  - **Beta mercaptoethanol or dithiothreitol**: reducing agents that prevent illegitimate disulfide bond formation
  - **Non-specific protein**: Bovine serum albumin (BSA)
  - **Other cofactors**, eg. ATP, NADH: some enzymes need these for function
- ✓ “ 10X” reaction buffer is ten times too concentrated. Make a 1/10 dilution for “ 1X” , the working concentration

## Enzyme structure/activity is pH sensitive: buffer is essential

Ideal biochemical buffers:

- $pK_a$  ( $\log_{10}$  of the acid dissociation constant) between 6 and 8  
(buffering capacity is greatest when  $pH = pK_a$ )
- Chemically unreactive
- Polar (soluble, not membrane permeable)
- Non-toxic
- Inexpensive
- Buffering minimally influenced by temperature or salt

## Tris: widely used but not perfect

- Tris: pKa is 8.0, so buffering is weak below pH 7.5 and above pH 9
- Tris is toxic to many types of mammalian cell cultures
- *Tris solution pH changes with temperature.* pH falls by 0.03 units for each degree C increase (pH 8.0 at 25°C becomes pH 6.5 at 75°C)
- *Tris solution pH changes with concentration*  
Example: 100mM Tris, pH 8.0 → dilute to 10mM Tris, pH is now 7.9

## Other buffers, e.g. Good's buffers

- N-substituted aminosulfonic acids: HEPES, Tricine, BES, MOPS, MES
- Useful at pH below 7.5

# Proteins for manipulating DNA

## Specific tools and their uses

- a) *Nucleic acid polymerases*: make and repair DNA
- b) *Nucleases*: break down DNA or RNA
- c) *Restriction endonuclease*: cut the DNA backbone at a specific site
- d) *Ligase*: fix gaps in DNA backbone
- e) *CRISPR-Cas9*: a genomic homing device

# DNA polymerases: making copies, adding labels, or fixing DNA

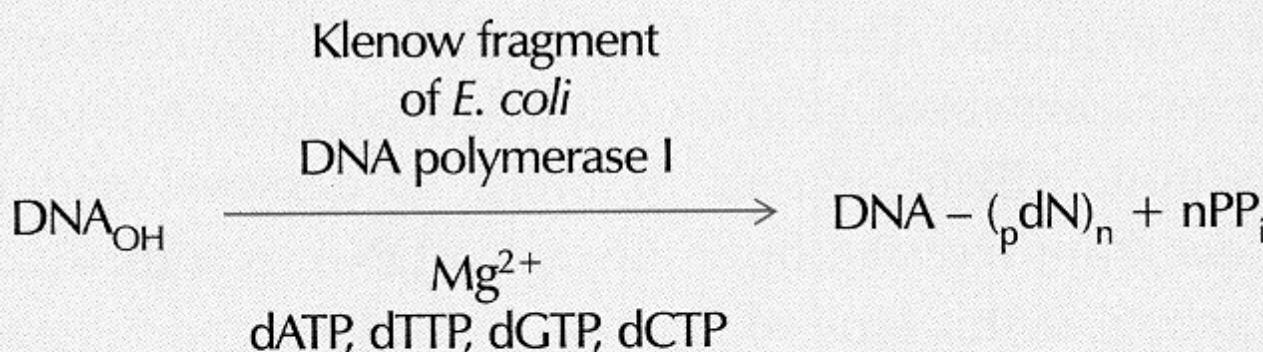
**Klenow fragment** of DNA polymerase – the C-terminal  
70% of E. coli DNA polymerase I

- Lacks a 5' → 3' exonuclease activity
- Uses include:
  - Synthesis of DNA from a 'primer'
  - Label DNA termini by filling in ends
  - Repair of ragged DNA ends
  - DNA sequencing

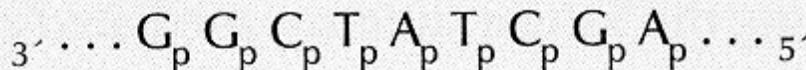
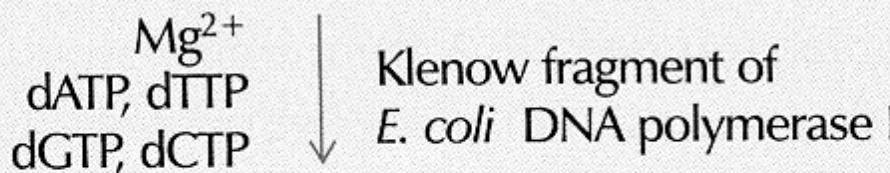
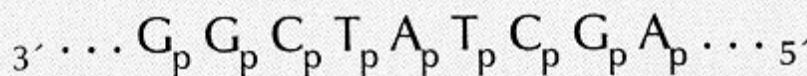
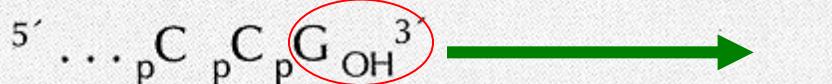
## Activity: 5' → 3' DNA polymerase

**Substrate:** Single-stranded DNA template with a primer containing a free 3'-hydroxyl group.

**Reaction:**



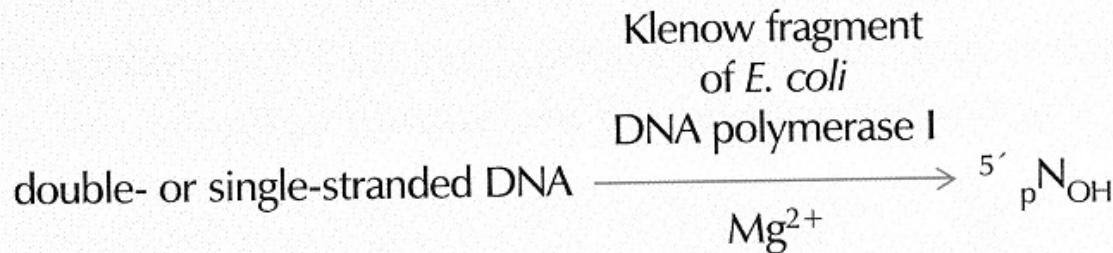
**For example:**



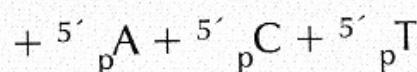
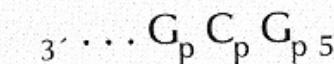
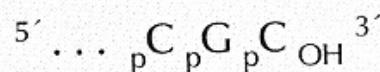
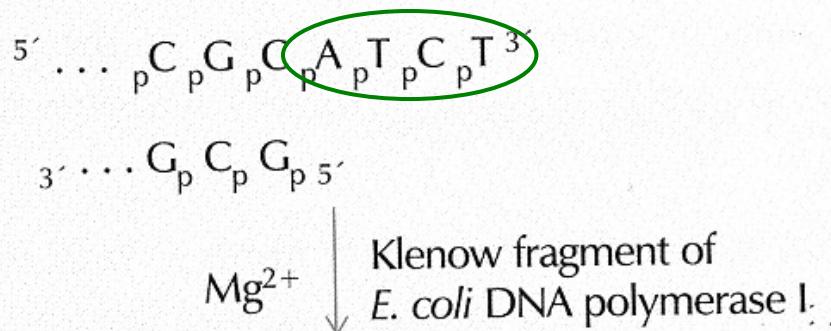
## Activity: 3' → 5' Exonuclease

**Substrate:** Double-stranded or single-stranded DNA degrades from free 3'-hydroxyl termini; exonuclease activity on double-stranded DNAs is blocked by 5' → 3' polymerase activity.

## Reaction:



## For example:



Make blunt-ended DNA (repair  
after mechanical fragmentation)

# DNA polymerases: for DNA sequencing

- **T7 DNA polymerase** (native) - highly processive, with highly active 3' → 5' exonuclease
- **T7 polymerase** (modified) --lack of both 3' → 5' exonuclease and 5' → 3' exonuclease
  - Ideal for sequencing, due to high processivity

# DNA polymerases for DNA amplification

## Thermostable DNA polymerases

- Taq: bacterial, high activity, higher mutation rate
- Archaeal DNA pols: lower activity, lower mutation rate
- PCR to amplify specific DNA sequences
- 'Cycle' sequencing: DNA sequencing with temperature cycles

# DNA polymerases for isothermal DNA amplification

## Phi29 DNA polymerase

- Highly processive
- Low mutation rate
- Strand displacement activity (no 5' to 3' exonuclease)
- Useful in “ WGA” (whole genome amplification)

# Special DNA polymerases

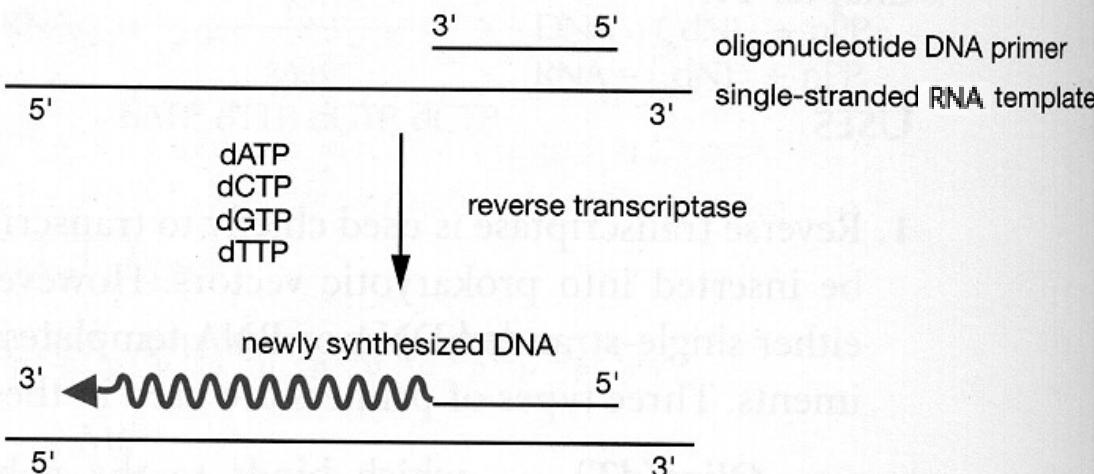
## **Reverse transcriptase: a retroviral protein**

- Makes DNA from an RNA template
- Used for making cDNA copies of RNA transcripts
- Detect, quantify RNA

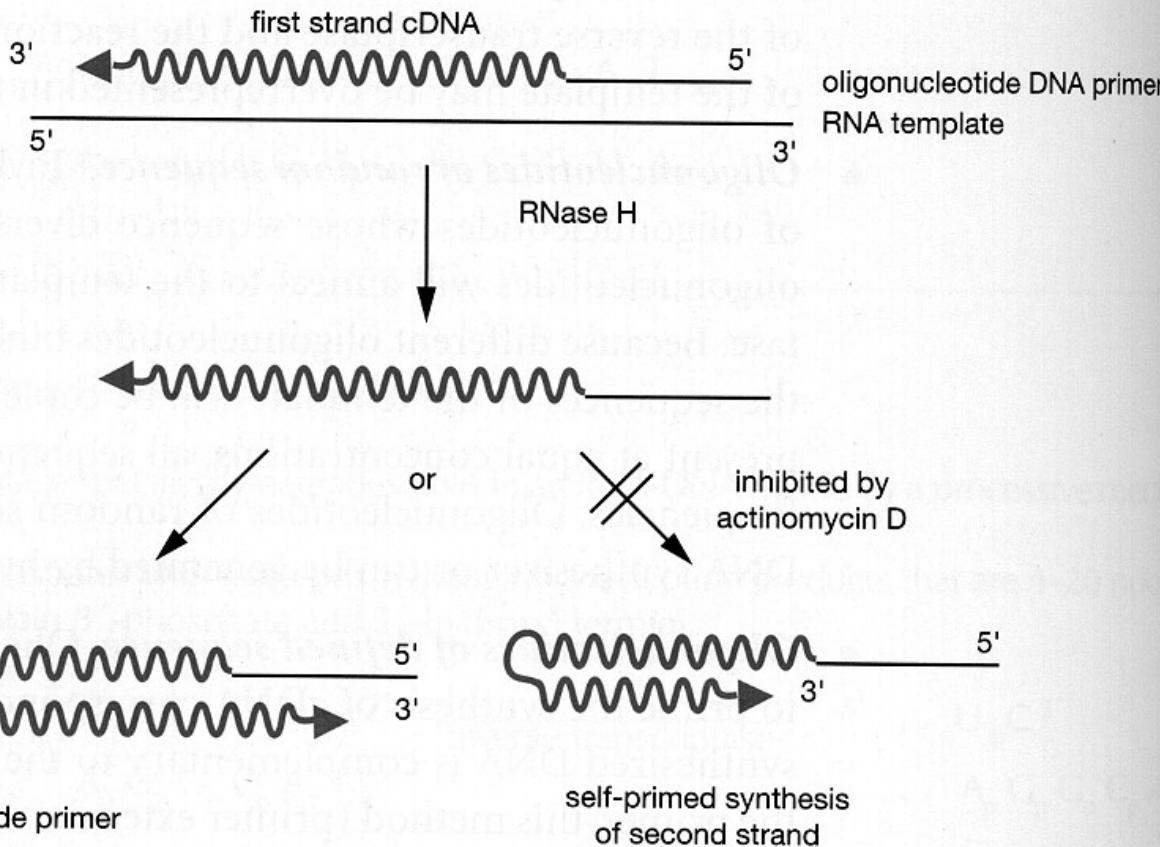
## **Reverse transcriptase:** has some issues

- The Km for dNTPs is very high (relatively non-processive, not good for long RNAs)
- Can make a DNA copy of RNA or DNA
- Can make a double stranded DNA by itself, but inefficiently
- To get clonable, double stranded DNA from RNA, the “second-strand” synthesis is usually done with DNA polymerase and a primer

## How RT works



B



# Special DNA polymerases

**Terminal transferase: makes new DNA without a template**

- template-independent DNA polymerase
- Incorporates dNTPs onto the 3' ends of DNA chains
- Used for adding homopolymer tails to the 3' ends of DNA strands (makes DNA fragments more easily clonable)

# RNA polymerase: T7

- Single-subunit RNA polymerase (from bacteriophage), no transcription factor required
- Highly specific promoter sequence determinants, and no cross-promoter recognition by cellular RNA polymerases
- Control transgene expression in a bacterial or eukaryotic host (place transgene under the control of a T7 RNAP promoter, and control expression of the T7 RNAP gene)
- Very active in vitro (makes lots of RNA easily)

# Proteins for manipulating DNA

- 1) Enzymes and other proteins require appropriate buffers and solution conditions for proper function
- 2) Specific tools and their uses
  - a) *Nucleic acid polymerases*: make and repair DNA
  - b) *Nucleases*: break down DNA or RNA
  - c) *Restriction endonuclease*: break DNA at a specific site
  - d) *Ligase*: seal breaks in DNA
  - e) *CRISPR-Cas9*: a genomic homing device

# Nucleases

- Exonucleases
  - Remove nucleotides one at a time from a DNA molecule
- Endonucleases
  - Break phosphodiester bonds within a DNA molecule
  - Include restriction enzymes

# Applications of exo- and endonucleases

Application	Recommended Enzyme(s)
Removal of 3' overhangs	T4 DNA Polymerase* + dNTPs
5' overhang treatment Fill in Cleavage	T4 DNA Polymerase* + dNTPs Klenow + dNTPs Mung Bean Nuclease
Removal of oligonucleotides post PCR	Exonuclease I
Removal of Chromosomal DNA in plasmid preparations	Lambda Exonuclease (Exonuclease I can be added to remove ssDNA generated by Lambda Exonuclease)
Removal of DNA in RNA preparations	DNase I
Chromatin Immunoprecipitation (ChIP) analysis	Micrococcal Nuclease
Generating ssDNA from linear dsDNA If 5' → 3' polarity required If 3' → 5' polarity required Best general choice	Lambda Exonuclease Exonuclease III Lambda Exonuclease

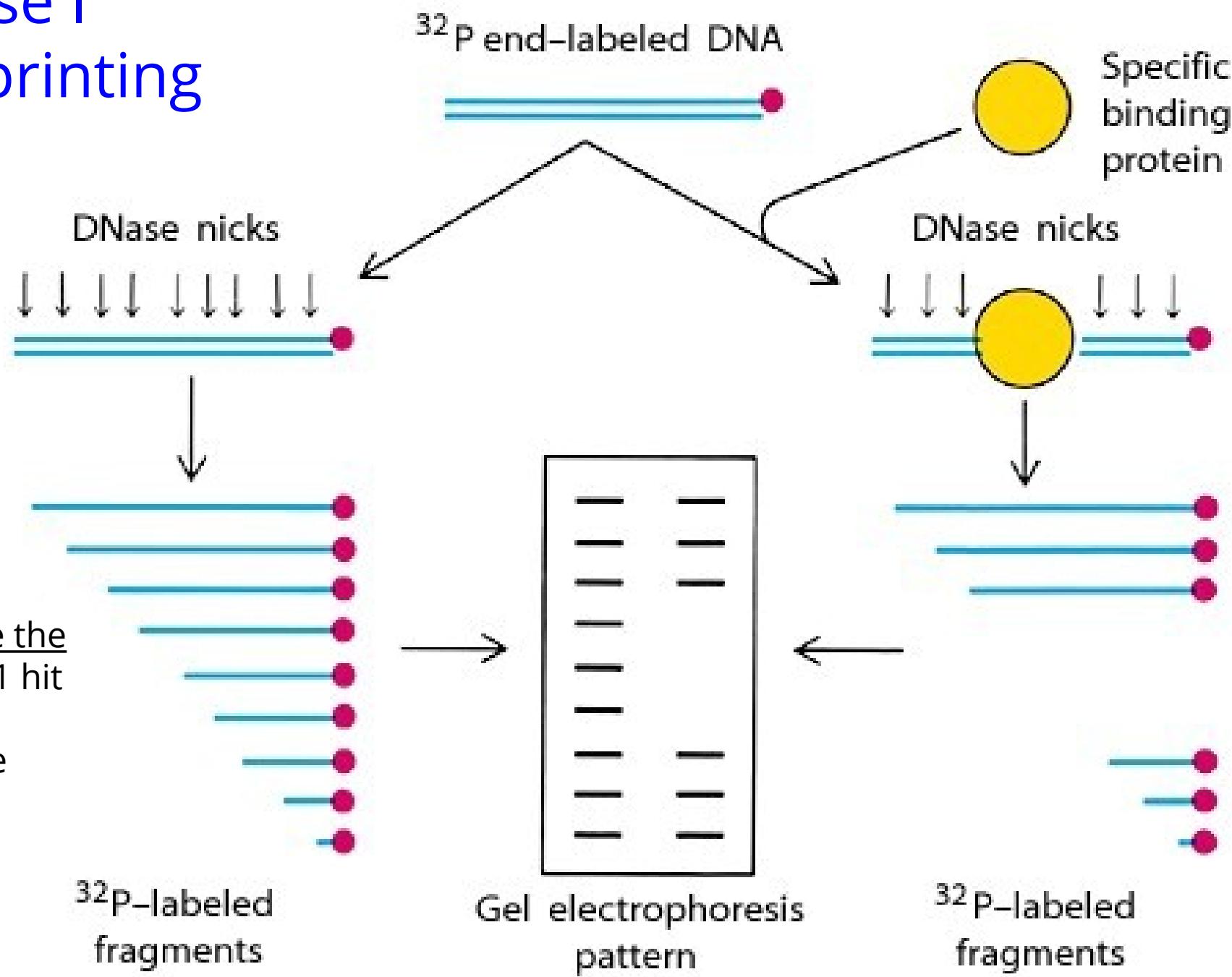
\* T4 DNA Polymerase has a strong exo- activity.

# Endonucleases

## Dnase I (deoxyribonuclease I)

- Cleaves double-stranded DNA randomly (also cleaves single-stranded DNA)
- Gets rid of double stranded DNA when only RNA or proteins are desired
- Reduces viscosity of cell lysates
- Useful in defining binding sites for DNA binding proteins

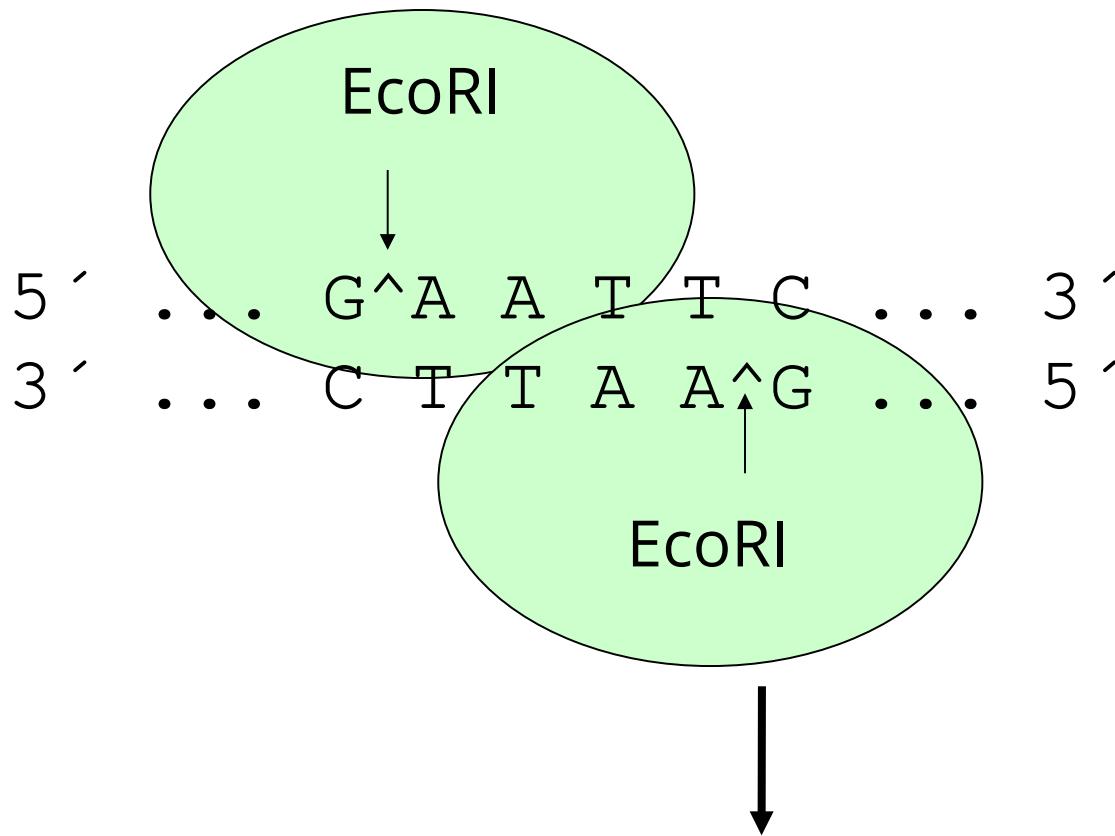
# DNAse I footprinting



# Type II endonucleases

- Target a specific, short DNA sequence
- Cut DNA at (or close to) that sequence
- DNA ends have 5' -phosphates, 3' -hydroxyls
- Useful for cloning purposes

# A type II restriction enzyme: EcoRI



5' ... G<sup>^</sup> 3'  
3' ... C T T A A 5'

5' A A T T C ... 3'  
3' ^G ... 5'

## Many type II enzymes, with unique target sequences

### 4-base recognition site:

AluI      5' ... AG<sup>^</sup>CT ... 3'      blunt ends

MspI      5' ... C<sup>^</sup>CGG ... 3'      5' overhang (2 bp)

### 6-bases

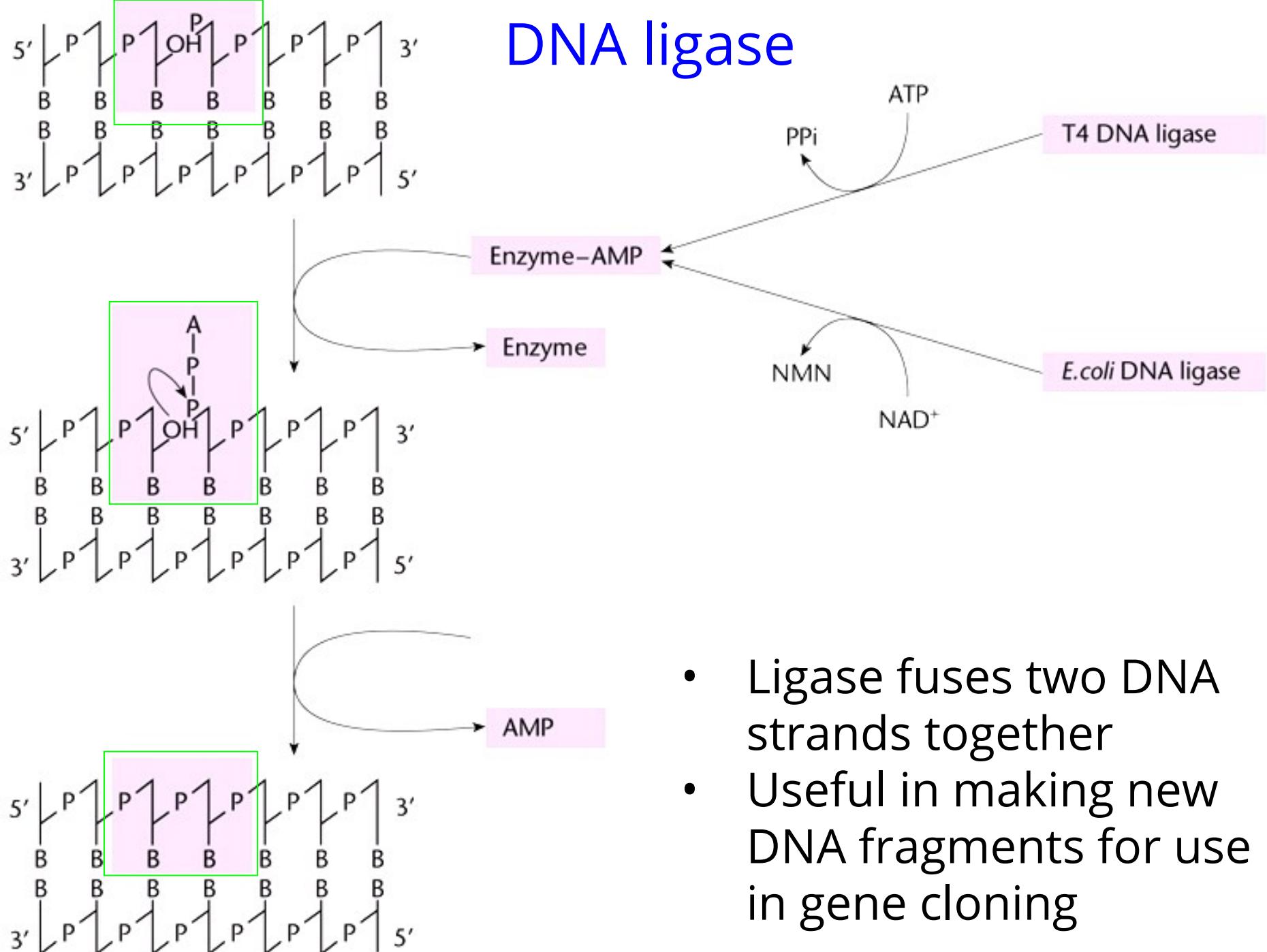
PvuII      5' ... CAG<sup>^</sup>CTG ... 3'      blunt ends

KpnI      5' ... GGTAC<sup>^</sup>C ... 3'      3' overhang (4 bp)

### 8-bases

NotI      5' ... GC<sup>^</sup>GGCCGC ... 3' 5' overhang (4 bp)

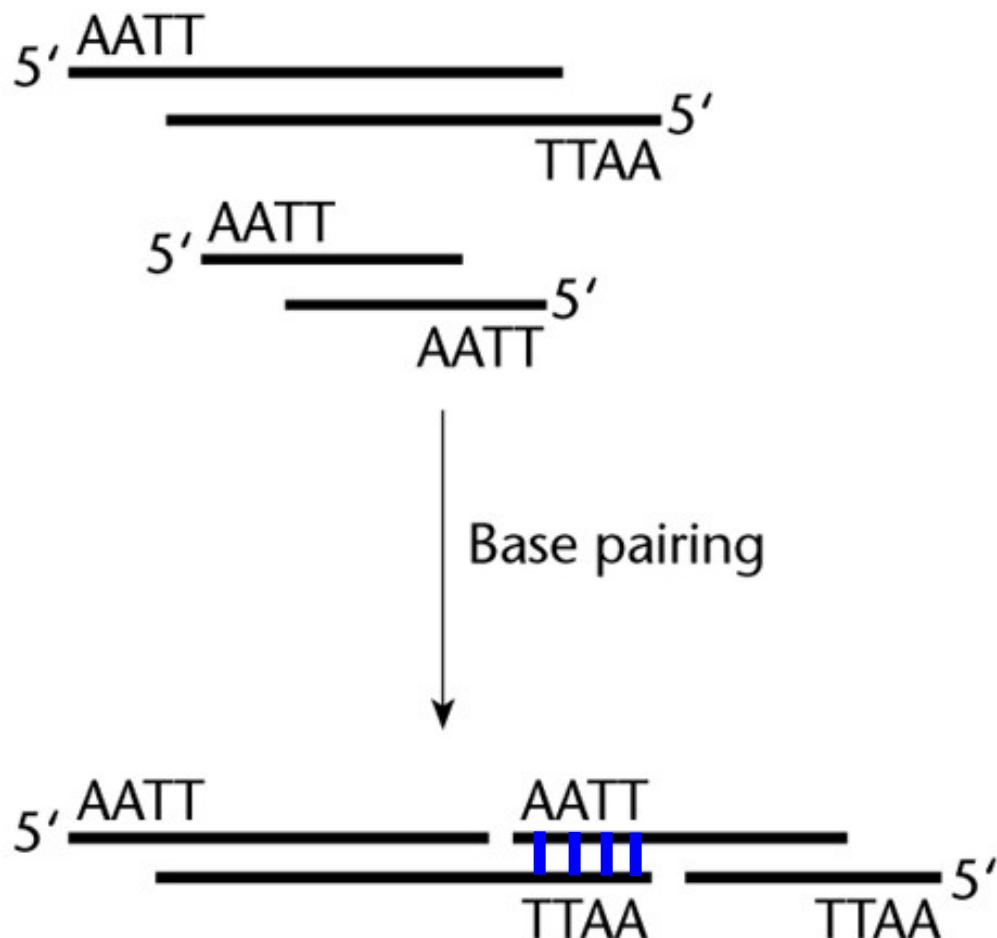
# DNA ligase



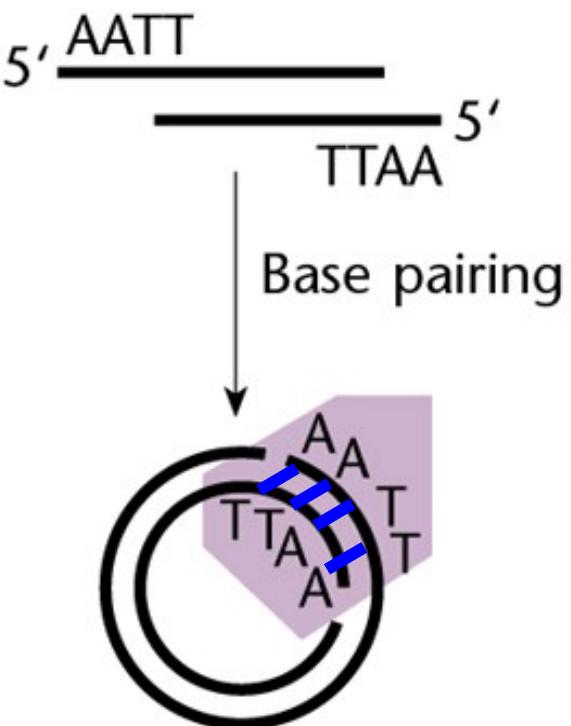
- Ligase fuses two DNA strands together
- Useful in making new DNA fragments for use in gene cloning

# Ligation of cohesive ends (overhangs)

## Intermolecular association



## Intramolecular association



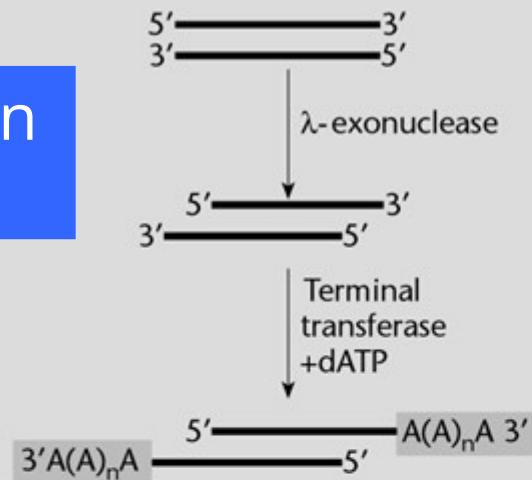
Base pairing helps in ligation reactions

# Cloning techniques

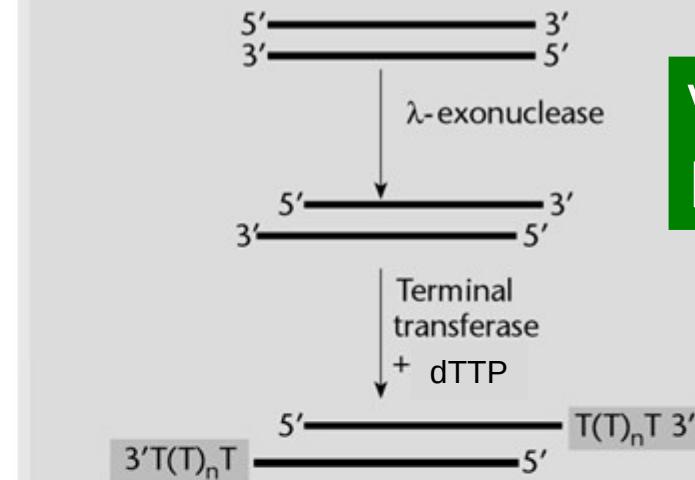
- A) The importance of the ends of the DNAs – make foreign DNA sequences more ligate-able
- B) Directional cloning – generate easily cloned PCR fragments
- C) Cloning by hybridization – new developments

# Terminal transferase: add polynucleotide tails to foreign DNA and vector DNA

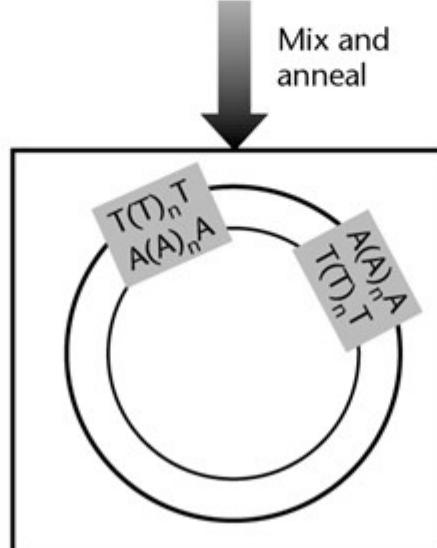
Foreign  
DNA



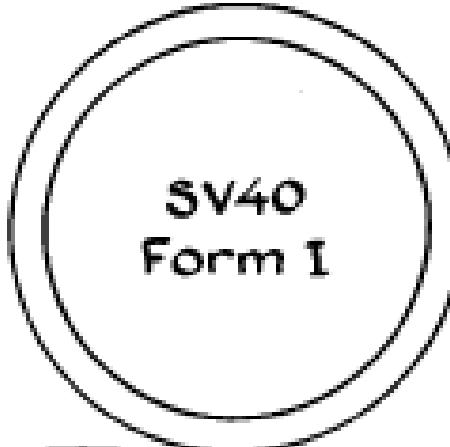
Vector  
DNA



Mix and  
anneal

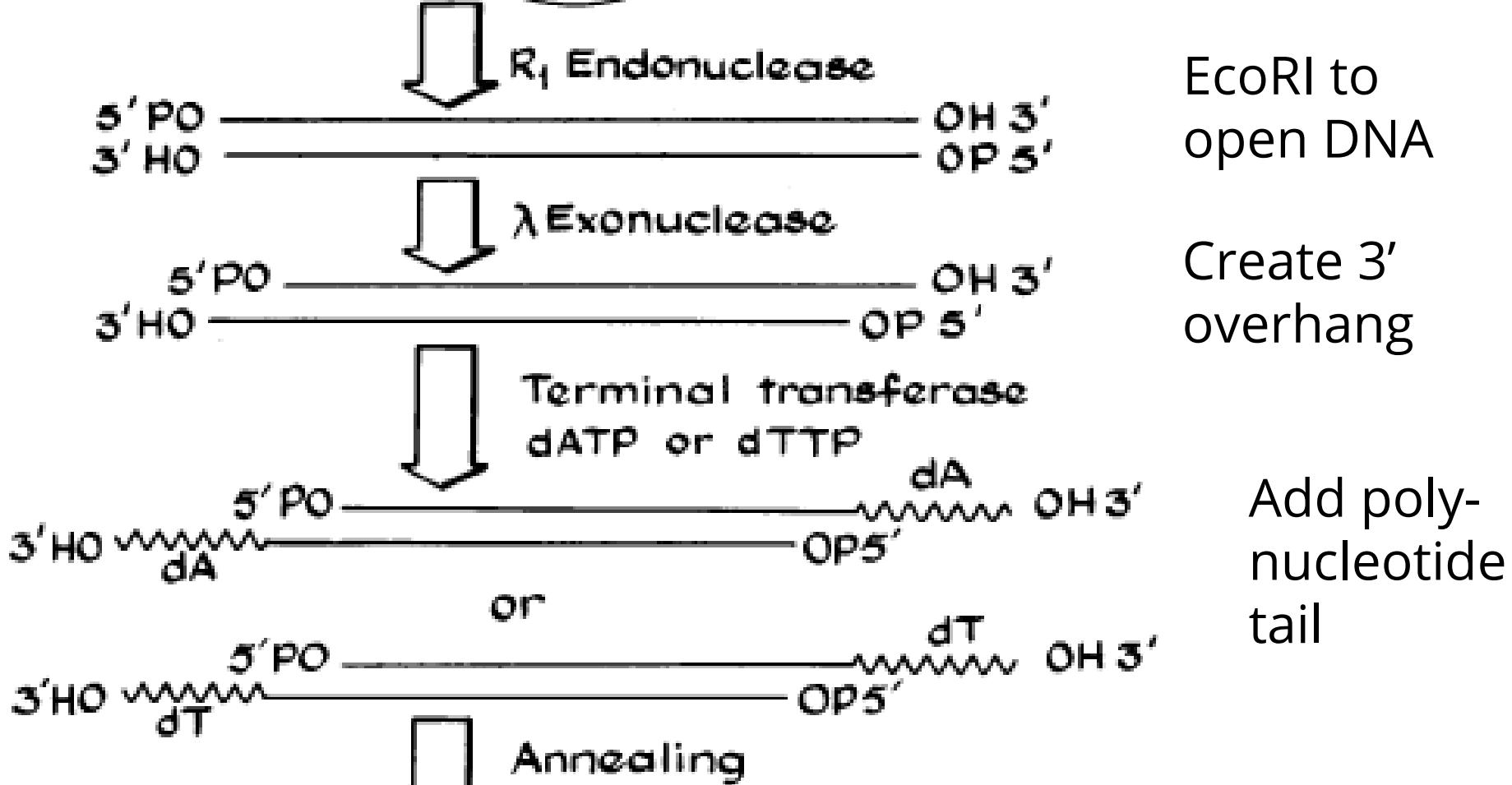


The first  
technique for  
joining two DNA  
molecules



Jackson, Symons  
and Berg, 1972

[https://youtu.be/u\\_10gnpxl](https://youtu.be/u_10gnpxl)



The first  
technique for  
joining two DNA  
molecules

Jackson, Symons  
and Berg, 1972

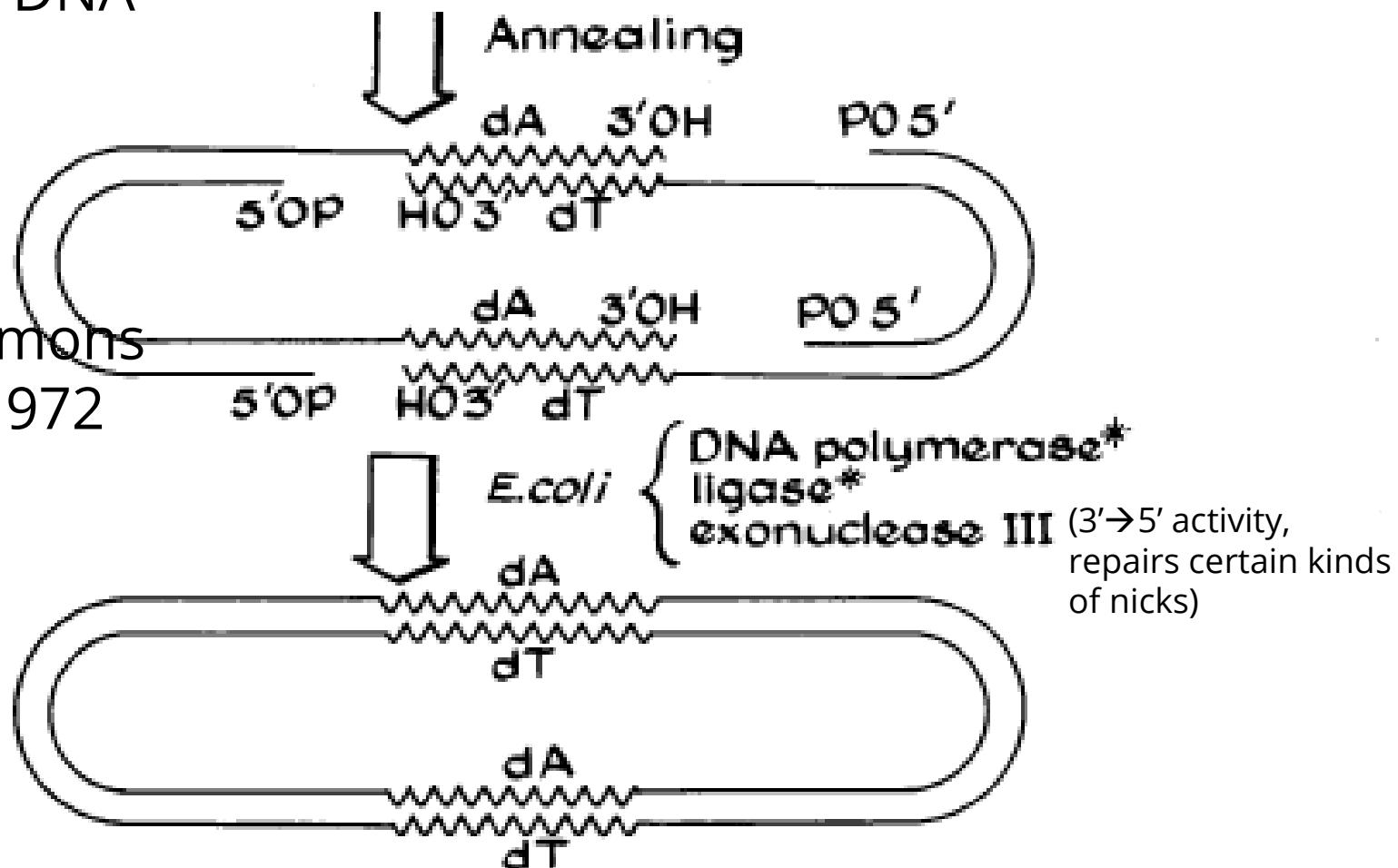
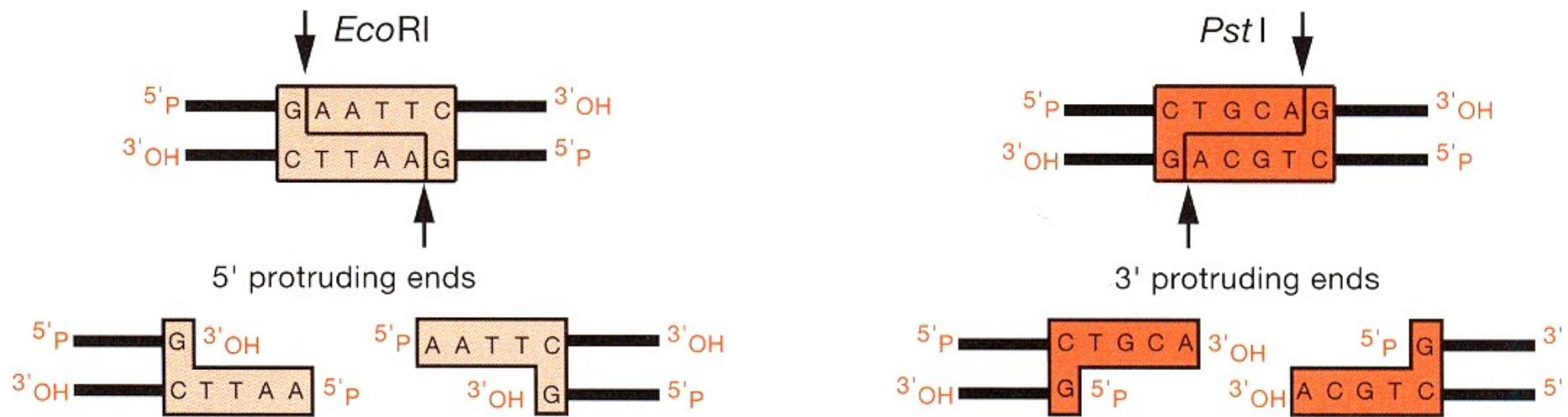


FIG. 1. General protocol for producing covalently closed SV40 dimer circles from SV40(I) DNA.

# Cloning techniques

- A) The importance of the ends of the DNAs – make foreign DNA sequences more ligate-able
- B) Directional cloning – generate easily cloned PCR fragments
- C) Cloning by hybridization – recent developments

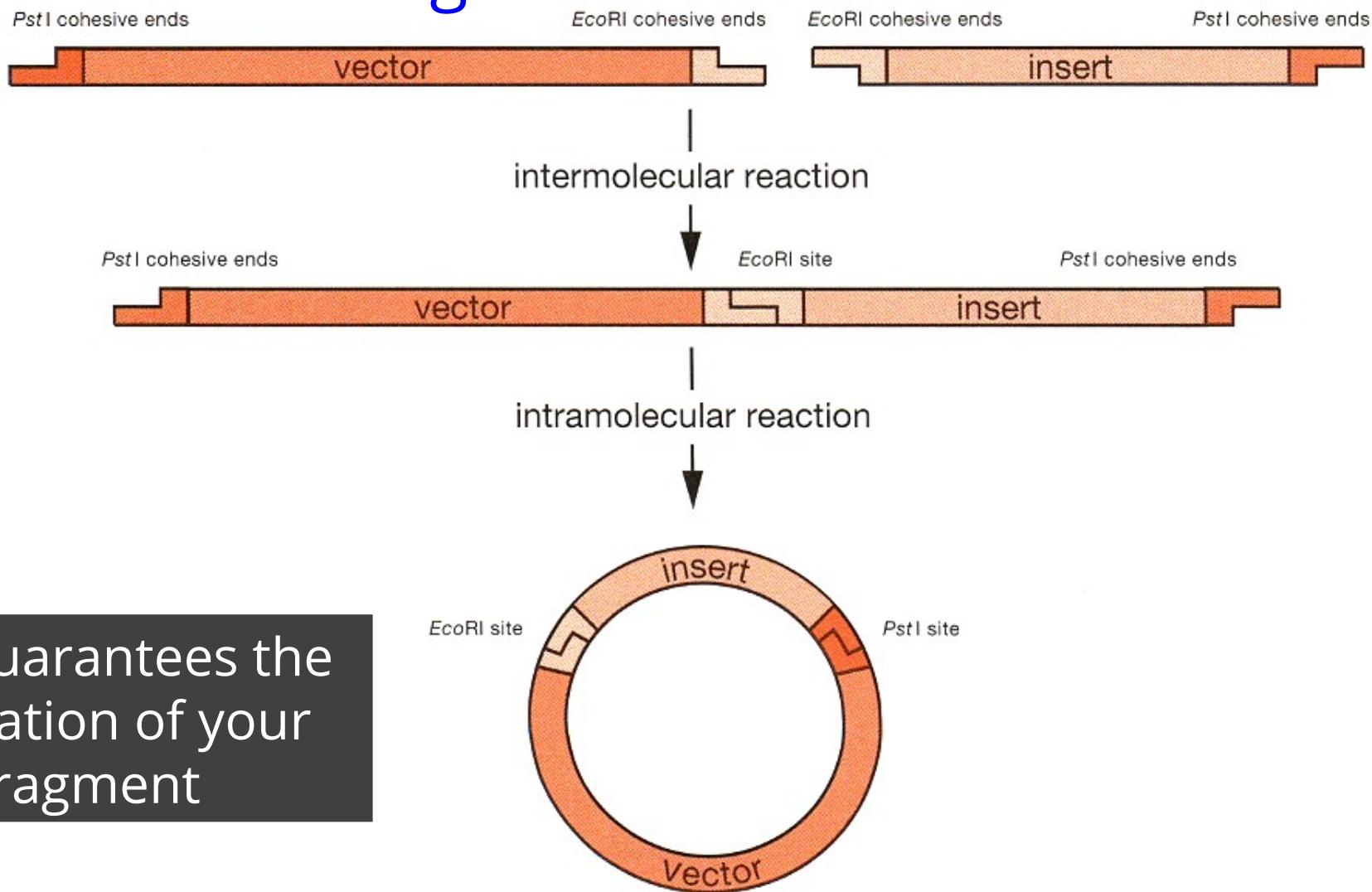
# Directional cloning



**FIGURE 1-2** Cloning 5' and 3' Protruding Ends

These sticky ends will not base pair with each other

# Directional cloning



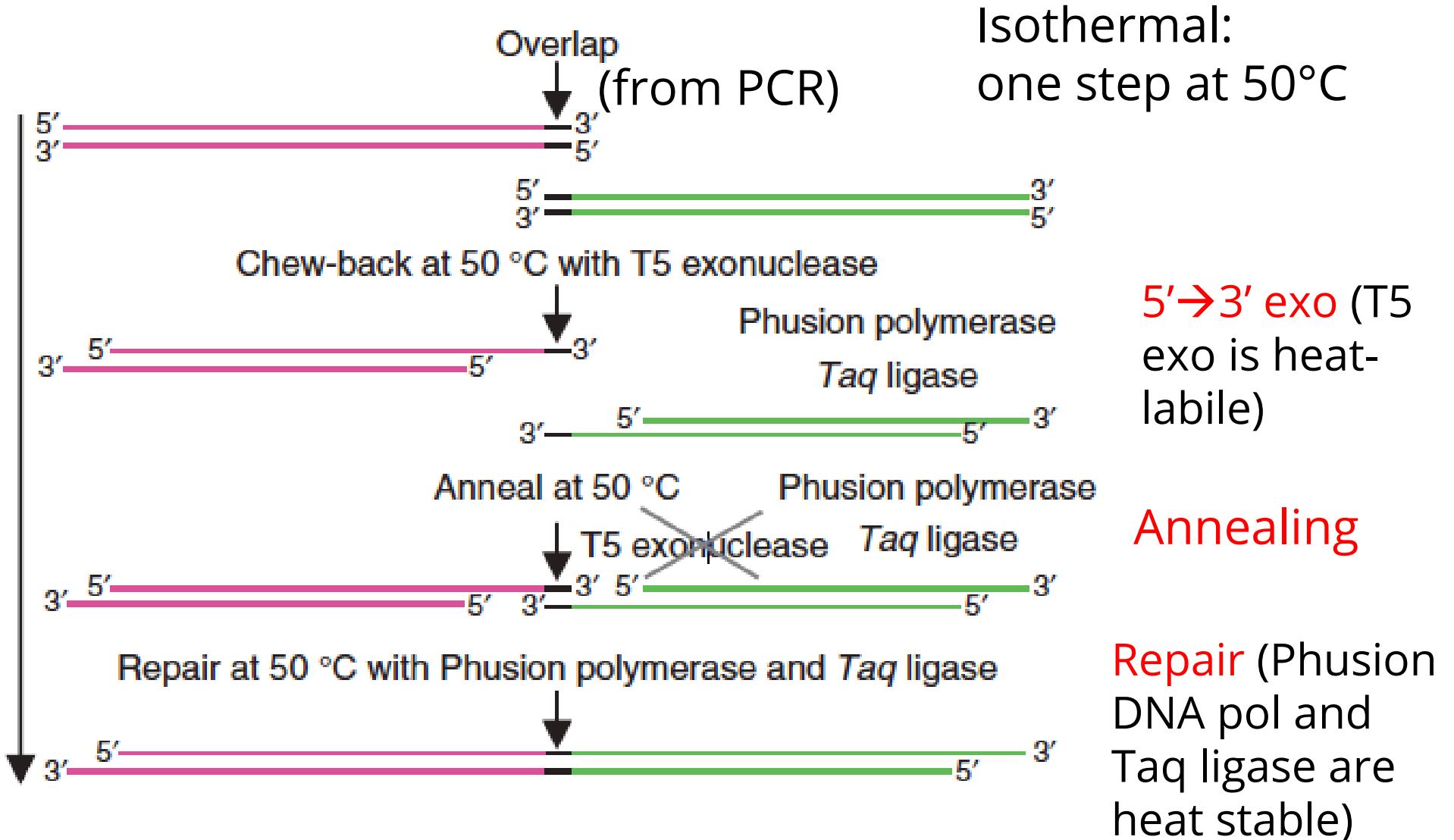
**FIGURE 1-5** Directional (Forced) Cloning in Plasmid Vectors

Vector sequences are represented by darker shading, and insert sequences by lighter shading.

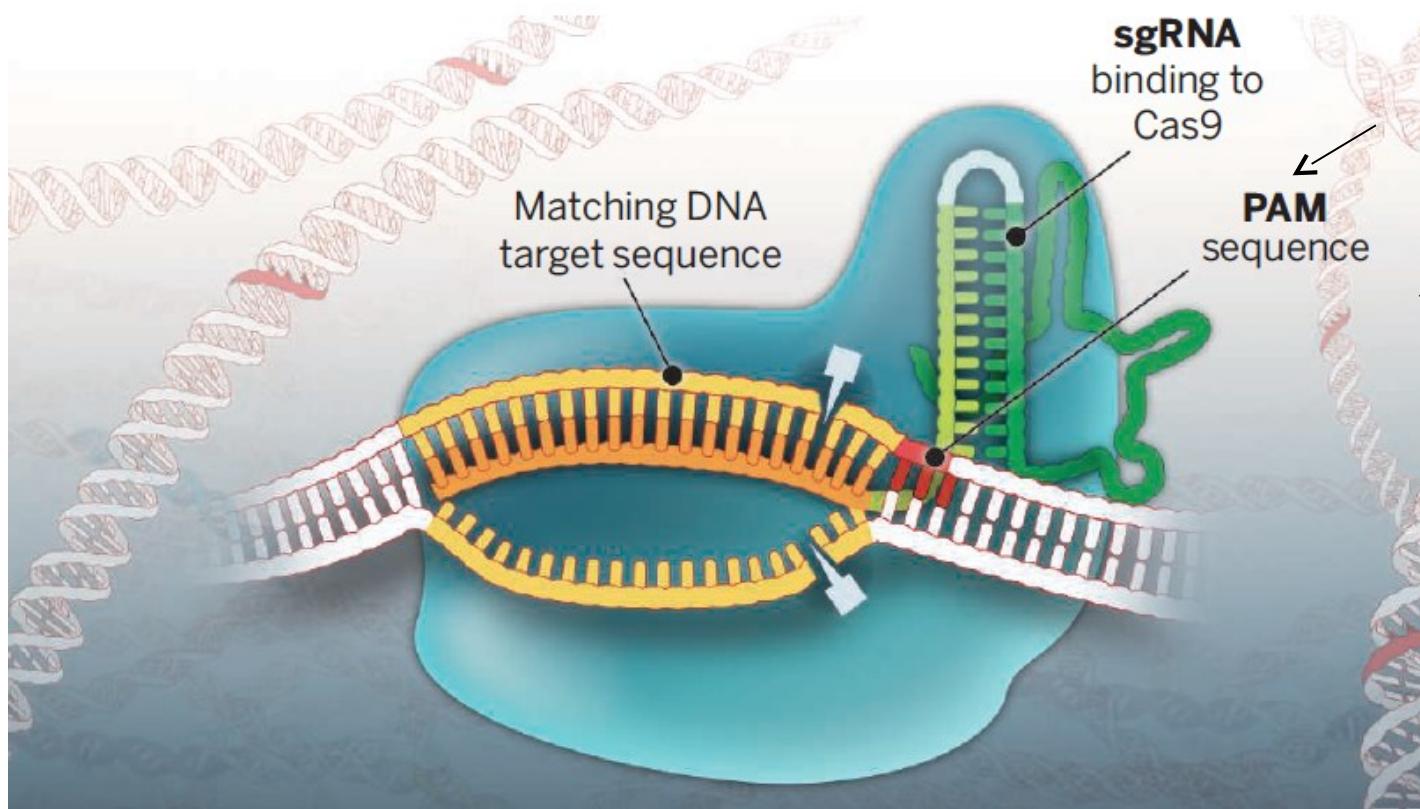
# Cloning by hybridization

- PCR product made using primers with 5' sequences matching plasmid cloning site
- DNA insert and plasmid are treated with nuclease to generate 5' or 3' overhangs, typically ~25 bases
- Base-pairing (hybridization) between plasmid and DNA insert sequence forces assembly
- Gaps in DNA backbone are corrected by DNA polymerase and DNA ligase (Gibson assembly)

# Gibson DNA assembly: make synthetic genes, pathways, or entire genomes.



## Type II CRISPR-Cas9: an RNA-guided nuclease



PAM = protospacer adjacent motif:  
required for Cas9 to bind DNA so strand invasion can occur by the guide RNA

The RNA-target interactions are very stable, and can also provide a tethering platform for proteins or RNAs, provided the nuclease activity is shut down

# 2020 Nobel Prize in Chemistry

Emmanuelle Charpentier, Max Planck Institute, Berlin  
Jennifer Doudna, University of California-Berkeley

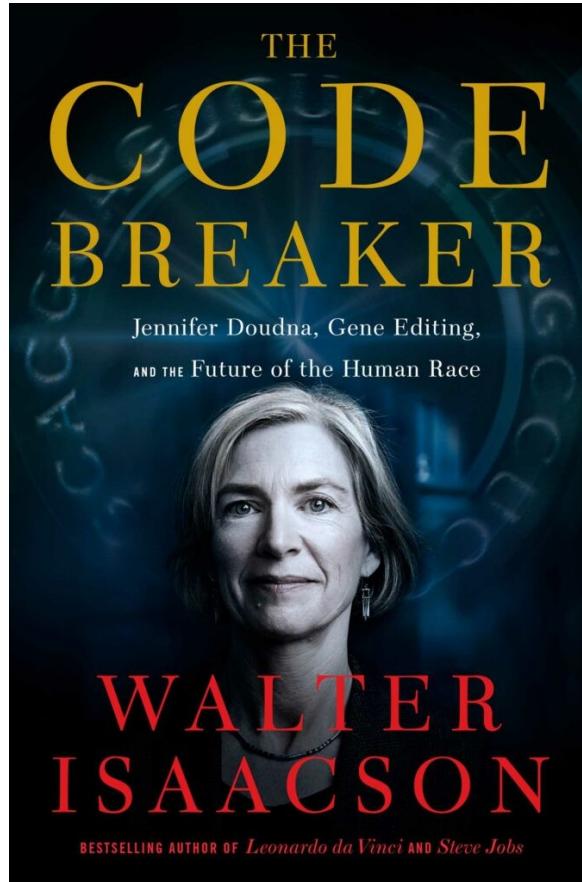


<https://www.nature.com/articles/d41586-020-02765-9>

The Nobel Prize in Chemistry 2020 was awarded jointly to Emmanuelle Charpentier and Jennifer A. Doudna "for the development of a method for genome editing."

<https://www.nobelprize.org/prizes/chemistry/2020/summary/>

Also see: "Code Breaker", book by Walter Isaacson (2021)



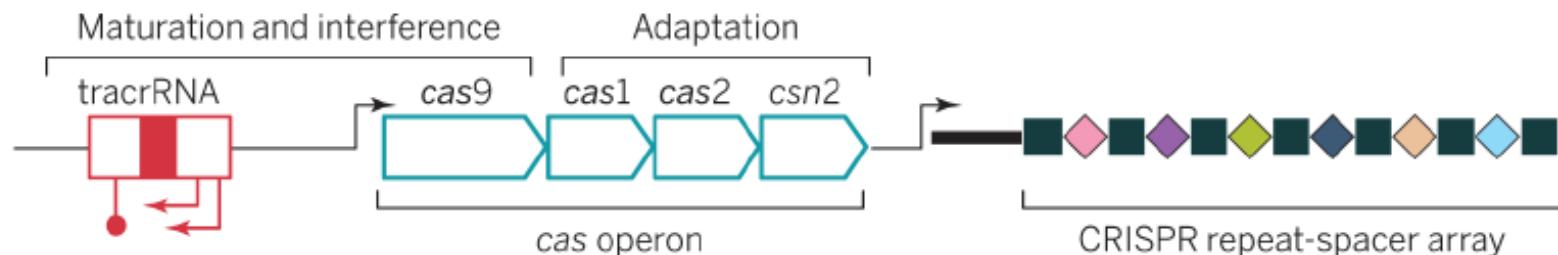
<https://datebook.sfchronicle.com/books/review-in-code-breaker-how-jennifer-doudna-became-a-pioneer-of-genes>

<https://www.wired.com/story/the-code-breaker-is-the-crispr-chronicle-you-need-to-read/>

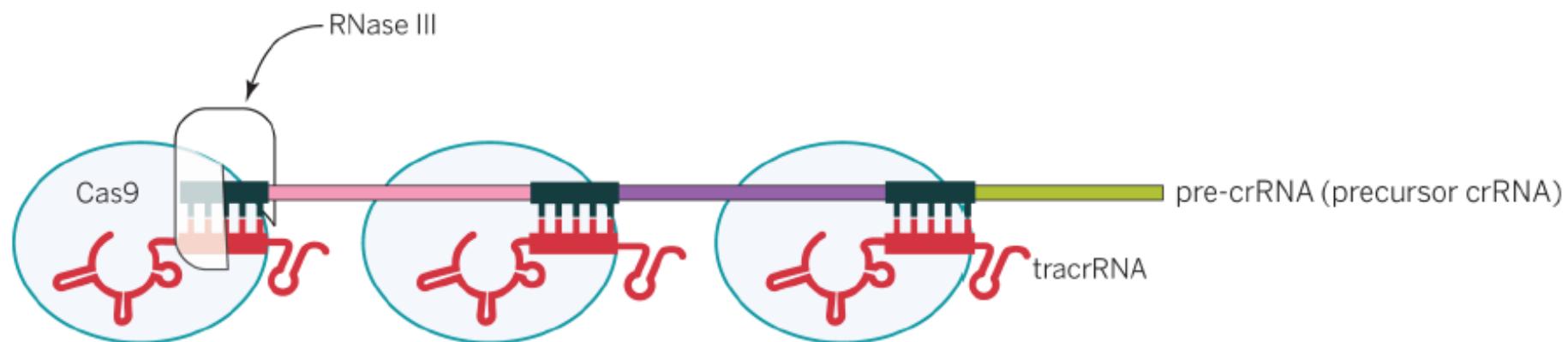
# CRISPR-Cas9 comes from prokaryotes

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

## A Genomic CRISPR locus



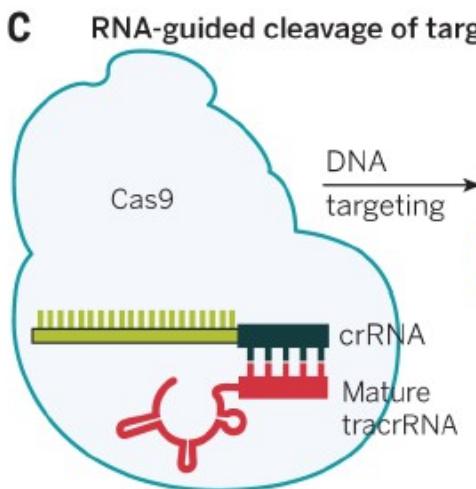
## B tracrRNA:crRNA co-maturation and Cas9 co-complex formation



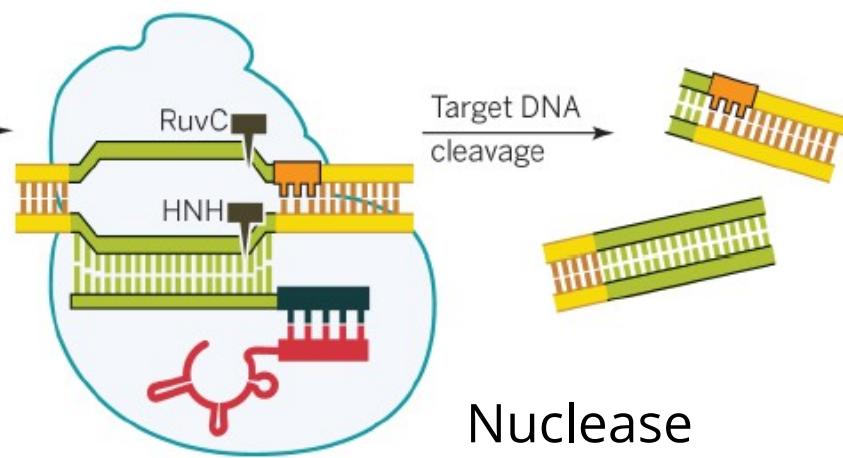
Tracr RNA interacts with Cas9, and this complex interacts with precursor crRNA to make the mature enzyme

# Cas9-RNA machinery in action

Cas9 finds PAM  
(protospacer adjacent motif)

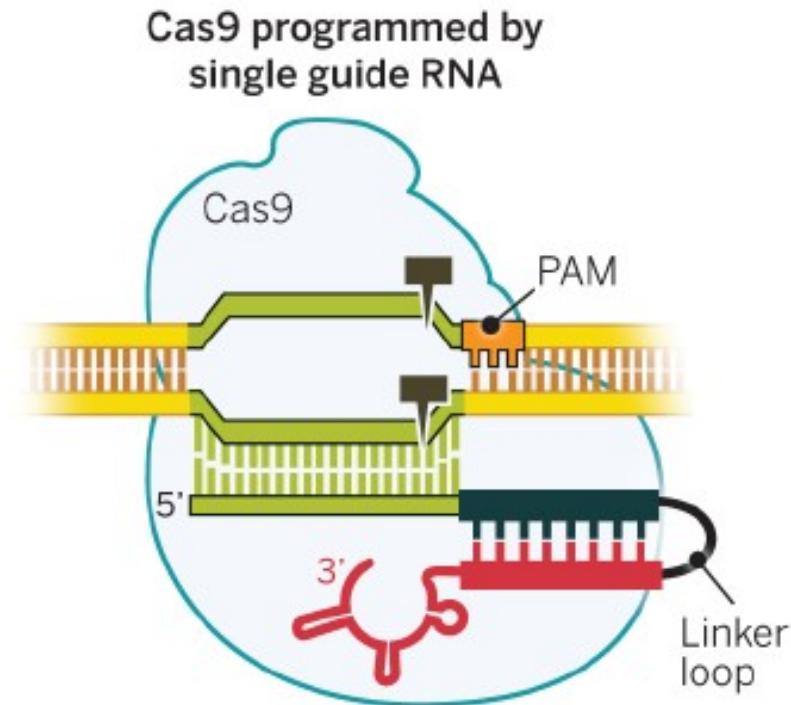
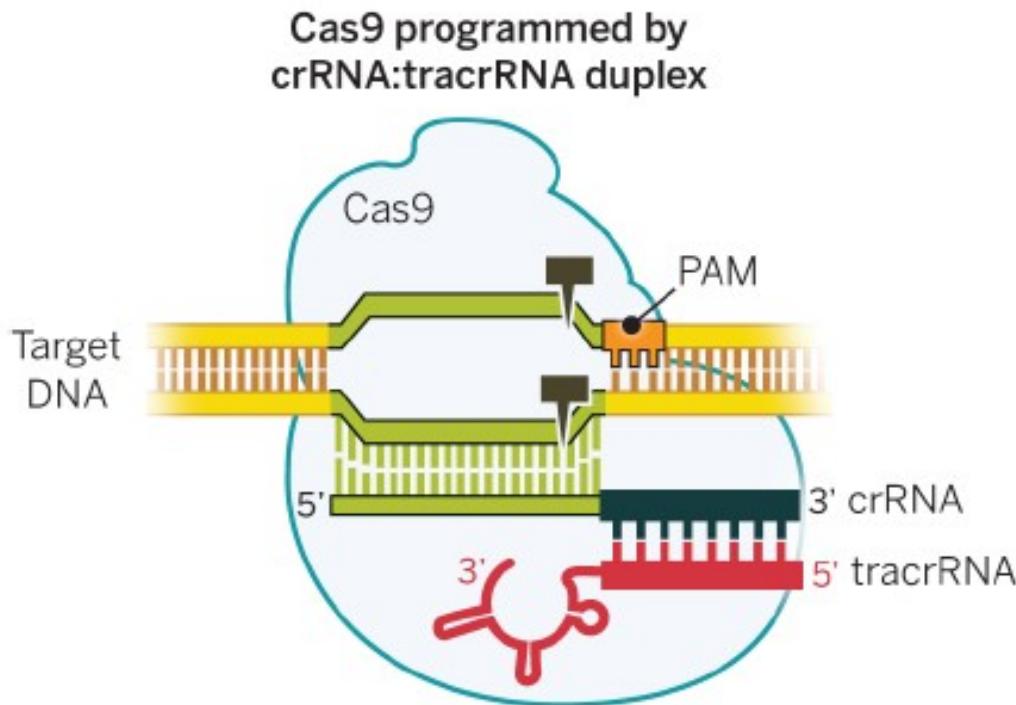


R-loop forms for gRNA hybridization



Nuclease activity cuts DNA

# An important innovation in utilizing this machine for engineering: fusion of crRNA and tracrRNA

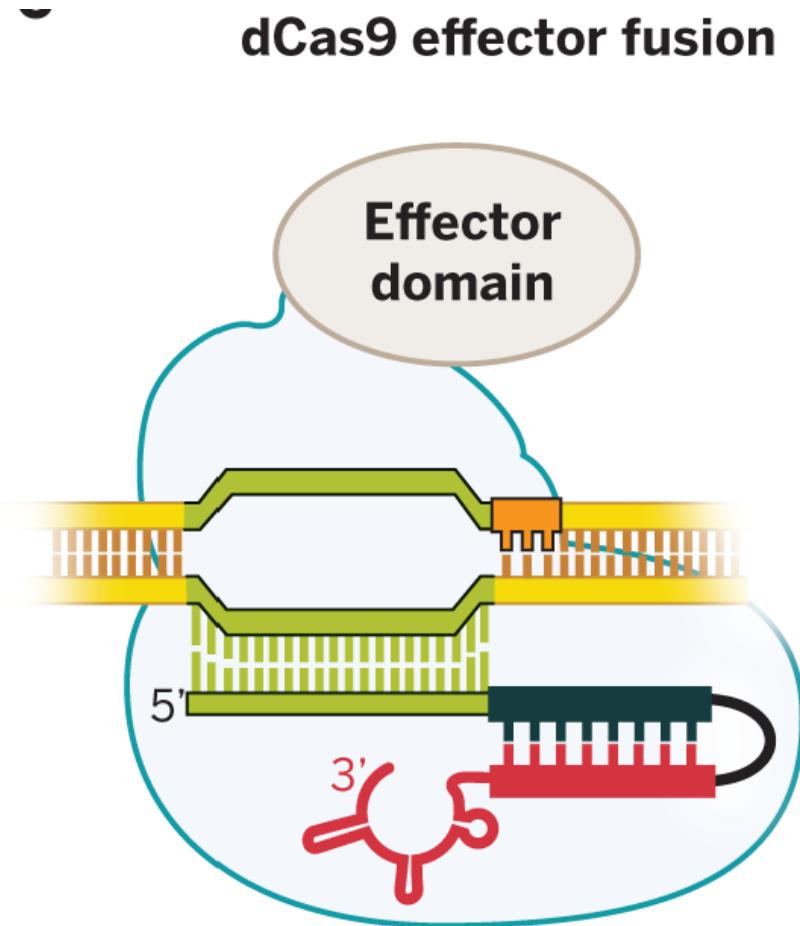


Normally, crRNA and tracrRNA are separate

They can be joined by a linker and the machine still functions

# Cas9 as a programmable DNA binding protein

Both nucleases can be inactivated, making Cas9 a DNA binding protein that can be programmed, and can take effector domains any place in the genome



# Proteins for manipulating DNA

- 1) Enzymes and other proteins require appropriate buffers and solution conditions for proper function
- 2) Specific tools and their uses
  - a) *Nucleic acid polymerases*: make and repair DNA
  - b) *Nucleases*: degrade DNA or RNA
  - c) *Restriction endonuclease*: break DNA at a specific site
  - d) *Ligase*: seal breaks in DNA
  - e) *CRISPR-Cas9*: a genomic homing device

When quantifying a double-stranded DNA sample (150 base pairs) in the lab using a spectrophotometer, the undiluted DNA gives an A<sub>260</sub> of 0.100.

- a) What is the concentration of your DNA sample, assuming that 1 A<sub>260</sub> = 50 micrograms/milliliter DNA
  
- b) If you have 85 microliters of this DNA sample left, how many nanograms of the DNA do you have ?

# Amplification of DNA *in vitro*

- I. Components of the PCR reaction
- II. Applied PCR/amplification techniques
  - a) Reverse transcription PCR (for RNA measurements)
  - b) Quantitative real-time PCR
  - c) PCR of long DNA fragments
  - d) Detection of an RNA virus: SARS nCov-2
    - a) PCR
    - b) LAMP
  - e) Whole genome amplification (WGA)

# Guide to readings: PCR

1) **"Discovering Life in Yellowstone... Where Nobody Thought It Could Exist"**: the story of the discovery of *Thermus aquaticus* by Brock and Freeze

2) **15 MC4 PCR:**

- a) Introduction to PCR, DNA pols, primer design (p. 455-69)
- b) Basic PCR protocol, and troubleshooting (p. 470-76)
- c) PCR topics: Hot Start, Touchdown, Taq (p. 477-83, 533-6)
- d) PCR primer design w/ Primer3Plus (p. 564-70)

3) **16 MC4 quantitative PCR**: Theory and practice

4) **Whole genome amplification** (2003). Protocols for indiscriminate amplification of DNA

5) **Guide to Primer3**. A short guide for the primer design program " Primer3" (complementary to section D above)

6) **EUA-CDC-Panel-IFU**. Applied PCR: the CDC protocol for its kit that detects SARS nCoV 2 RNA

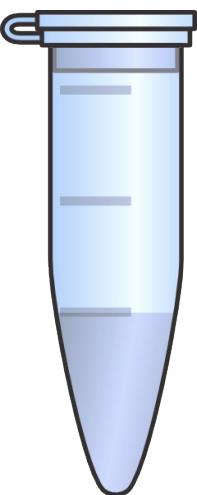
# The need for specific DNA segments

- DNA for **cloning** and **sequencing**, or for *in vitro* studies
- Confirm the identity of engineered DNA constructs
- Monitor gene expression
- Detect a genetic disorder
- Detect a microbe
- Identify an individual
- Etc.

# Multiple round DNA synthesis by thermal cycling -- PCR

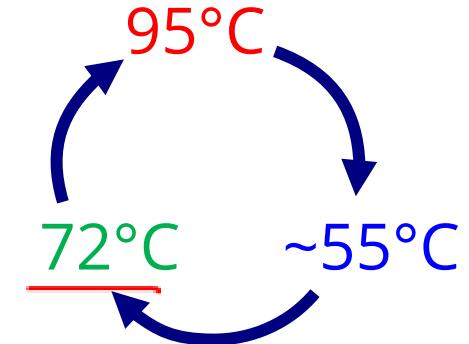
- Polymerase Chain Reaction -- first described in 1971 by Kleppe and Khorana
- Re-description and first successful use in 1985 (Mullis)
- Massive amplification of specific sequences that have defined endpoints
- Fast, powerful, adaptable, and simple
- Many many applications

# PCR: What you need in the tube



1. **Template DNA** that contains the “ target sequence”
2. **Primers:** short oligonucleotides that define the ends of the target sequence
3. Thermostable **DNA polymerase**
4. Buffer, dNTPs
5. Temperature cycles (provided by a thermal cycler)

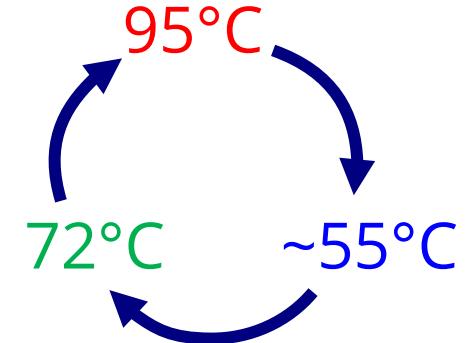
# PCR temperature cycling



## Denaturation:

- denature template strands (**95°C** for 2-5 minutes)
- can also add your DNA polymerase at this temp. for a “hot start”, which prevents false priming in the initial round of DNA replication

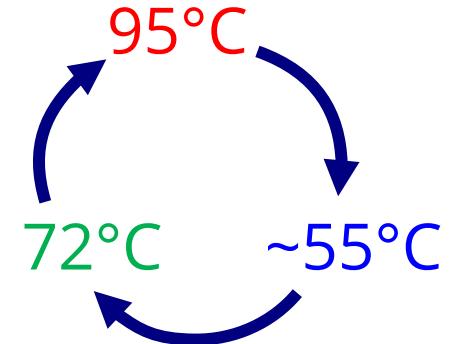
# PCR temperature cycling



## Annealing:

- The primers base pair (anneal, hybridize) to the template DNA
- The default T is around 55°C
- This temperature variable is the most critical one for getting a successful PCR reaction – the best variable to start with when trying to optimize a PCR reaction for a specific set of primers
- Annealing temperatures can go as low as 40-45°C, but non-specific annealing can be a problem

## A typical PCR program (part II):



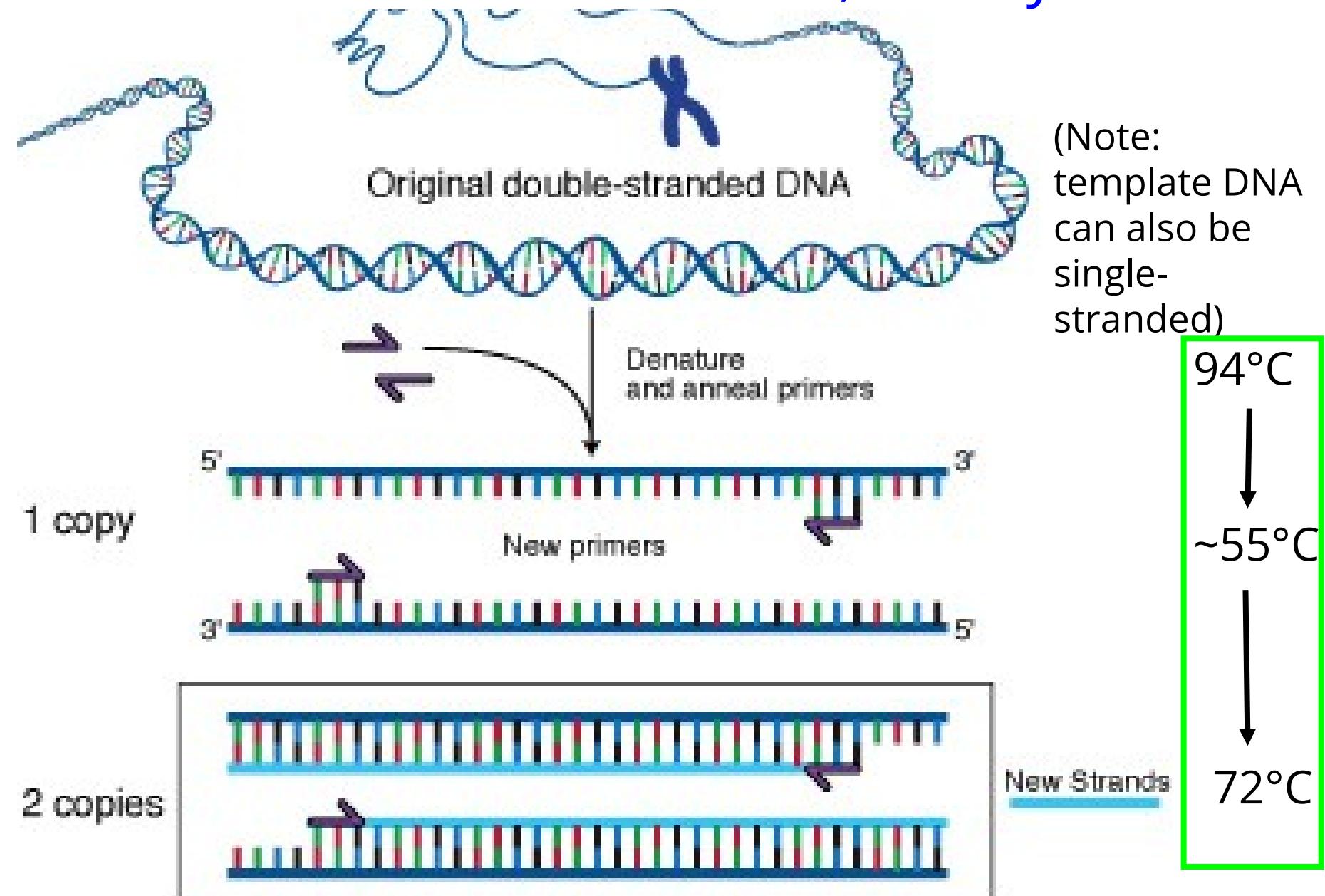
### Extension:

- DNA polymerase extends the primer, making a copy of the DNA template
- Generally 72°C, allows enzymatic activity of many thermostable DNA polymerases

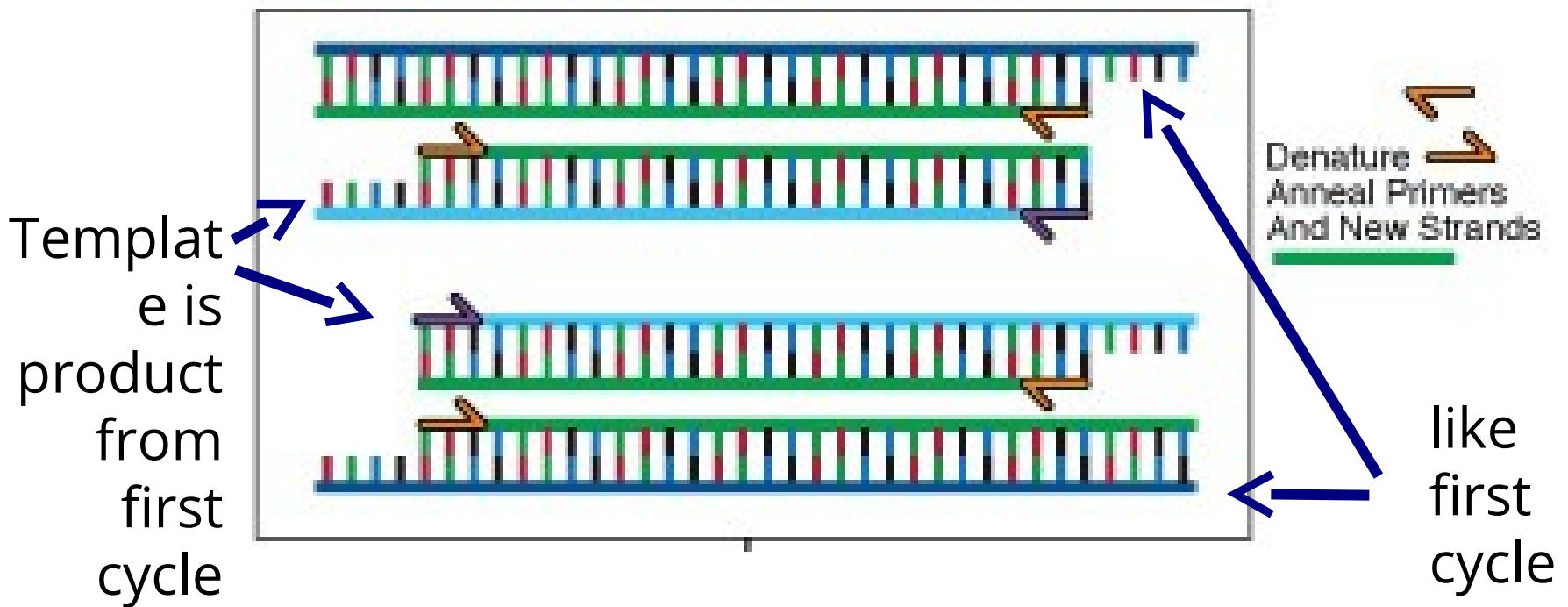
### Number of cycles:

- Each cycle repeats the temperature series:  
94 → ~55 → 72
- 20 to 30 cycles is typical

# How it works: PCR reaction, first cycle

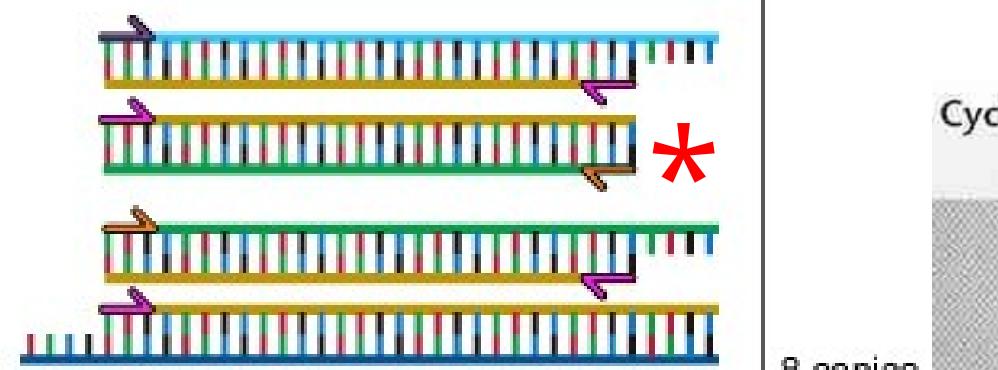
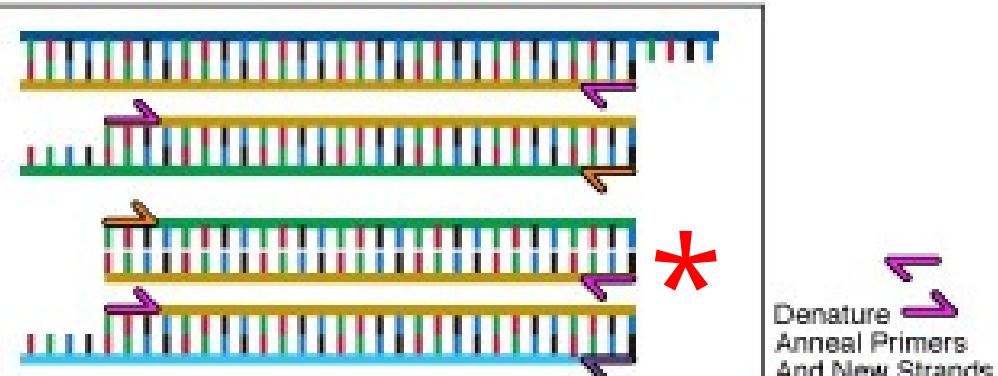


# PCR reaction, second cycle



# R reaction, third cycle: the first finished products

\* These products serve as templates for future rounds: copies of copies



↓  
20 -30 cycles

↓  
Millions and Millions of copies

Cycle number	Number of double-stranded target molecules
1	0
2	0
3	2
4	4
5	8
29	134,217,728
30	268,435,456

PCR animation:  
<https://dnalc.cshl.edu/resources/animations/pcr.html>

# How to choose primers for PCR?

- Should be ~18-25 nucleotides (can be longer)
- Calculated melting temperature ( $T_m$ ) should be nearly identical for both primers
- If possible, avoid inverted repeat sequences and self-complementary sequences in the primers
- If possible: avoid complementarity between primers (' primer dimers' )
- Use software to help, e.g. " Primer3Plus" : you can choose " detection" or " cloning" to define the type of primers you want

<http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/>

Design a pair of primers (10 nucleotides each) that will amplify DNA from this template.

Label the 5' and 3' end of each primer

5' gtgttgttatt tgtctgaaga gtatccacct 3'  
3' cacaacataaa acagacttct cataggtgga 5'

# The primers for PCR are often modified

## 1) Tags

- Biotin
- Fluorescent tag
- See for example: IDT.com

## 2) Extra non-complementary sequences may be added to the primer (at the 5' end only!)

- For cloning purposes
- For mutagenesis
- To give a specific sequence identifier tag

# Thermostable DNA polymerases from thermophile microbes

- Bacterial
  - Taq, from *Thermus aquaticus* (discovered by Dr. Tom Brock, see "Discovering Life in Yellowstone....")
  - High efficiency, but low fidelity
  - Excellent for routine reactions and small PCR products
- Archaeal
  - Pfu, from *Pyrococcus furiosus*
  - Lower efficiency, but high fidelity'
  - also good for routine reactions and best for cloning
    - 3' → 5' exonuclease activity provides very high fidelity
    - Very stable to heat

# Thermal cycling

## I. Standard heat block

- “ramp” times 5-10 seconds to change temperature, 30 cycle PCR lasts 2-3 hours.
- Advantage: easily automated, heat blocks can PCR up to 384 samples at a time
- Disadvantage: relatively slow (1-3 hours)

## II. Capillary tubes with heat source

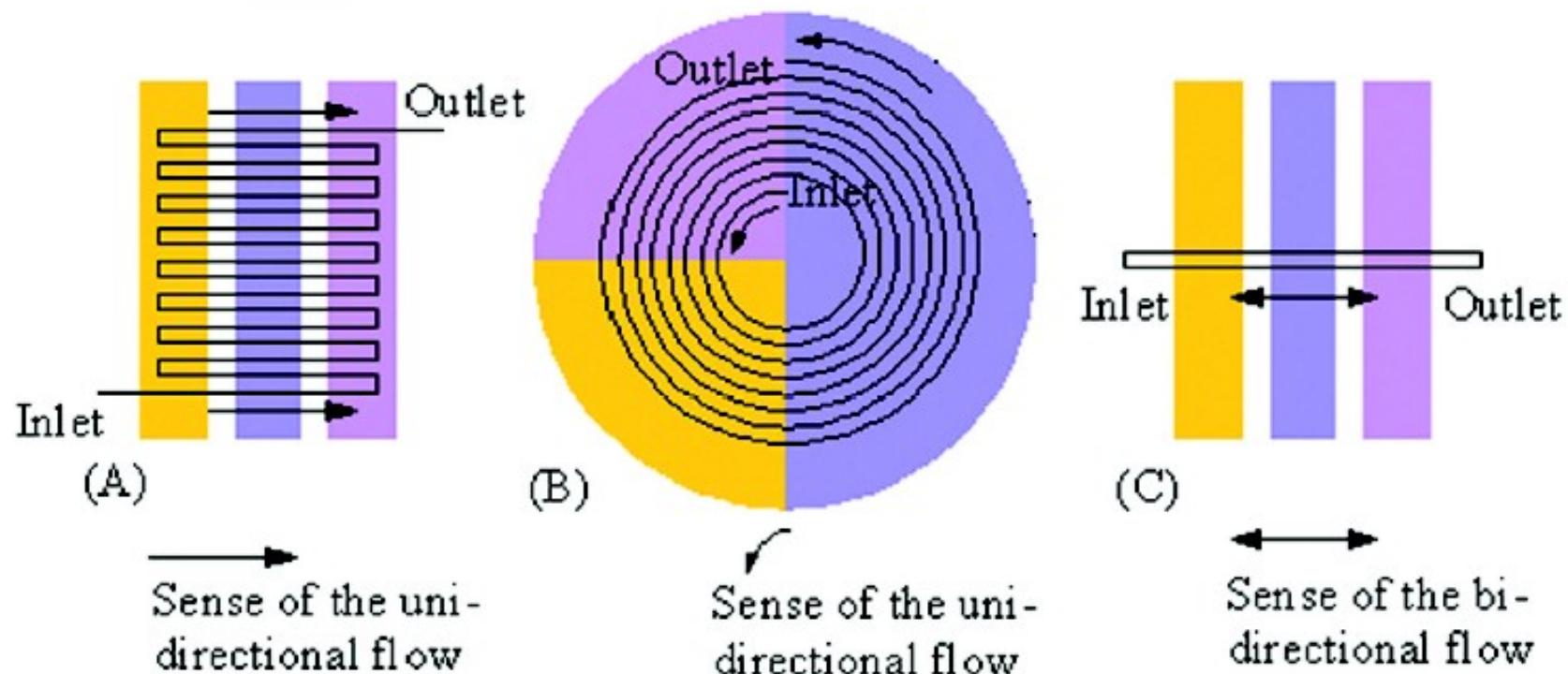
- heated and cooled by blasts of air
- 30 cycle-PCR done in < 30 minutes, but it's limited to only a few samples at a time

## III. Continuous flow

- channels force liquid through temperature gradients, *very* fast (seconds to minutes)

## Super-fast chip-based PCR

- [Yellow] Denaturation temperature
- [Purple] Extension temperature
- [Magenta] Annealing temperature



**Figure 2.** Continuous-flow PCR. (A) The serpentine channel continuous-flow PCR. (B) The spiral channel-based continuous-flow PCR. The sample is introduced at the inlet and pumped unidirectionally towards the outlet. (C) The straight channel oscillatory-flow PCR. The sample is introduced in the inlet and pumped back and forth in a straight channel. Temperature zones are provided by three heaters.

# Hot Start of PCR reactions

- non-specific priming occurs at low temperature (room temp.) -- the non-specific priming could give artifactual amplification as temperature rises in the PCR tube
- Withhold some component of the reaction until the denaturing temperature is reached (94°C)
  - Wait until 94°C to add enzyme  
or
  - Enzyme bound to an inactivating enzyme antibody that releases at high temperature  
or
  - Wax beads containing Mg++ that can only be released at high temp

# Touchdown PCR: improve target specificity

Allows you to selectively amplify only the best sequences (with the least mismatches) while minimizing non-specific PCR products

- Start with 2 cycles at an annealing temperature about 5-10°C higher than the calculated primer melting temperatures.
- Progressively reduce the annealing temperature by 1°C at 1 or 2 cycle intervals
- Final cycles of PCR done at annealing temp 2-5°C lower than calculated annealing temp

Useful if your primers are not 100% complementary to your template DNA (e.g. degenerate oligos), or when there are

# Difficult PCR? Be sure to include controls

	Primers	Bystander DNA	template DNA	Known target DNA	Expected result
Your template	+	-	+	-	?

Template DNA:  
The DNA being tested

## Positive controls

1	+	-	-	+	Band
2	+	+	-	+	Band

Known target DNA: known to contain primer recognition sequences

## Negative controls

3	+	-	-	-	No
4	+	+	-	-	No

Bystander DNA: not recognized by

Deviation from expected results gives ideas for troubleshooting

# Trouble-shooting a failed PCR

- No PCR product
- Very little product
- Wrong-sized DNA bands on gel
- Etc.

## Remedies:

- Temperature or solution conditions may need changing
- Template DNA may have contaminants – repurify DNA
- Primers may need to be re-designed

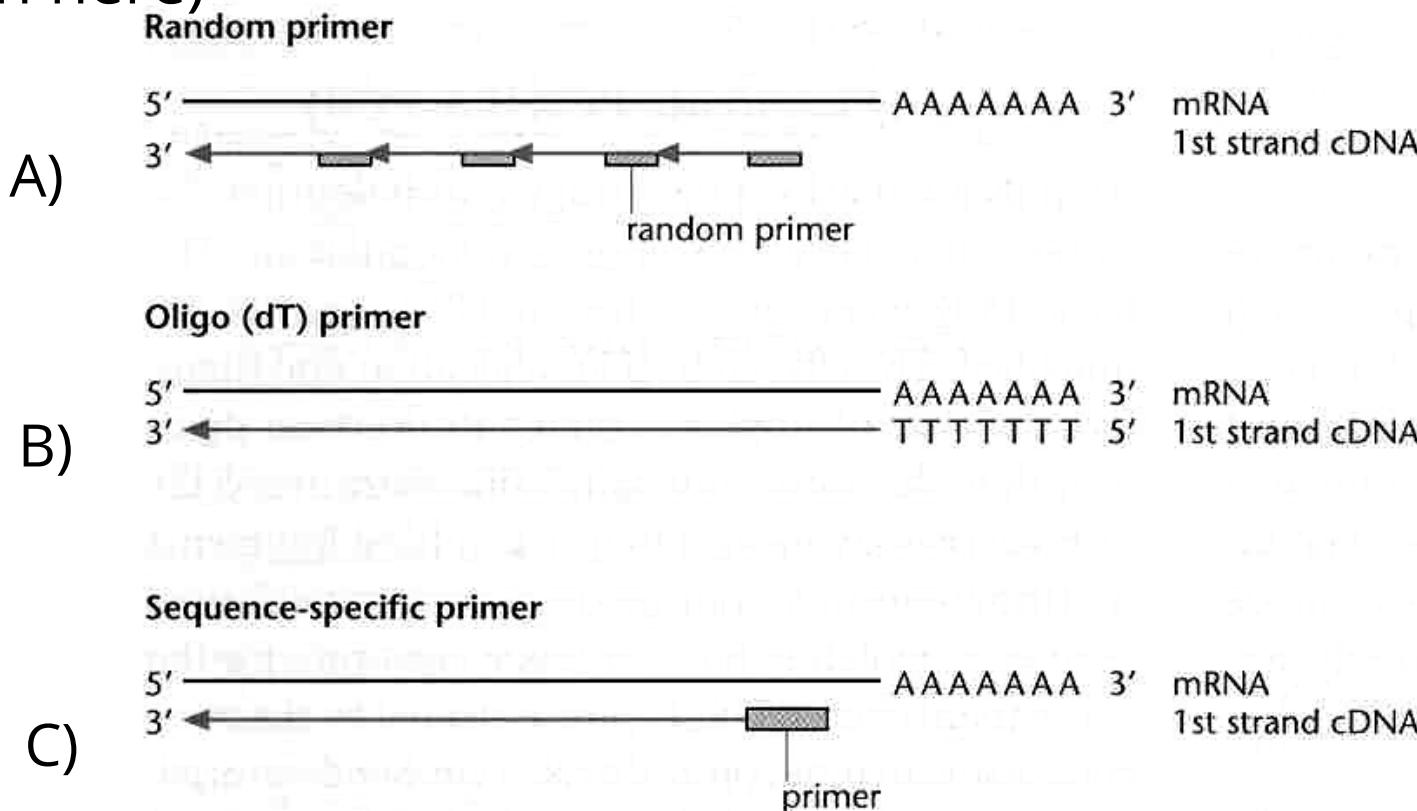
(see 15 MC4 PCR, p. 473-6)

## II. Specific applications for DNA amplification

- A. Reverse transcription PCR (for RNA measurements)
- B. Quantitative (real-time) PCR
- C. PCR of long DNA fragments
- D. Whole genome amplification

# Detection of RNA (gene expression, or RNA virus): reverse transcription followed by PCR

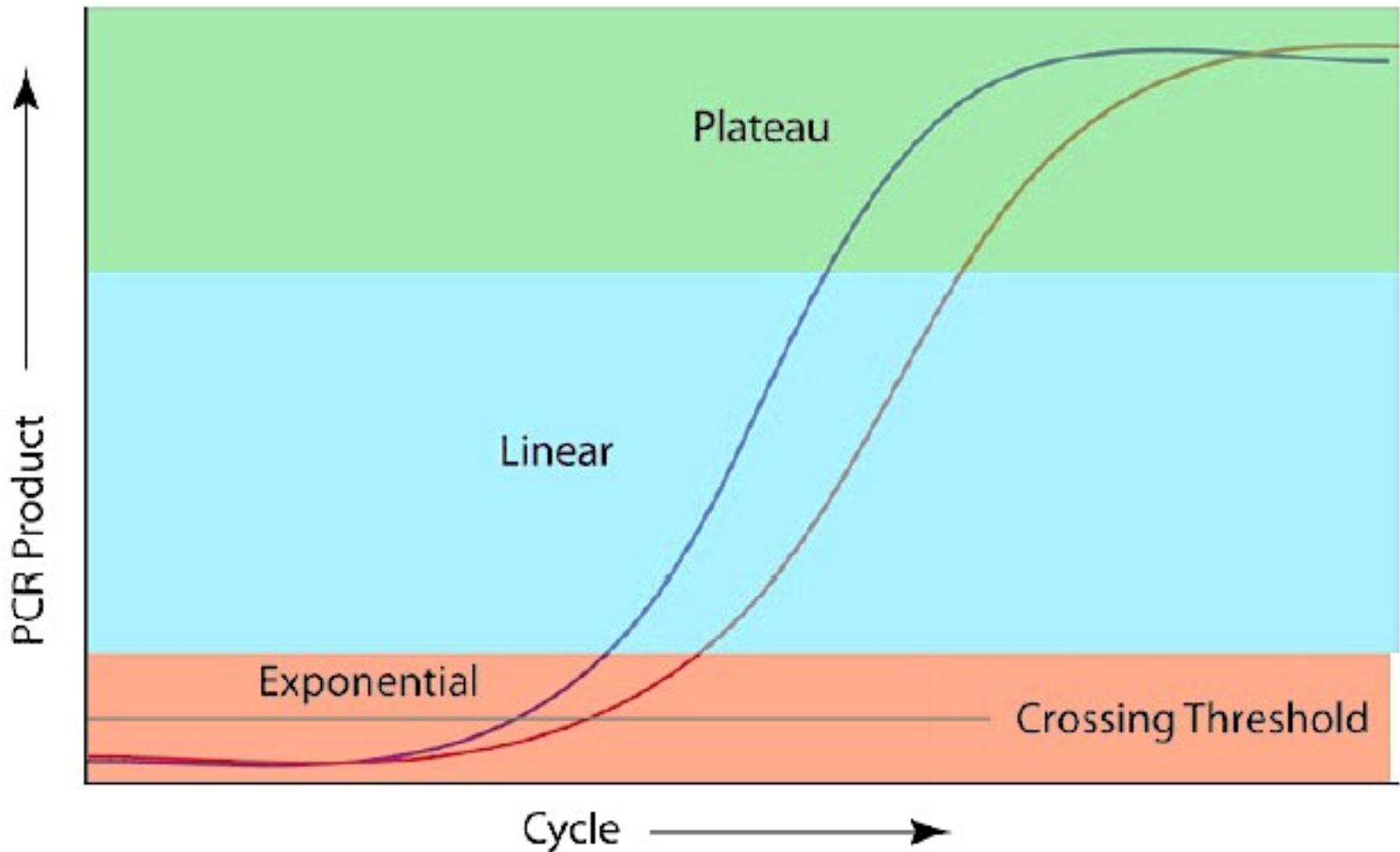
**Step 1:** make cDNA with reverse transcriptase (three ways shown here)



**Step 2:** normal PCR (from cDNA) using gene-specific primers

NOT QUANTITATIVE (end point DNA level doesn't report RNA levels)

## PCR reaction progress as temperature cycling progresses



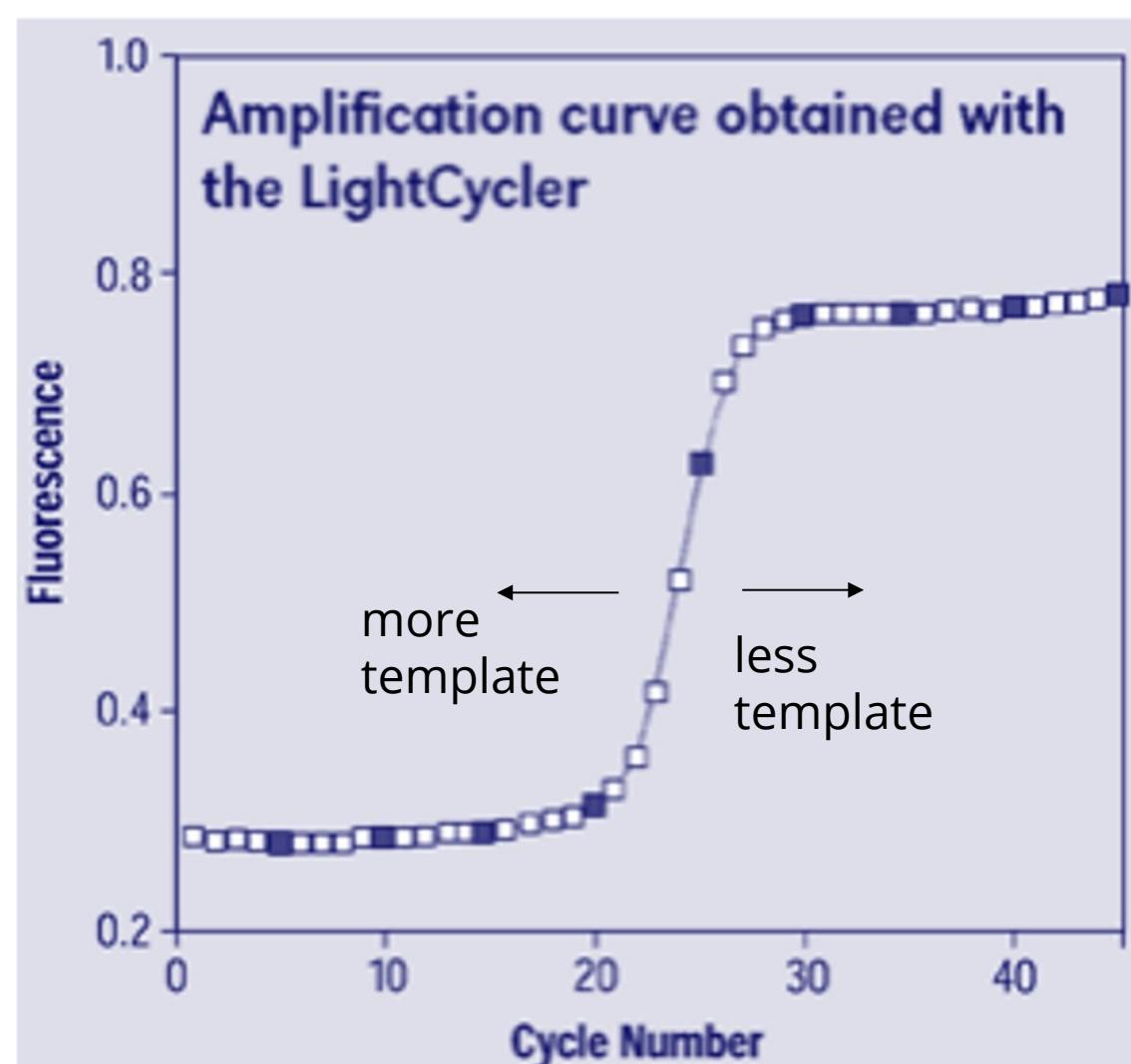
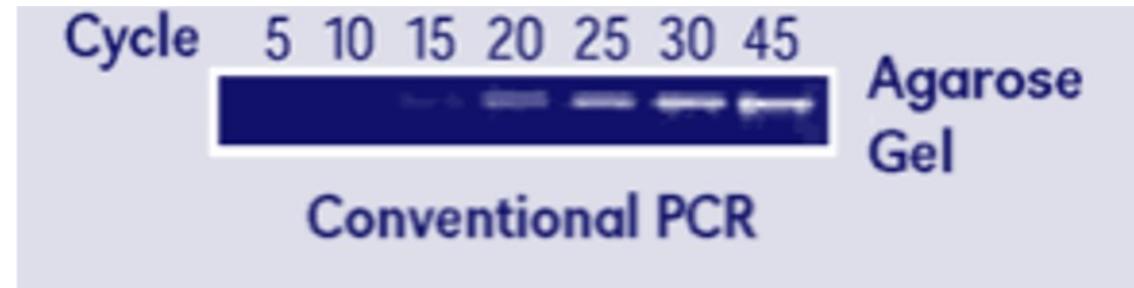
Each line represents a different template RNA/cDNA  
Which one is present in the highest quantity?

# Quantitative Real Time (QRT) PCR

Fluorescence measurements are done simultaneously with PCR temperature cycles

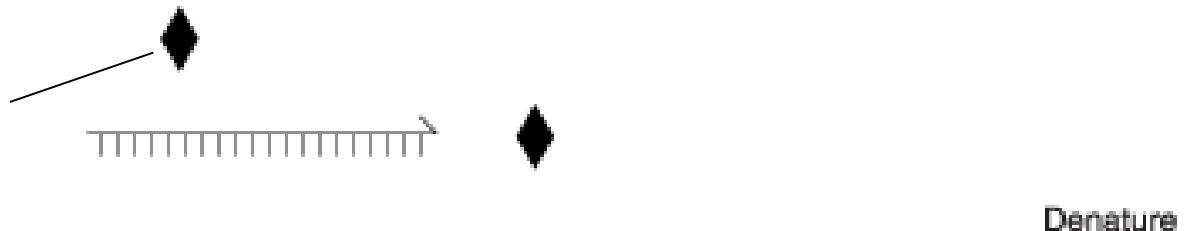
Instantaneous measurement of product levels

Position of the center of the curve changes depending on the amount of template RNA/cDNA. *Variations of over 5 or 6 orders of magnitude can be*

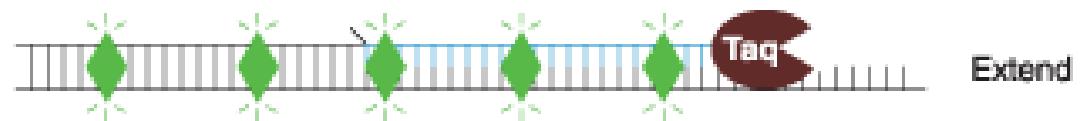
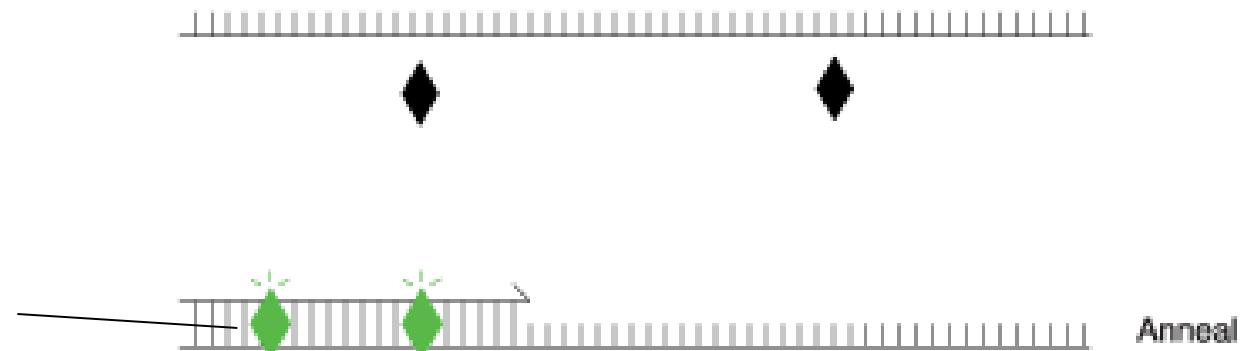


# Dye-binding: more DNA, more fluorescence

Non-fluorescing SYBR  
green dye



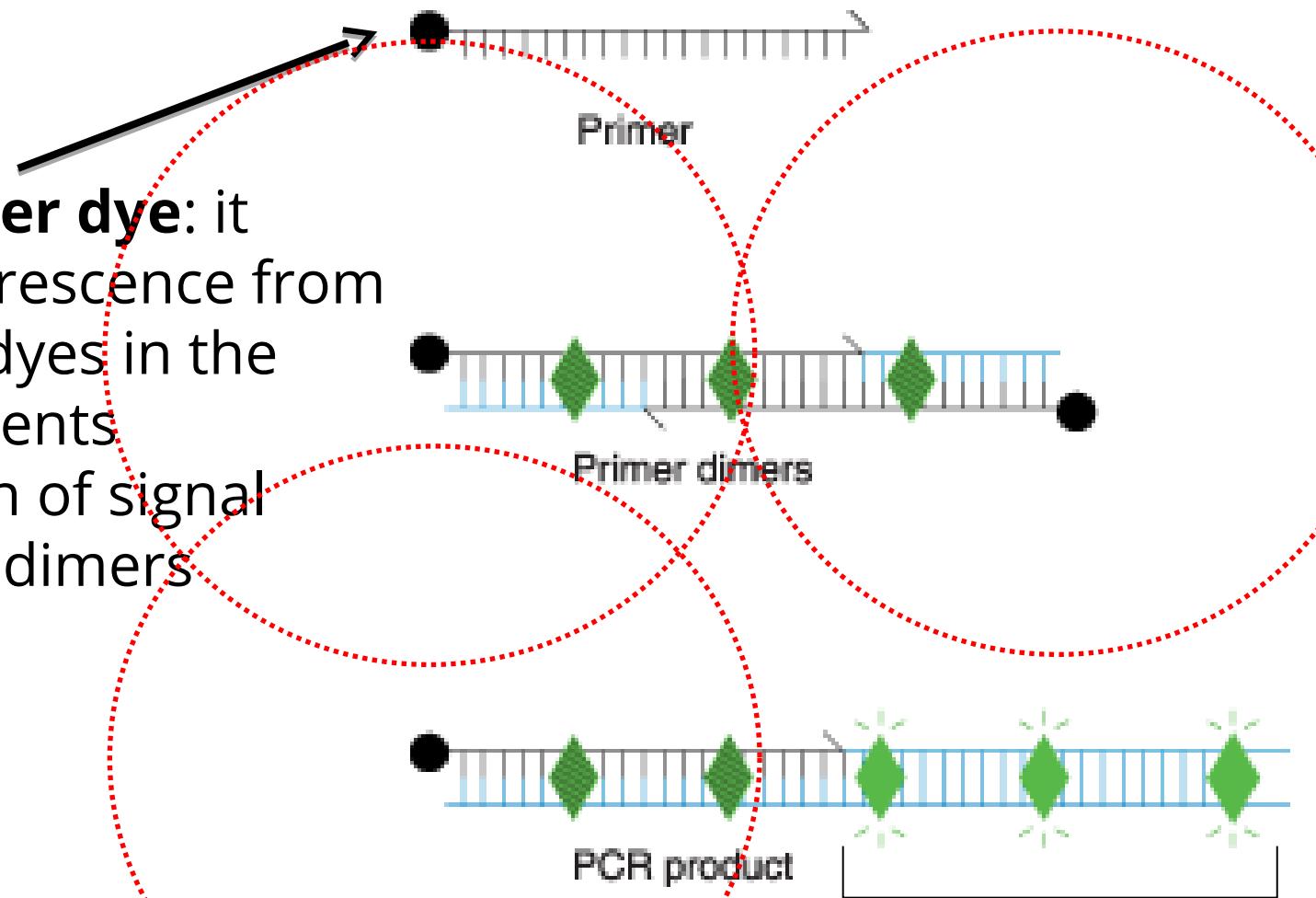
Fluorescing SYBR  
green dye



Low cost, but detects all DNA, including artifacts

# Primer dimers can give false signal with dye binding

**QSY quencher dye:** it absorbs fluorescence from SYBR green dyes in the vicinity--prevents accumulation of signal from primer dimers

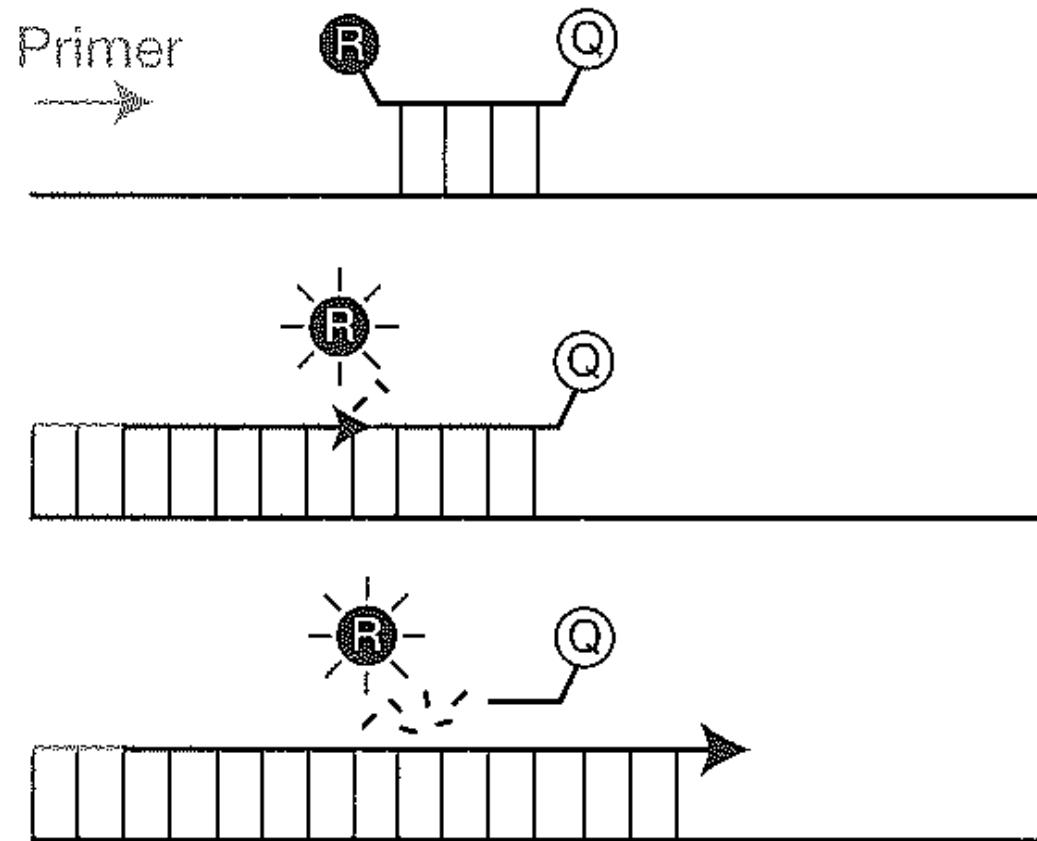


This can be avoided by short-range 'quenching' of fluorescence

## Fluorescent probes: removal of 'quencher' based on product accumulation

### TaqMan probes

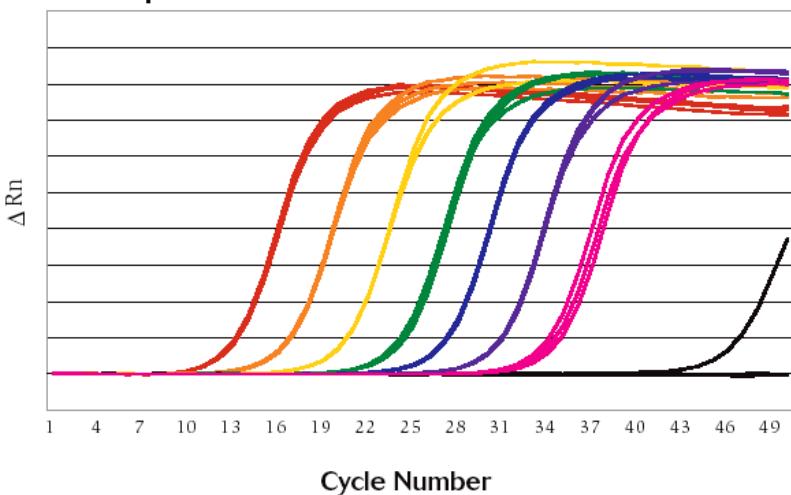
The more target DNA there is, the more probe anneals, the more it is cleaved (by Taq polymerase  $5' \rightarrow 3'$  exonuclease activity), the more fluorescence is produced



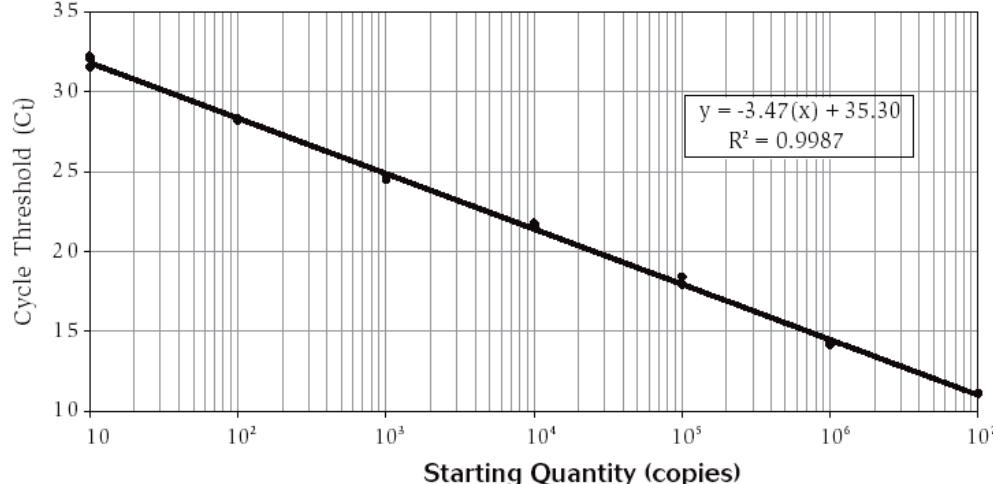
Each experimental target needs its own probe

# Control amplifications allow precise quantitation:

A. Amplification Plot



B. Standard Curve



Real-time quantitative PCR of 10-fold serial dilutions ( $10^7$  to 10 copies) of pCR2.1 plasmid were performed using primers specific to the Kanamycin resistance gene (200 nM each) with Platinum® SYBR® Green qPCR SuperMix-UDG and ROX Reference Dye. Reactions were incubated for 2 min. at 50°C, then 2 min. at 95°C, followed by 50 cycles of 95°C for 15 sec.; 60°C, 30 sec. using the ABI PRISM® 7700.

Standard curve: based on the cycle at which “ threshold” (of detection) is reached for a specific number of DNA molecules

(From the Invitrogen website)

Recommendations for the CDC regarding detection of SARS n-CoV 2 infection: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/testing-overview.html>

- Many commercial tests available
- CDC offers its own test:  
<https://www.cdc.gov/coronavirus/2019-ncov/lab/virus-requests.html>
- RT-PCR based

# Detection of 2019-nCov in patient samples (CDC protocol)

- Collect sample, extract RNA (Trizol reagent treatment)
- Add primers for
  - (step 1) Reverse transcriptase, making DNA copy of the RNA genome
  - (step 2) PCR amplification of nucleocapsid gene using primers specific to 2019-nCoV version of the gene
  - (step 3) Detection of the amplified DNA using a “TaqMan” probe approach (fluorophore revealed when probe binds target and gets degraded)
    - <https://www.biostarsearchtech.com/support/videos/real-time-pcr-probe-animation-video>

## 2019-Novel Coronavirus (2019-nCoV) Real-time rRT-PCR Panel Primers and Probes

Name	Description	Oligonucleotide Sequence (5'>3')	Label <sup>1</sup>	Working Conc.
2019-nCoV_N1-F	2019-nCoV_N1 Forward Primer	5'-GAC CCC AAA ATC AGC GAA AT-3'	None	20 µM
2019-nCoV_N1-R	2019-nCoV_N1 Reverse Primer	5'-TCT GGT TAC TGC CAG TTG AAT CTG-3'	None	20 µM
2019-nCoV_N1-P	2019-nCoV_N1 Probe	5'-FAM-ACC CCG CAT TAC GTT TGG TGG ACC-BHQ1-3'	FAM, BHQ-1	5 µM

Primers and probe for nucleocapsid protein gene

Control primers and probe for human RNase P gene

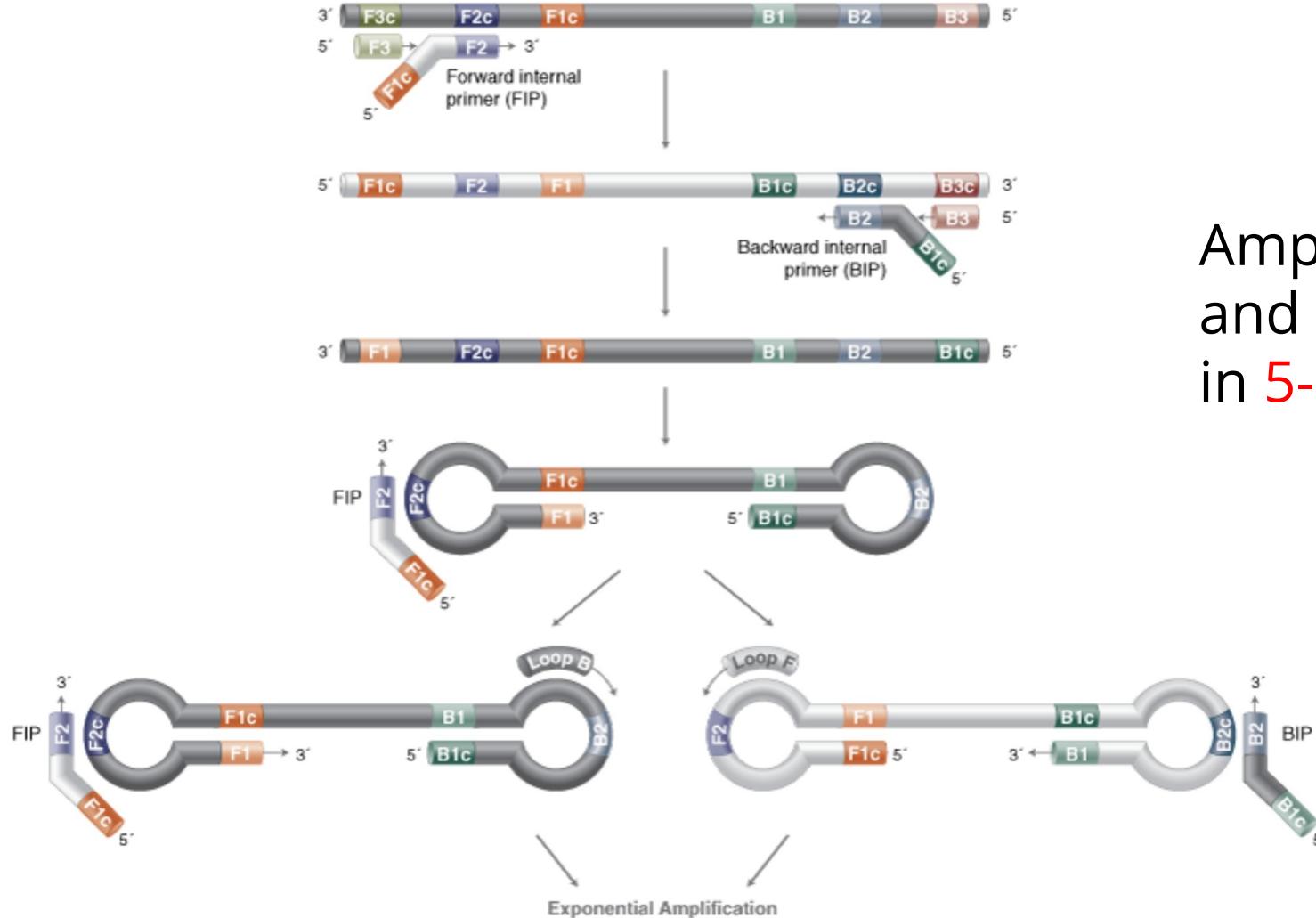
RP-F	RNAse P Forward Primer	5'-AGA TTT GGA CCT GCG AGC G-3'	None	20 µM
RP-R	RNAse P Reverse Primer	5'-GAG CGG CTG TCT CCA CAA GT-3'	None	20 µM
RP-P	RNAse P Probe	5'-FAM – TTC TGA CCT GAA GGC TCT GCG CG – BHQ-1-3'	FAM, BHQ-1	5 µM

<sup>1</sup>TaqMan® probes are labeled at the 5'-end with the reporter molecule 6-carboxyfluorescein (FAM) and with the quencher, Black Hole Quencher 1 (BHQ-1) (Biosearch Technologies, Inc., Novato, CA) at the 3'-end.

Note: Oligonucleotide sequences are subject to future changes as the 2019-Novel Coronavirus evolves.

Other tests: Isothermal amplification allowing rapid amplification of target DNA

## LAMP: loop mediated isothermal amplification



Amplification  
and detection  
in 5-15 min

# PCR of long sequences (>2 kb)

Long DNAs can be challenging to amplify

- Discontinuity (breakage) within the target DNA sequence reduces number of 'good' templates
- DNA polymerase is not 100% processive (it falls off or degrades before finishing)
- Misincorporation by error prone DNA polymerases causes mutations in the product: longer sequence = more mutations

# PCR of long sequences (>2 kb)

Some changes to protocol to assist in long PCR

- Make sure DNA is exceedingly clean & prepared without breakage
- Use DNA polymerase “ cocktail” : Taq for high activity, and Pfu for proofreading activity (it can correct Taq’s mistakes)
- Increase time of extension reaction (5-20 minutes, compared to the standard 1 minute for short PCRs). The rule of thumb is 1 minute per kilobase of DNA amplified
- Use DNA polymerases engineered to be more processive

# Whole genome amplification: multiple displacement amplification (MDA)

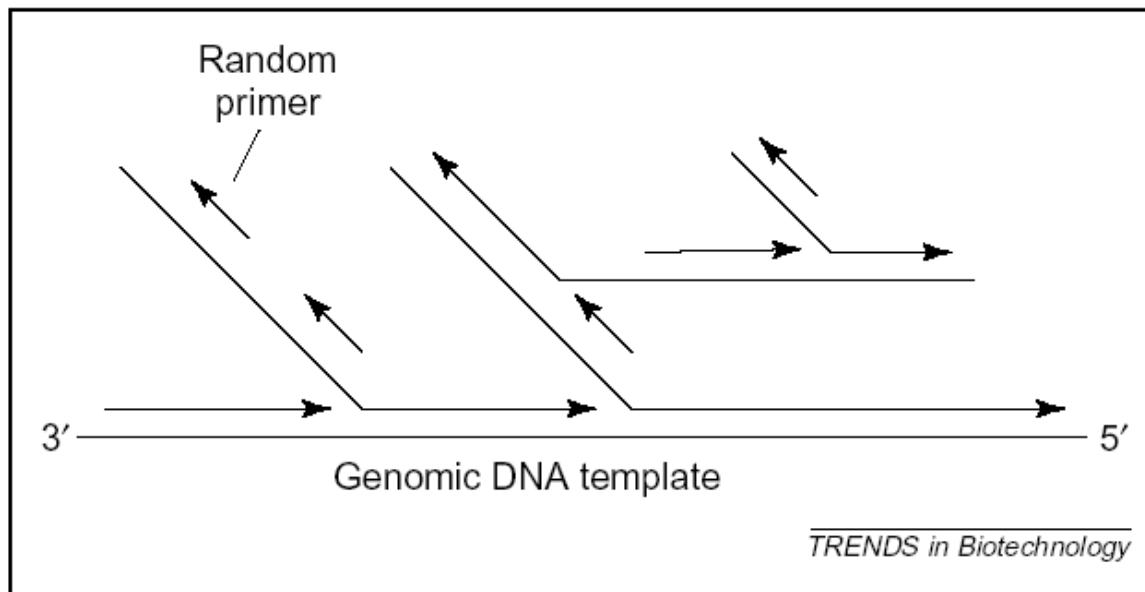
## How it works:

Strand-displacement amplification used by rolling-circle replication systems.

- Phi29 DNA polymerase (very low error rate)
- *Bacillus stearothermophilus* DNA polymerase (Bst) is also sometimes used (higher error rate)
- Primers: random hexamer (6 nucleotide)
- Incubation temperature: 30°C. No cycling.

Applications: forensics, *in utero* disease diagnosis, microbial diversity surveys, single cell genome

# Whole genome amplification : multiple displacement amplification (MDA)



**Figure 1.** Multiple displacement amplification reaction. DNA synthesis is primed by random hexamers. Exponential amplification occurs by a 'hyperbranching' mechanism. Unlike PCR, which requires thermal cycling to repeatedly melt template and anneal primers, the  $\phi$ 29 DNA polymerase acts at 30°C to concurrently extend primers as it displaces downstream DNA products.

20-30 micrograms human DNA can be recovered from 1-10 copies of the human genome

Products arise from a random sampling of the available template – this is the least biased method for amplification of



# In vitro amplification of DNA



- I. Components of the PCR reaction
- II. A few advanced applications of PCR
  - a) Reverse transcription PCR (for RNA measurements)
  - b) Quantitative real-time PCR
  - c) PCR of long DNA fragments
  - d) Whole genome amplification (WGA)

Design two PCR primers that will amplify **the entire** portion of this DNA fragment. Indicate the position of these primers relative to the DNA, the sequence of the primers, and their 5' and 3' ends.

5' GTGAATAAGCAAAAGGTTGCCTGCTGTGAATCTGCGGAAC TTATTGATCCA  
GAGAGGGGGAAATAGTCTGTGCCAAGTGC GGTTATGTAATAGAAGAGAACATAATTGA  
TATGGGT CCTGAGTGGCGTGCTTTGATGCTTCTCAAAGGGAACGCAGGTCTAGAACTG  
GTGCACCAGAAAGTATTCTTCTTCATGACAAGGGGCTTCAA CTGAAATTGGAATTGAC  
AGATCGCTTCCGGATTAATGAGAGAGAACATGTACCGTTGAGGAAGTGGCAGTCCAG  
ATTAAGAGTTAGTGATGCAGCAGAGAGGAACCTAGCTTTGCCCTAAGTGAGTTGGATA  
GAATTACTGCTCAGTAAAACCTCCAAGACATGTAGAGGAAGCTGCAAGGCTGTAC  
AGAGAGGCAGTGAGAAAGGGACTTATTAGAGGTAGATCTATTGAGAGCGTTATGGCGGC  
ATGTGTTACGCTGCTTGTAGGTTATTAAAAGTTCCCAGGACTCTGGATGAGATTGCTG  
ATATTGCTAGAGTTGATAAAAAGGAAATTGGAAGAAGTTACAGATTGCGAGAAAT  
CTCAATTAACTCCCCAAAAACTATTTGTCAAGCCA ACTGATTATGTA AAT AAAATTG  
GGATGAGCTCGGATTAAGTAAAAAGTTAGGAGAAGAGCTATTGAAATTCTTGTAGGAGG  
CTTATAAAAAGGGGGTTAACTAGTGGTAAGAGTCCAGCTGGTTAGTAGCAGCAGCCCTA  
TACATAGCTTCTTATTGGAGGGAGAGAACACAAAGAGAACAGTTGCCAAGTTGC  
TAGAGTAACTGAAGTGA C TGTGAGAAATAGATACAAGGAGCTCGTAGAGAACAGTTGAAGA  
TTAAAGTT CCTATAGCATGA 3'

Add extra sequences, containing restriction sites, to primers to help make them easier to clone. Indicate where those sequences should go.

One of the primers:

5'      3'

You attempt the PCR, but you run the gel and you see no DNA products. Suggest two things that may have gone wrong.

# DNA sequencing methods

- I. Chain termination (ddNTP) sequencing
- II. “Next generation” sequencing
- III. Sequencing genomes

## Guide to readings:

- 1) 17 MC4 *DNA sequencing*. Intro to sequencing techniques. Also protocol on “shotgun” sequencing
- 2) 18 MC4 *Next generation sequencing*. Advances in sequencing that have allowed very high ‘throughput’
- 3) 10 years of Next gen. A review of next generation sequencing technology over its first 10 years.
- 4) Nanopore sequencing 2012 and 2016. A revolutionary shift in sequencing approaches
- 5) Panda genome perspective 2010
- 6) Genome sequencing futures 2021

# DNA sequencing in biology

- Genomic DNA:
  - all of the DNA available for an organism to use -- an important tool for studying biology (pathogens, crops, economically important microbes, etc.)
  - Sequences of genes, and also positioning of genes and sequences of regulatory regions and features
  - Human genomes: how much variability from person to person, or from normal to cancerous cells?
- RNA sequence (via cDNA): which genes are expressed, and how much are they expressed?
- Recombinant DNA projects: keep track of constructions, follow progress of experiments

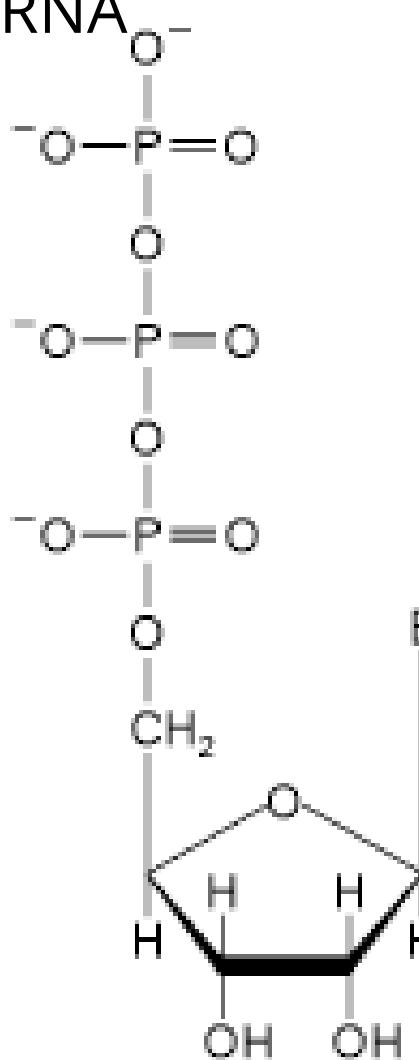
# Methods for DNA sequencing

- A. Sanger dideoxy (primer extension/chain-termination) method: the original protocol for genome sequencing, adaptable, scalable to large sequencing projects
- B. Next generation sequencing: many reactions at the same time
- C. Sequencing a genome – break the DNA, sequence it, and put it back together

# for dideoxy sequencing:

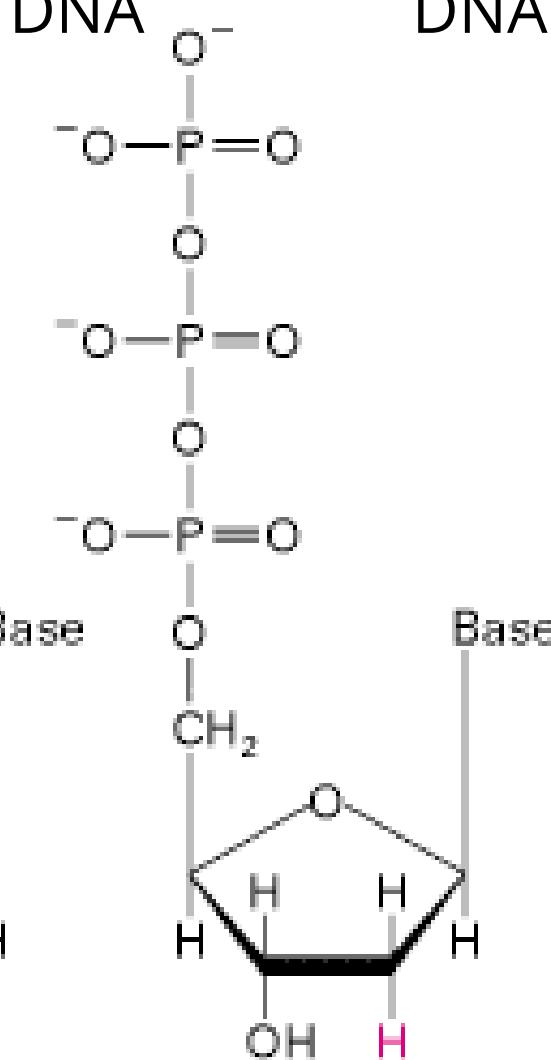
- 1) DNA template
- 2) An oligonucleotide primer for DNA synthesis
- 3) DNA polymerase
- 4) Deoxynucleoside triphosphates and  
dideoxynucleotide triphosphates

RNA



Ribonucleoside  
triphosphate

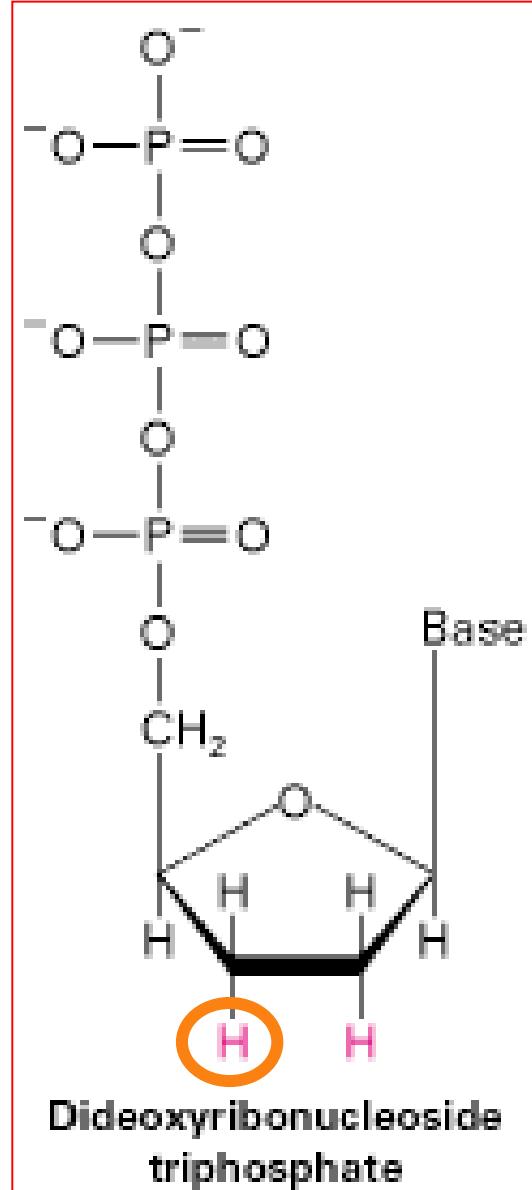
DNA



Deoxyribonucleoside  
triphosphate

rNTP

DNA



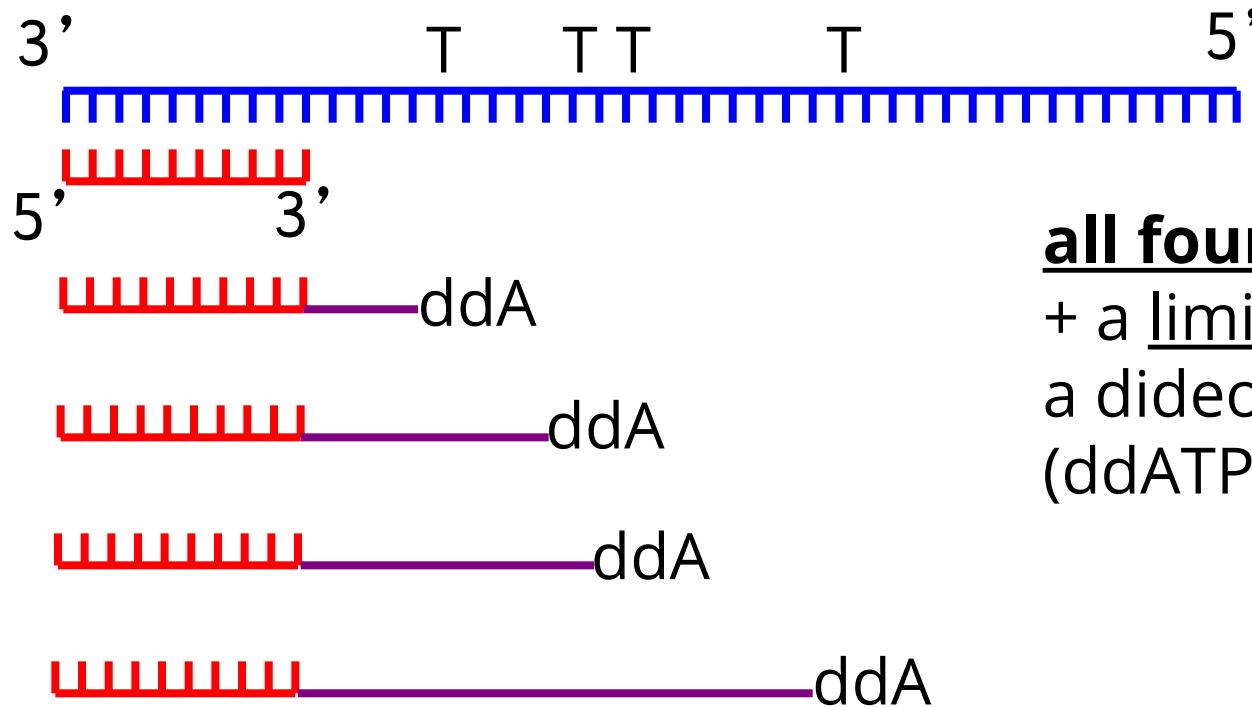
Dideoxyribonucleoside  
triphosphate

ddNTP: no 3' -OH

# DNA polymerase for sequencing

- Highly processive, NO exonuclease activity
- Able to use dideoxy NTPs relatively efficiently
- Sometimes thermostable DNA pols are useful in sequencing

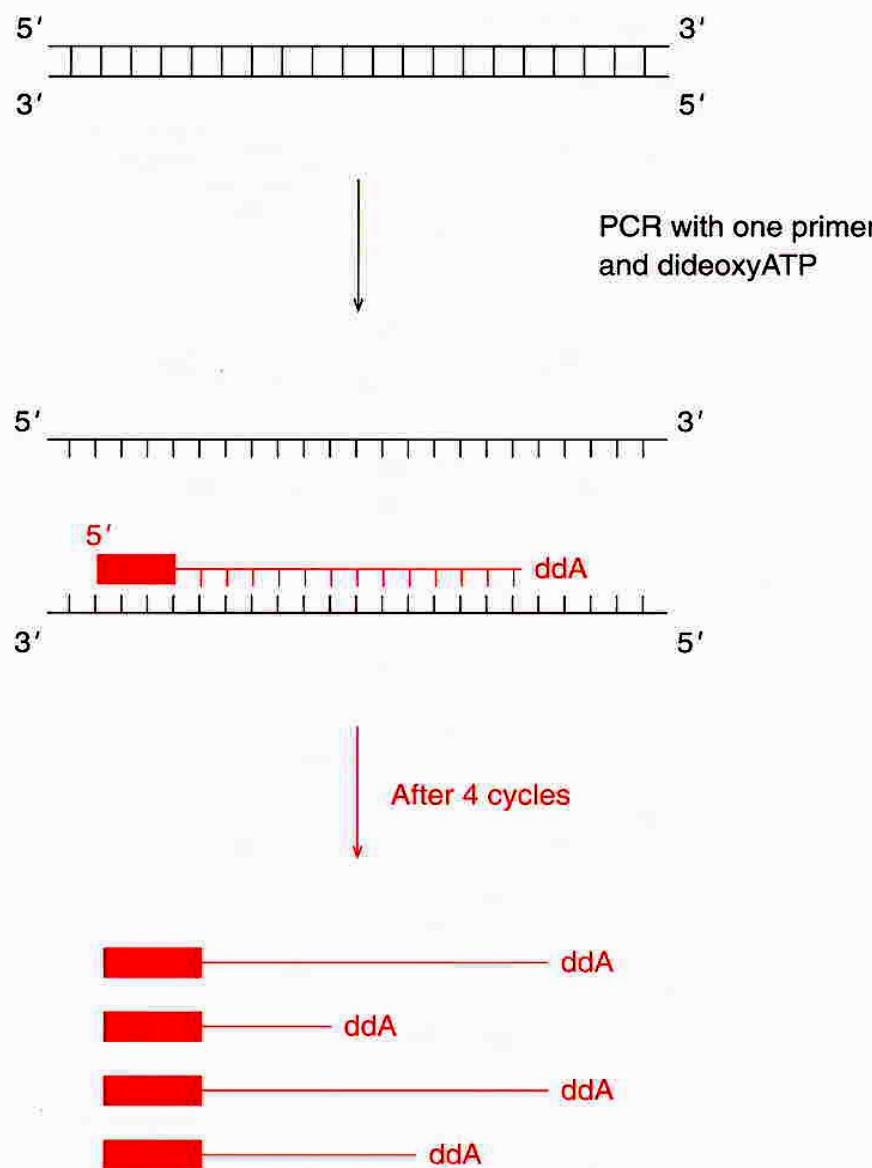
# Sanger dideoxy sequencing: chain termination of DNA synthesis



**all four dNTPs**  
+ a limited amount of  
a dideoxy NTP  
(ddATP)

ddATP in the reaction: anywhere there's a T in the template strand, occasionally a ddA will be added to the growing strand

# Cycle sequencing: denaturation occurs during temperature cycles



94°C:DNA denatures

45°C: primer anneals

60-72°C: thermostable  
DNA pol extends primer

Repeat 25-35 times

# Detection of the DNA fragments

- Radioactivity
  - Radiolabeled primers (kinase with  $^{32}\text{P}$ )
  - Radiolabelled dNTPs ( $\gamma^{35}\text{S}$  or  $^{32}\text{P}$ )
- Fluorescence
  - ddNTPs chemically synthesized to contain a different fluorophore
  - Each fluorophore is a different color

# Analysis of sequencing products:

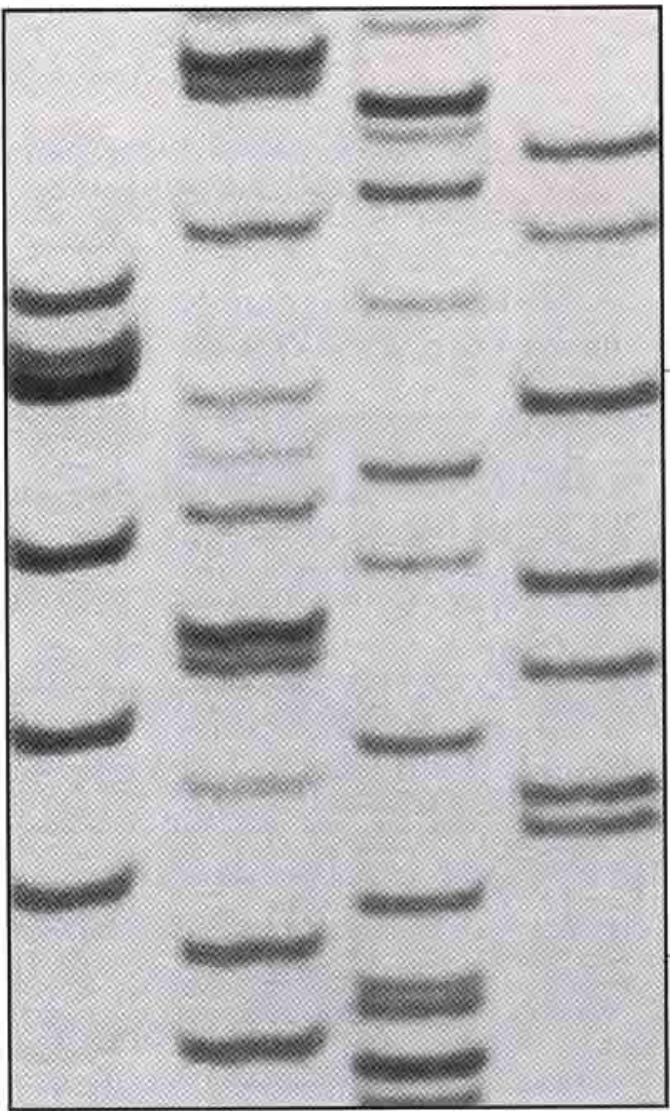
Polyacrylamide gel electrophoresis -- resolves of fragments differing by a single dNTP

- 'Slab' gels: as previously described
- Capillary gels:
  - narrow tubes filled with a gel matrix
  - only a tiny amount of sample needed
  - much faster than slab gels, best for “high-throughput” sequencing

Sequencing gel autoradiograph

Electrophoresis

A C G T



Chain terminator  
used

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



# Animation of cycle sequencing:

<https://dnalc.cshl.edu/resources/animations/cycseq.html>

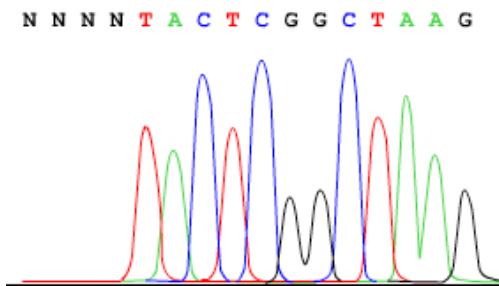
# Sequencing in a typical lab

It is rare for research labs to do their own large scale sequencing:

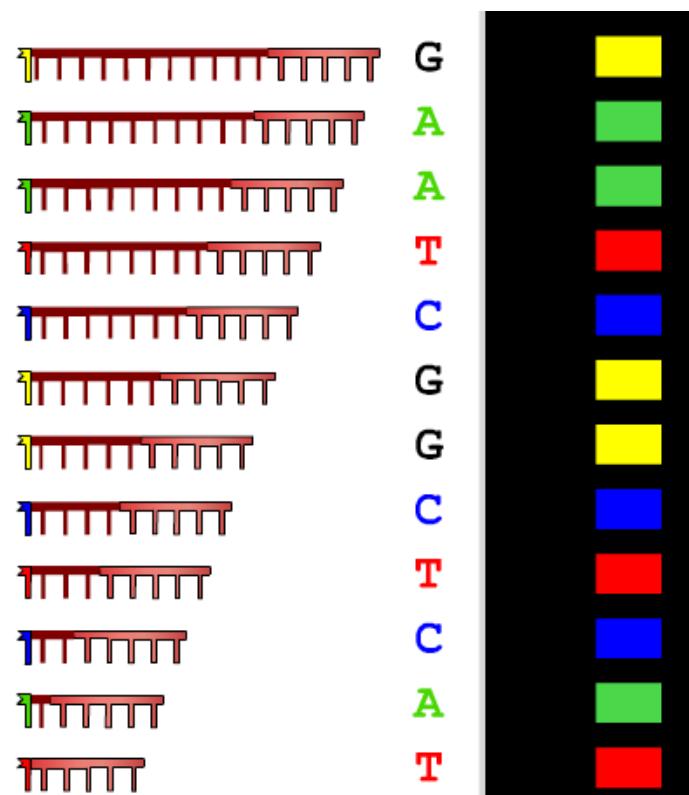
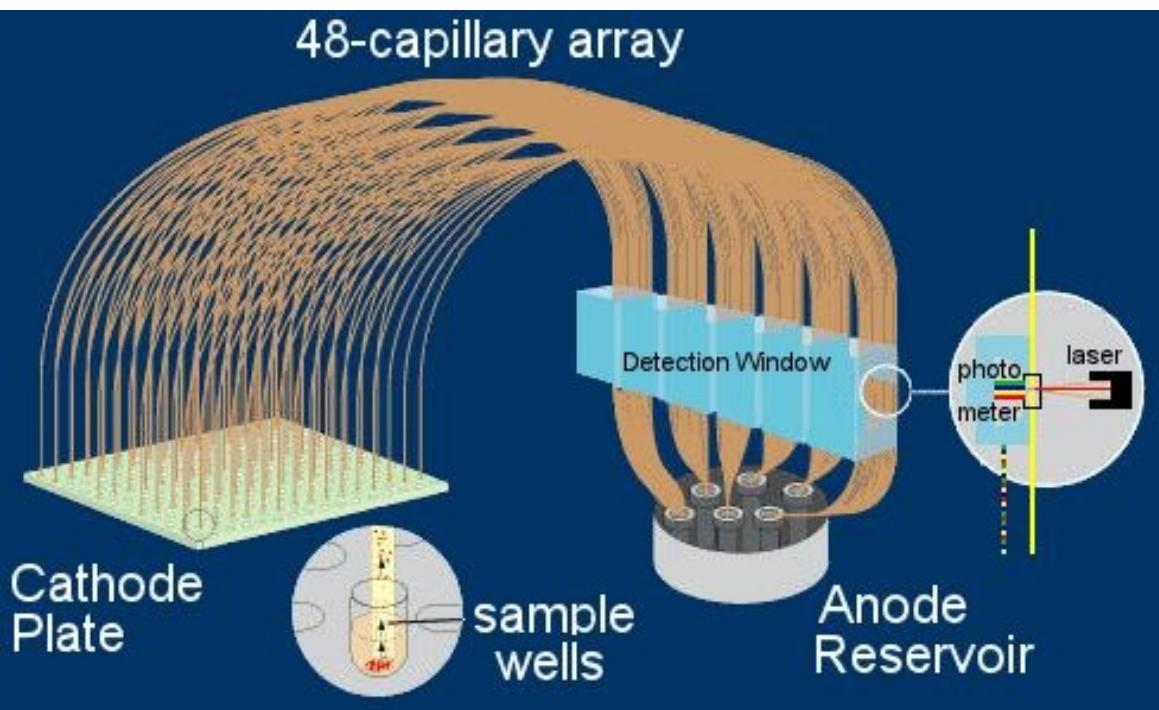
- costly equipment and materials
- time consuming protocols

Most labs send out for sequencing:

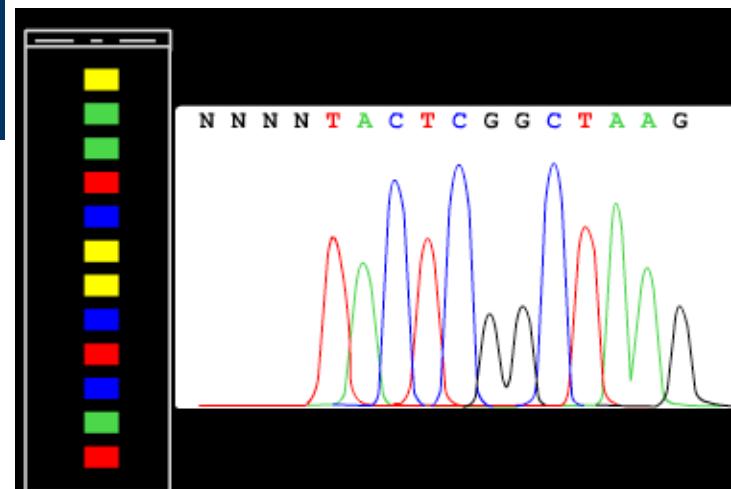
- You prepare the DNA (usually a plasmid or PCR product), supply the primer, company or university sequencing center does the rest
- The sequence is recorded by an automated sequencer as an “electropherogram”
- Viewable using ApE or other software



# An automated sequencer

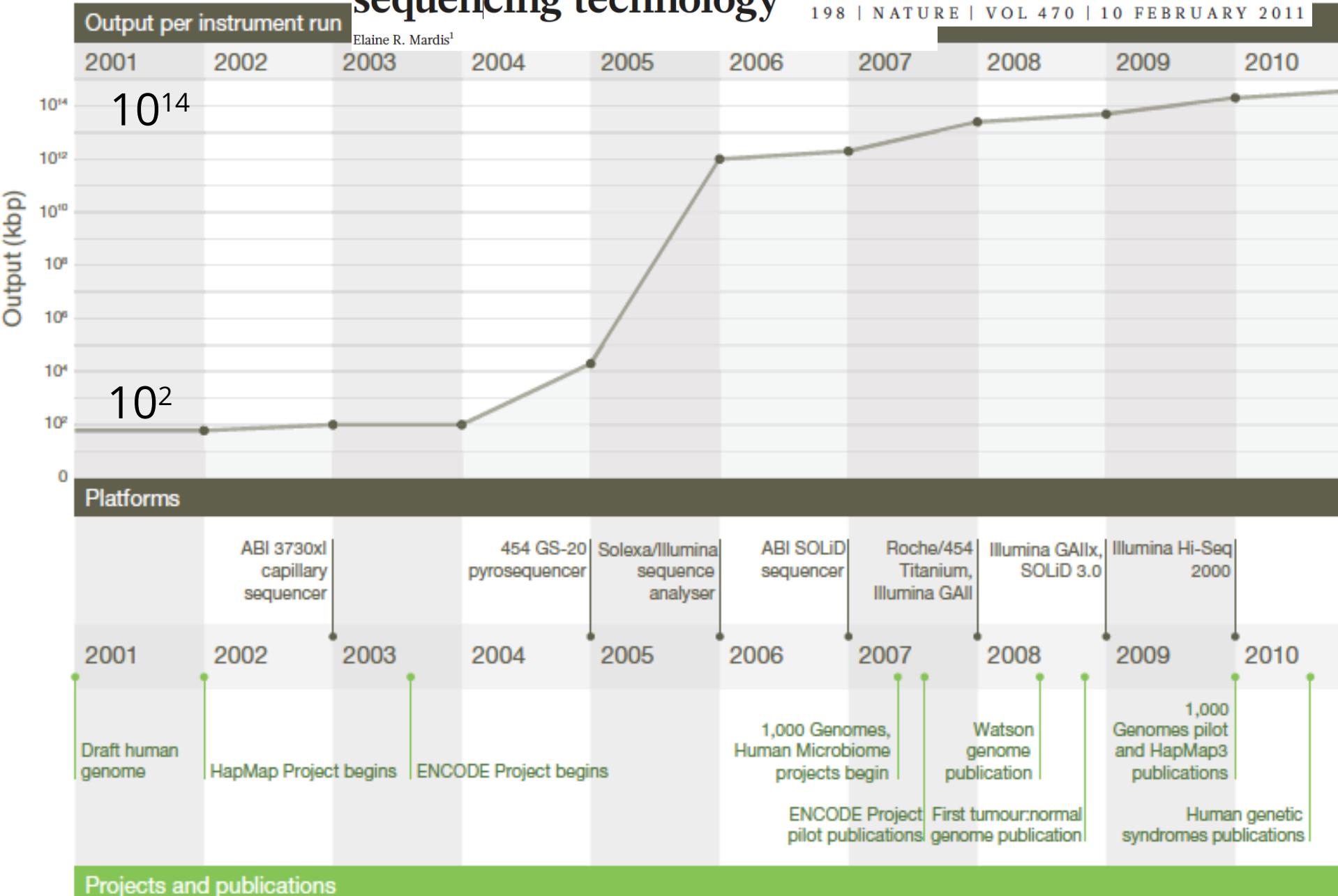


The data:  
electropherogram



# A decade's perspective on DNA sequencing technology

198 | NATURE | VOL 470 | 10 FEBRUARY 2011



## **Early sequencing: one DNA at a time**

Speed up by doing many DNA molecules at a time – arrays of sequencing reactions

## **Next generation sequencing: many reactions at once**

- 1) Pyrosequencing/ion torrent: dNTP addition detected by PPi chemistry or H<sup>+</sup> release
- 2) Sequencing by synthesis: fluor dye dNTPs are recorded over many rounds of sequencing
- 3) Ligation-mediated sequencing: short oligos are ligated to primers, which ligate in a sequence-dependent way
- 4) Pore sequencing: DNA through pore, record each base

# **“ pyrosequencing”**

Cut a genome to DNA fragments of 300 - 500 bp

Add adapters (short DNA handles) by ligation

Immobilize single strands on a very small bead (one piece of DNA per bead)

Amplify the DNA on each bead to cover each bead to boost the signal (error may creep in at this step)

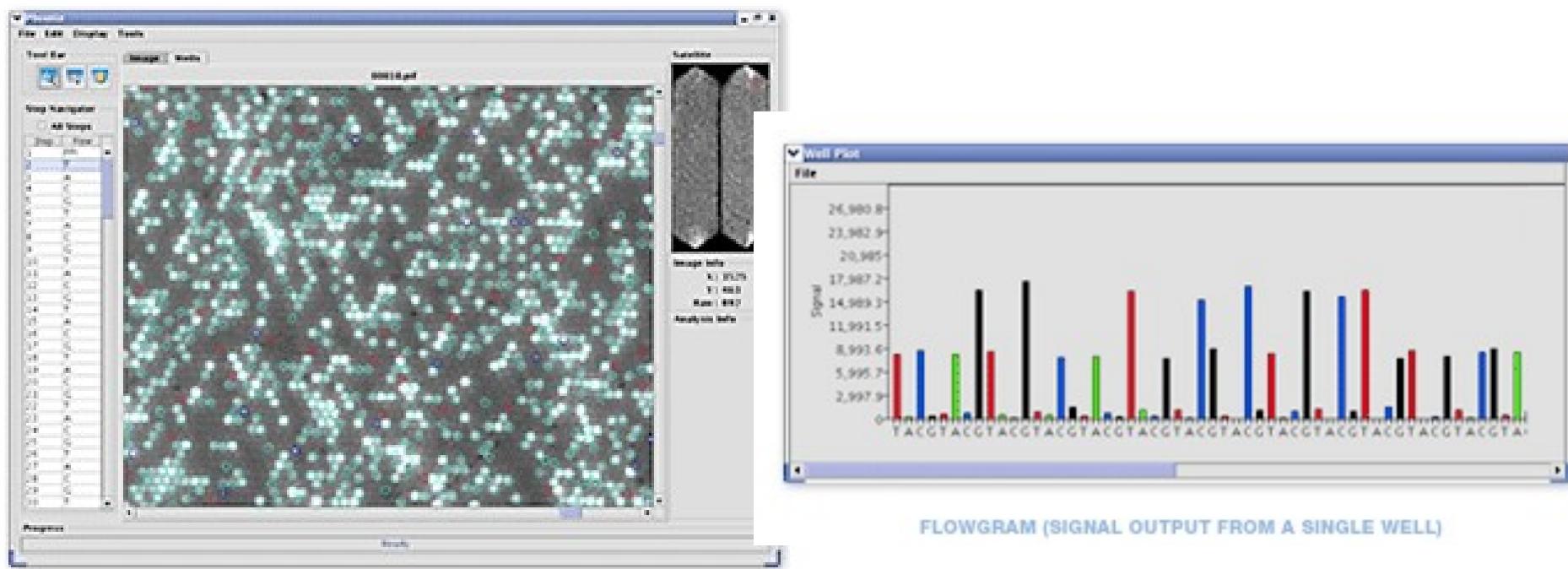
Separate each bead on a plate with up to 1.6 million wells

Sequence by primer extension. SHORT READS (50-150 bp)

Sequence by DNA polymerase-dependent chain extension, one base at a time in the presence of a reporter (luciferase)

Luciferase is an enzyme that will emit a photon of light in response to the pyrophosphate (PPi) released (and then added to Adenosine phosphosulfate) upon nucleotide addition by DNA polymerase

Flashes of light and their intensity are recorded

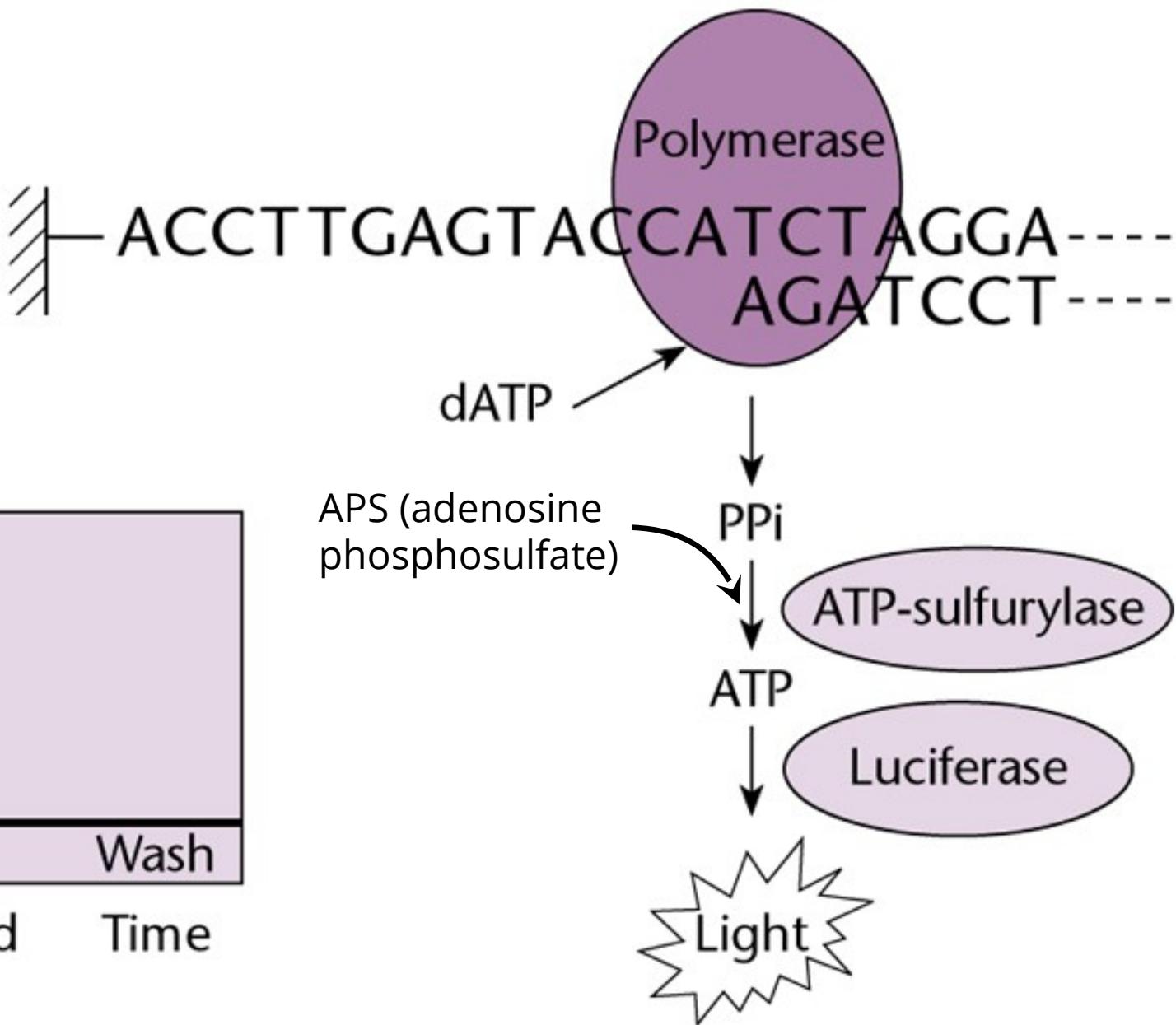
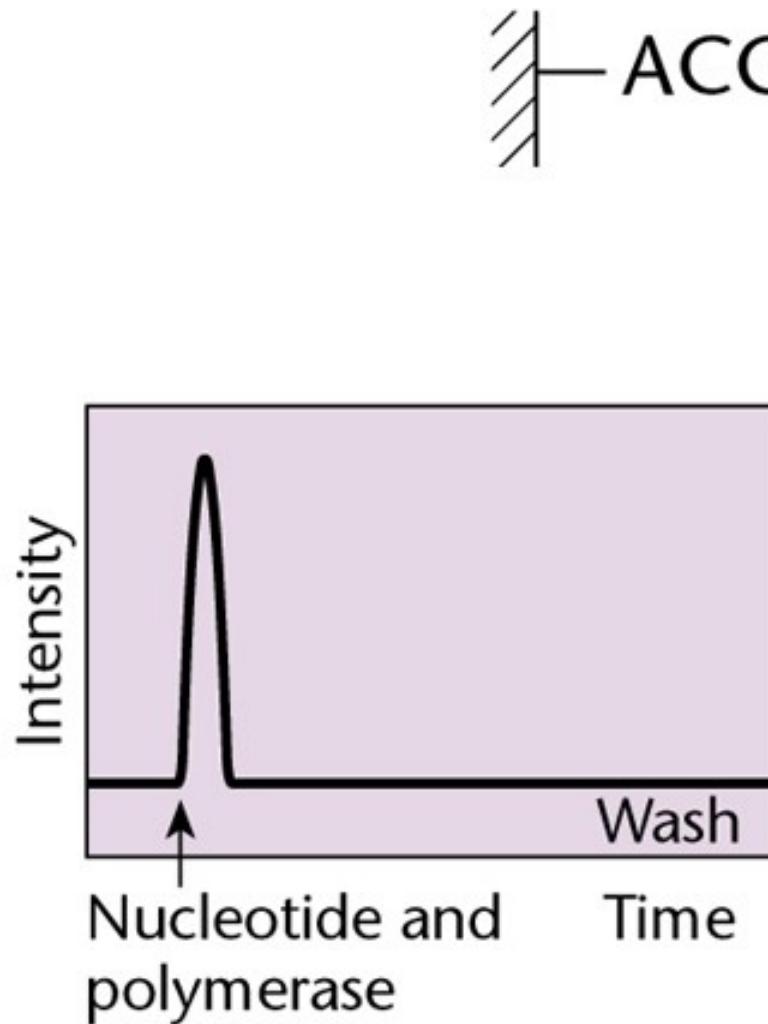


## DATA ANALYSIS: OUTPUT PACKAGE

Read length: about 200 bp

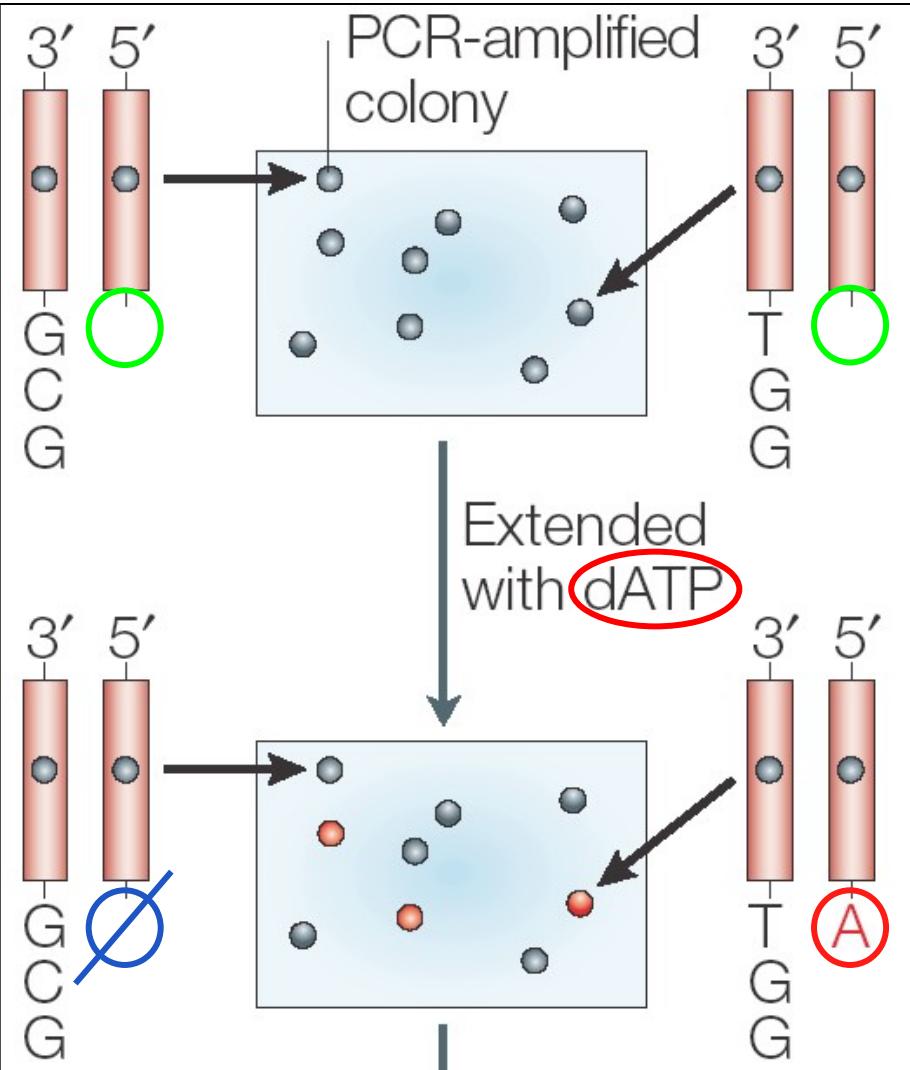
# How the extension reaction is detected (version A)

(a)



# Extension with individual dNTPs gives a readout

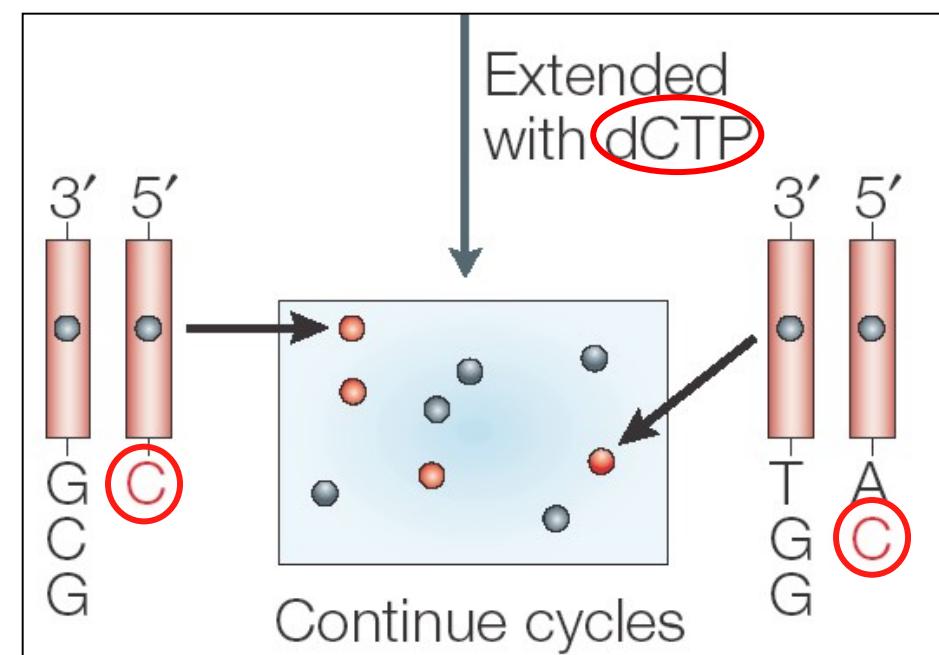
A



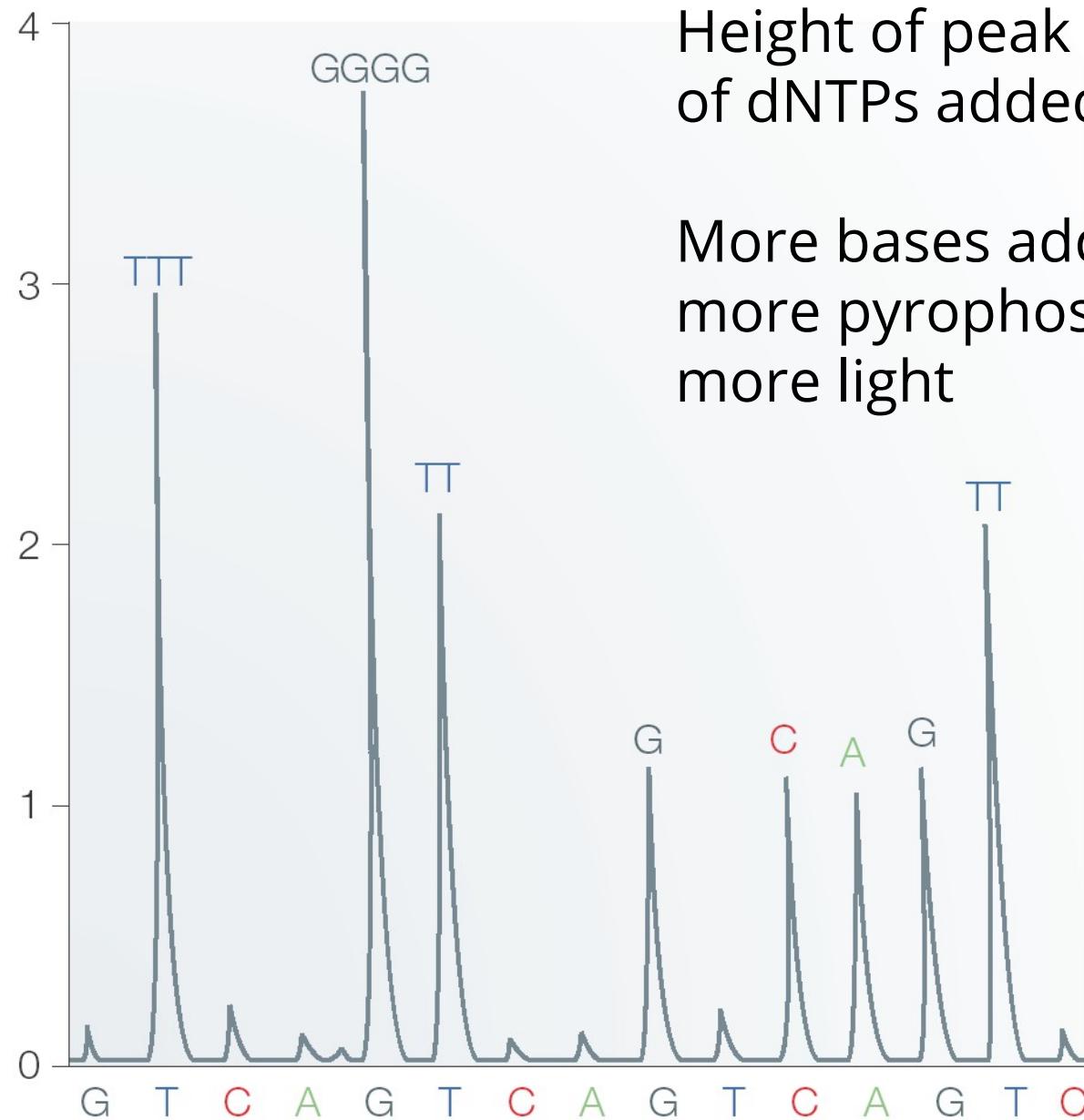
B

The readout is recorded by a detector that measures position of light flashes and intensity of light flashes

A



B



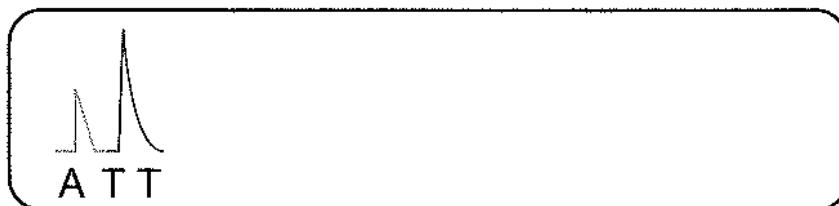
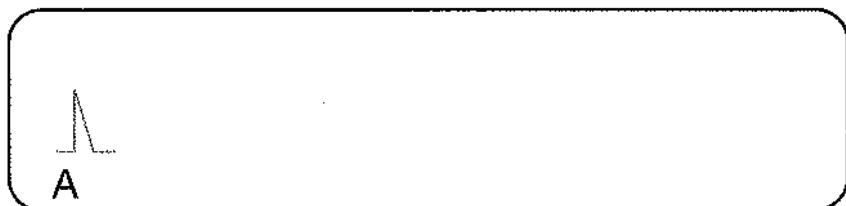
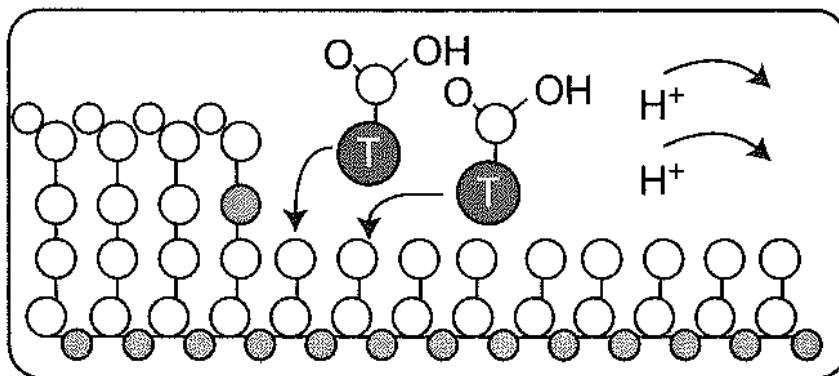
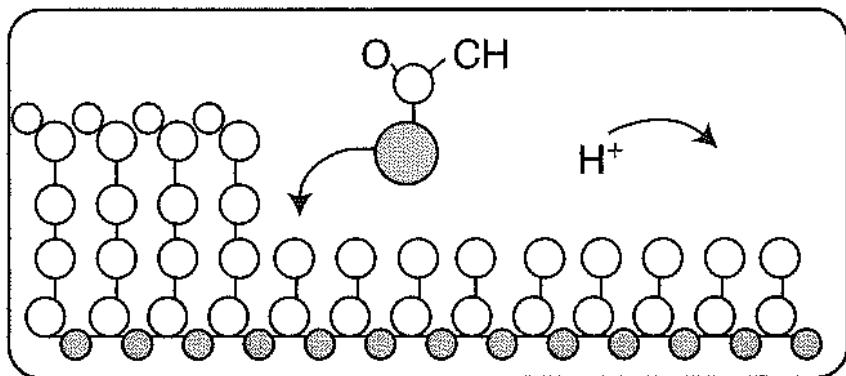
Height of peak indicates the number of dNTPs added

More bases added =  
more pyrophosphate =  
more light

This sequence: TTTGGGGTTGCAAGTT

Alternatively: detect pH change rather than light for each synthesis step

C Sequencing and base calling



## **Early sequencing: one DNA at a time**

Speed up by doing many DNA molecules at a time – arrays of sequencing reactions

## **Next generation sequencing: many reactions at once**

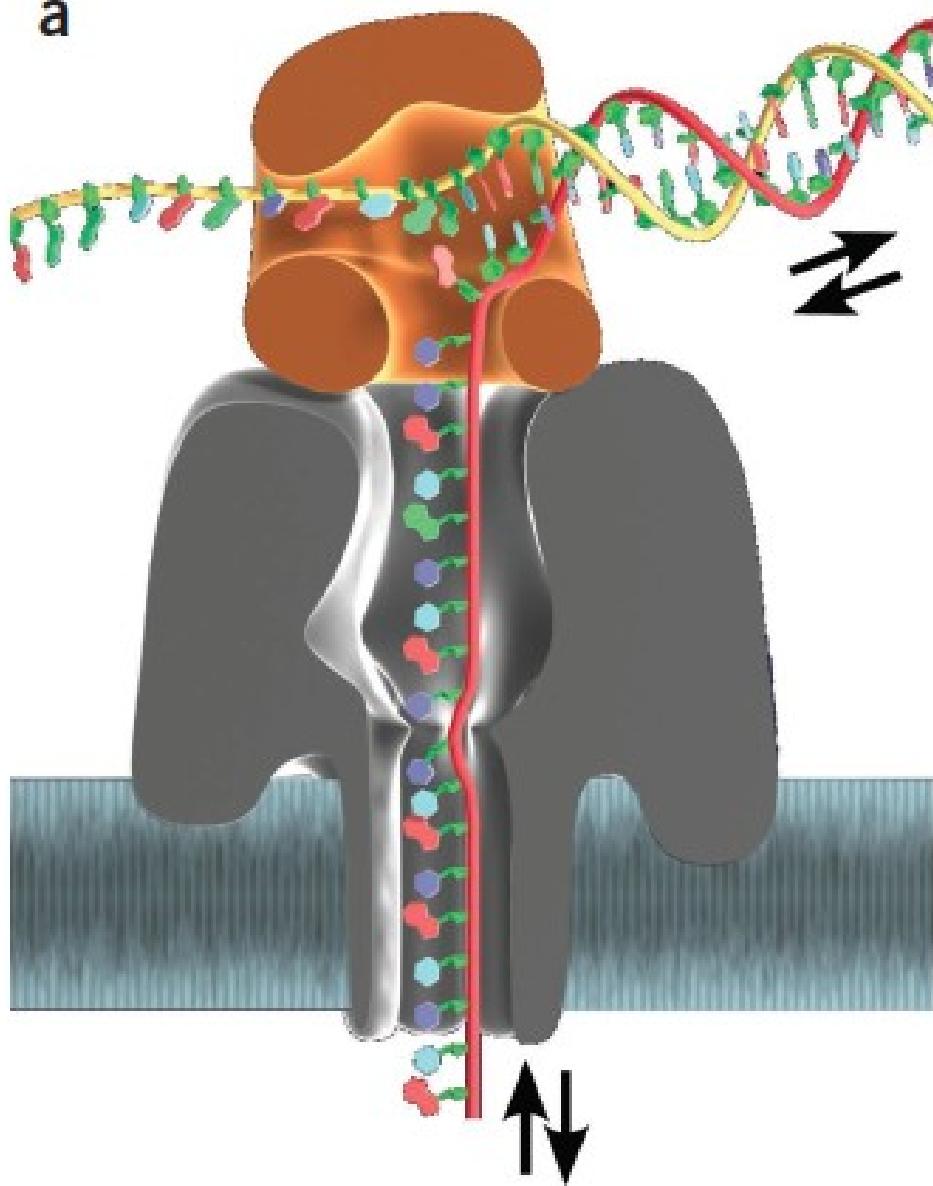
- 1) Pyrosequencing/ion torrent: dNTP addition detected by PPi chemistry or H<sup>+</sup> release
- 2) Sequencing by synthesis: fluor dye dNTPs are recorded over many rounds of sequencing
- 3) Ligation-mediated sequencing: short oligos are ligated to primers, which ligate in a sequence-dependent way
- 4) Pore sequencing: DNA through pore, record each base

**Nanopore sequencing:** controlled passage of DNA strands through pores.

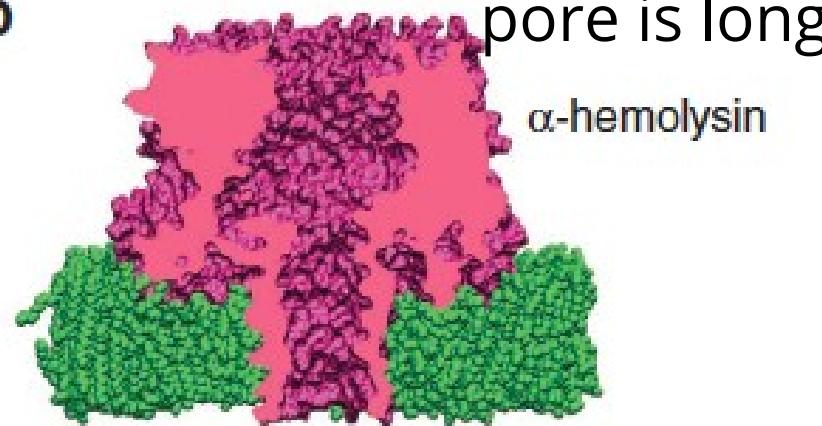
- 1) alpha hemolysin pore, through which ions can move (see movie)
  - thread ssDNA through a pore electrophoretically, remove “blocking oligo”
  - Phi29 DNA pol extends primer, drawing DNA through pore
  - Base passage through pore affects ion current amplitudes
- 2) mutated MspA pore
  - Same DNA polymerase approach
  - Shorter pore, better ion current data?

# DNA polymerase assisted translocation

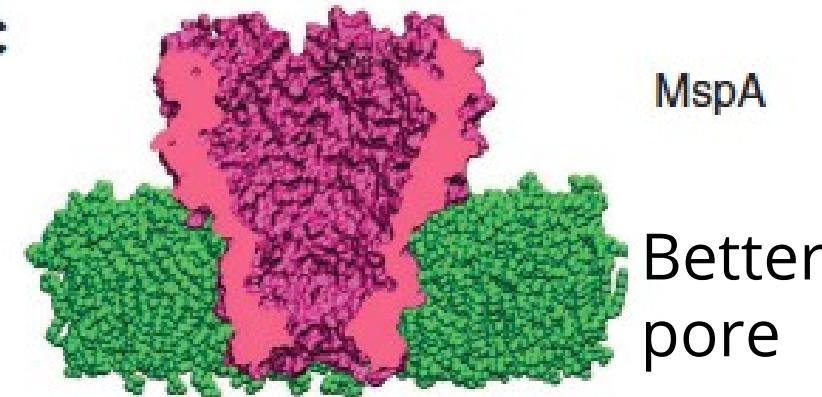
a



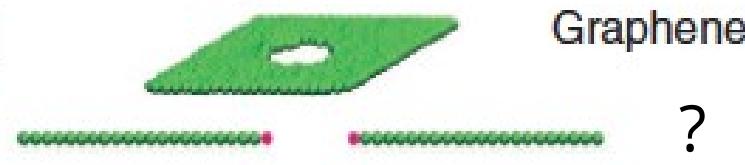
b



c



d



# Nanopore devices

- Oxford Nanopores: USB drive version of a nanopore sequencer in 2012
- Inaccuracies in base-calling, but multiple reads of the same sequence is helpful, and software for base calling is improving
- Very high speed sequencing, so it could be useful for speedy diagnosis in clinic: e.g. ID infectious agent to help give best treatment

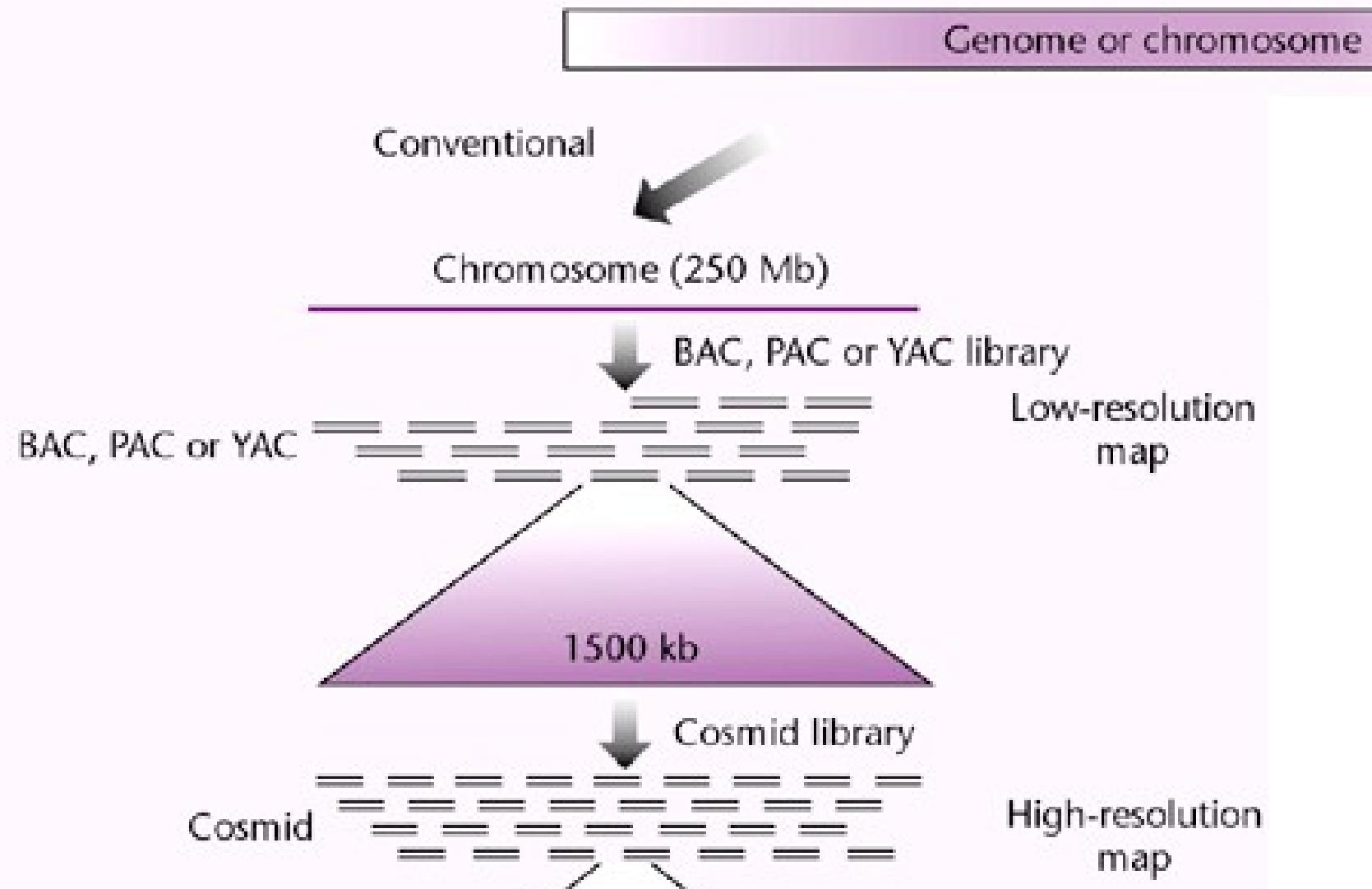


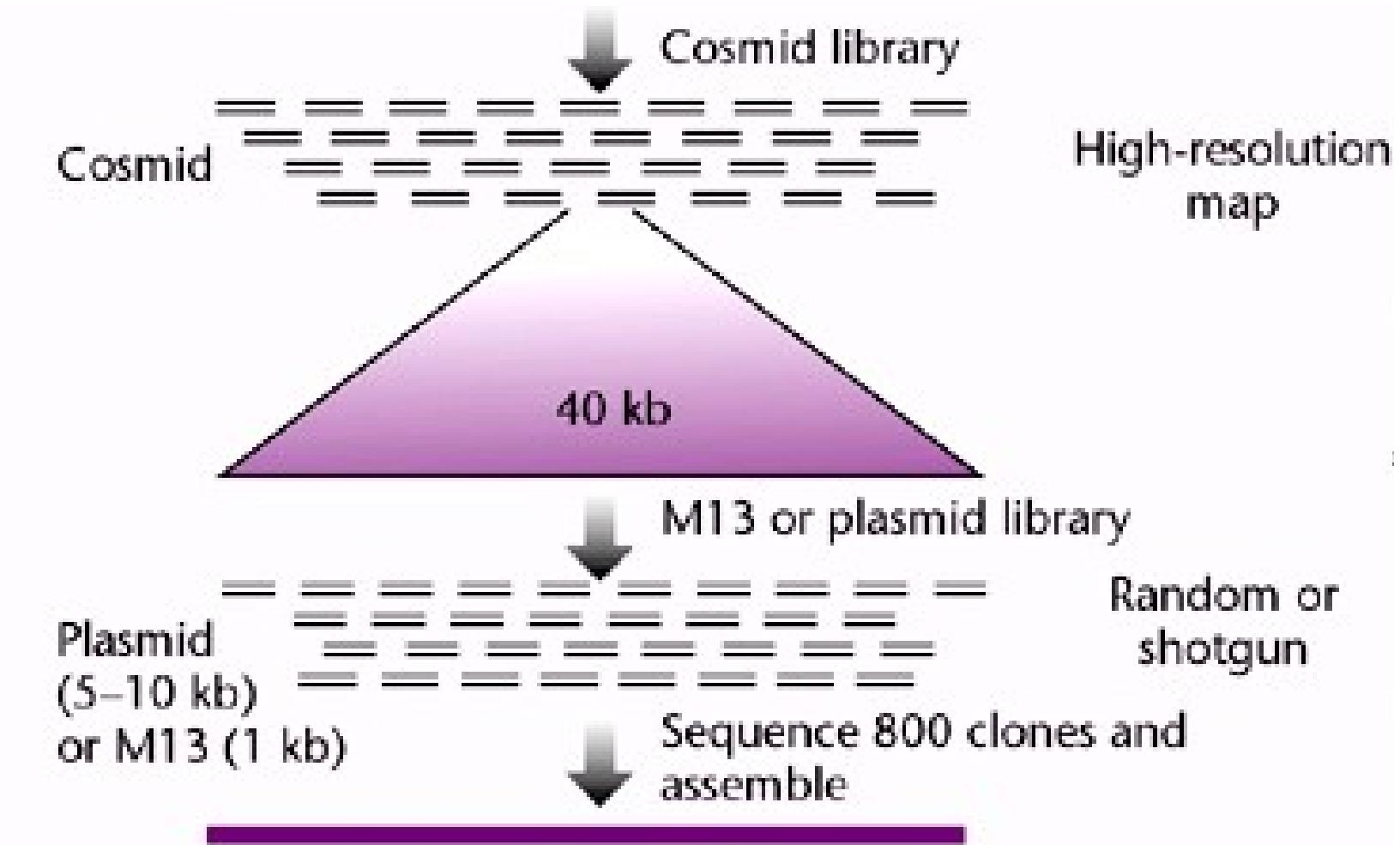
Movies

<https://youtu.be/GUb1TZvMWsw> , <https://youtu.be/hs0FdiTHMbc>

## Whole genome sequences:

Break up the genome, sequence pieces, re-assemble genome





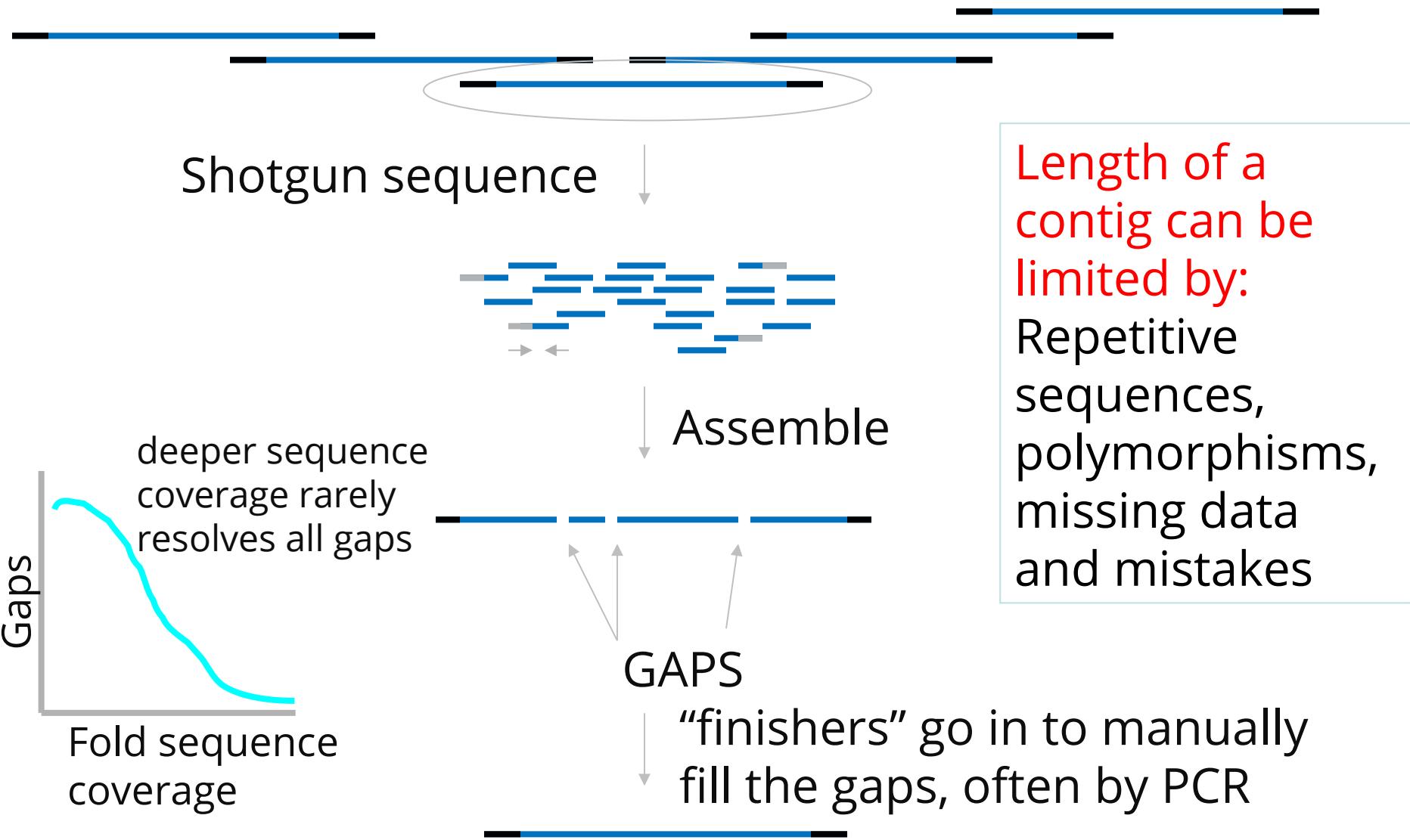
Sequencing is easy, mapping/assembly is more difficult

# Sequencing large pieces of DNA: “ shotgun” method

- Break DNA into small pieces (around 1000 base pairs), clone into a vector
- Sequence enough clones to ensure complete coverage (eg. sequencing a 3 million base pair genome would require **5x** to **10x** 3 million base pairs to have a reliable representation of the genome)
- Assemble genome through overlap analysis using computer algorithms and other methods. These **contiguous sequences blocks** are called '**contigs**'

## Clone based assemblies

— BAC insert  
— BAC vector



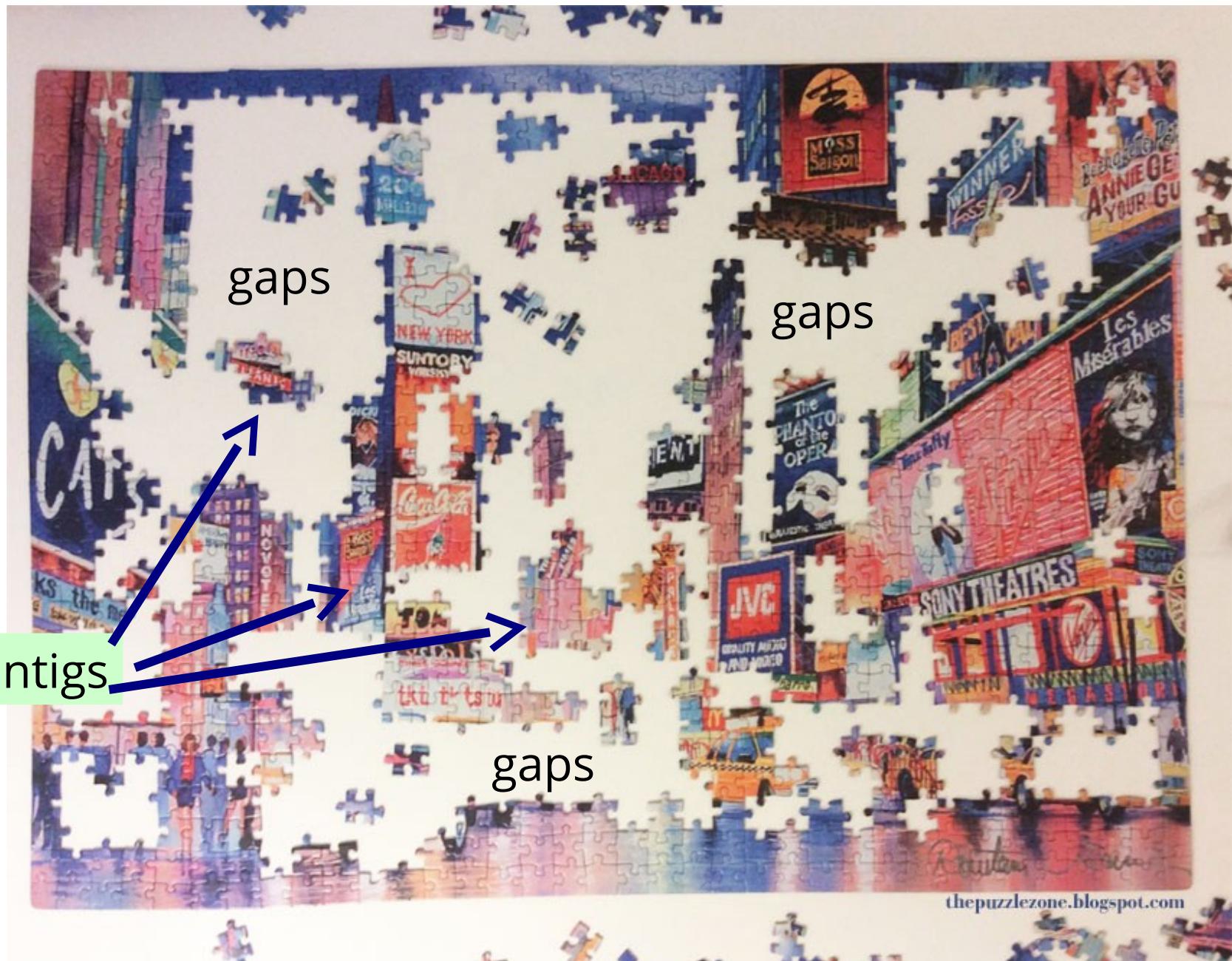
# Vocabulary

Contig: a sequence constructed from smaller, overlapping sequence, that contains no gaps

Typically build a contig from new reads, but also can include sequences found in GenBank/EMBL/DDBJ

Scaffold: a sequence constructed from smaller sequences which may contain gaps.

# Jigsaw puzzle / genome assembly



# Whole genomes are a challenge for 'next gen' sequencing

- Lots of sequencing reads, but short sequences, which requires much larger computational capacity for assembly

example: the human genome puzzle

- **Sanger (ddNTP) sequencing:**

- up to 1000 base pair reads of DNA sequence
- Need: ~30 million pieces, & ~8 copies of each piece (to account for errors)

- **Next gen sequencing:**

- about 100 base pair reads of DNA sequence
- Need: ~2 billion pieces, & ~100 copies of each piece

It is difficult to assemble whole genomes with next gen. technology. *Often used for 'resequencing'*

## 2010: giant panda genome sequencing

- This was the first high quality *de novo* genome sequence done using “next generation” sequencing
- 73-fold total coverage of the genome
- 2.4 Gigabase assembly (~94% of genome)
- The “contig N50” was 40 kilobases (50% of genome is found in contigs of 40 kilobases or greater length), typical for ‘finished’ genomes is 20-100 kb
- The genome assembly had more than 3,800 scaffolds (separated by gaps), this is quite high (by comparison, the dog genome has less than 100)
- (see perspective by Worley and Gibbs, 2010)

# Rapid genome sequencing in 2020: nCov-2

- Viral RNA isolated from bronchioalveolar lavage fluid (BALF)
- RNA was reverse transcribed to cDNA
- cDNA was fragmented, adaptors ligated, and amplified by PCR
- DNA was denatured, and single stranded DNA ligated to form circles, which were then amplified to make nanoarrays aka nanoballs (lots of copies of the same sequence)
- Each nanoball is put on a solid support in an array
- Nanoball spots sequenced by a method called combinatorial probe anchor ligation (cPAL)
- Allows sequence of 62-70 bases per nanoball

# combinatorial probe anchor ligation (cPAL)

Reading bases 1-5, e.g. position 5:



Common Probes  
(5th base set shown):

5 4 3 2 1	NNNN <del>A</del> NNNN
●	● NNNN <del>C</del> NNNN
●	● NNNN <del>G</del> NNNN
●	● NNNN <del>T</del> NNNN

# Methods for DNA sequencing

- A. Sanger dideoxy (primer extension/chain-termination) method: the original protocol for genome sequencing, adaptable, scalable to large sequencing projects
- A. Next generation sequencing: many reactions at the same time
- B. Sequencing a genome – break the DNA, sequence it, and put it back together

# The human genome: applications and implications

- 1) What is the human genome? History of the human genome sequencing project, & where we are now
- 2) Where is the human genome and how is it annotated?  
UCSC browser and 'tracks'
- 3) Genetic testing
- 4) Ethics and the genome

# Readings:

See links shown in class for more info

Also:

1. *DNA and insurance*: debate on the ethics of DNA testing by insurers
2. *Gene testing and anonymity*: debate on the overall value vs. risk of widespread genetic testing
3. *GINA and genomic medicine*: What effect has the Genetic Information Non-disclosure Act had on genetic testing and medicine?
4. *Genomics legal 2019*: genomics advances bring new legal challenges
5. *Genome injustice 2019*: Underrepresented groups push for more representation in genomics advances

# The human genome project: some milestones

- 1986:** "Human Genome Initiative" begins at US DOE
- 1992:** Complete, low resolution linkage map of genome
- 1995:** First complete genome (*Hemophilus influenzae*)
- 1999:** Human chromosome 22 finished
- 2000:** President Clinton announces completion of 'working draft' of human genome
- 2003:** Human genome project declared complete
- 2004:** Human gene count estimate: 20,000-25,000, function of more than half is unknown
- 2006:** Human chromosomes 1, 3, 8, 11, 12, 15, 17 completed
- 2010:** '1000 genomes consortium' to map human genetic variation

[https://web.ornl.gov/sci/techresources/Human\\_Genome/index.shtml](https://web.ornl.gov/sci/techresources/Human_Genome/index.shtml)

[http://web.ornl.gov/sci/techresources/Human\\_Genome/project/journals.shtml](http://web.ornl.gov/sci/techresources/Human_Genome/project/journals.shtml)

The world's largest collaborative biology project

# So whose genome was sequenced initially?

- Two groups worked to complete the genome assembly.
- The **publicly funded** group used DNA from an anonymized **group** of donors (two male, two female) from Buffalo, NY (where the DNA preparer was based)
- The **privately funded** group (Celera) used DNA from **five individuals** from an anonymized pool.
- Subsequently: the “1000 Genomes Project” included donors from diverse populations, for details see  
<http://www.internationalgenome.org/about>

<https://dnalc.cshl.edu/view/15327-The-public-Human-Genome-Project-s-DNA-donors-Eric-Lander.html>

Accessing the human genome: the UCSC browser

<http://genome.ucsc.edu/>

Current version: human reference sequence GRCh38  
(a.k.a. hg38) produced in December 2013, with updates  
frequently added. Most recent is GRCh38.p13, from  
2/28/19, although UCSC still uses .p12 from 12/21/17

[https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.39](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39)

User guide to UCSC browser:

<http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html>

Numerous helpful links there

# Statistics of hg38.p13 (2/28/19)

Number of regions with alternate loci or patches	358
Total sequence length	3,099,706,404
Total ungapped length	2,948,583,725
Gaps between scaffolds	349
Number of scaffolds	472
Scaffold N50	67,794,873
Scaffold L50	16
Number of contigs	998
Contig N50	57,879,411
Contig L50	18
Total number of chromosomes and plasmids	24
Number of component sequences (WGS or clone)	35,613

[https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.39](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39)

# Statistics of hg38.p12 (12/21/17)

Number of regions with alternate loci or patches	317
Total sequence length	3,257,319,537
Total assembly gap length	161,368,351
Gaps between scaffolds	349
Number of scaffolds	874
Scaffold N50	59,364,414
Scaffold L50	17
Number of contigs	1,535
Contig N50	56,413,054
Contig L50	19
Total number of chromosomes and plasmids	25
Number of component sequences (WGS or clone)	37,479

(from [https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.38](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38) )

# Getting to the data

Chromosome lengths

Total lengths

Ungapped lengths

N50s

Gaps

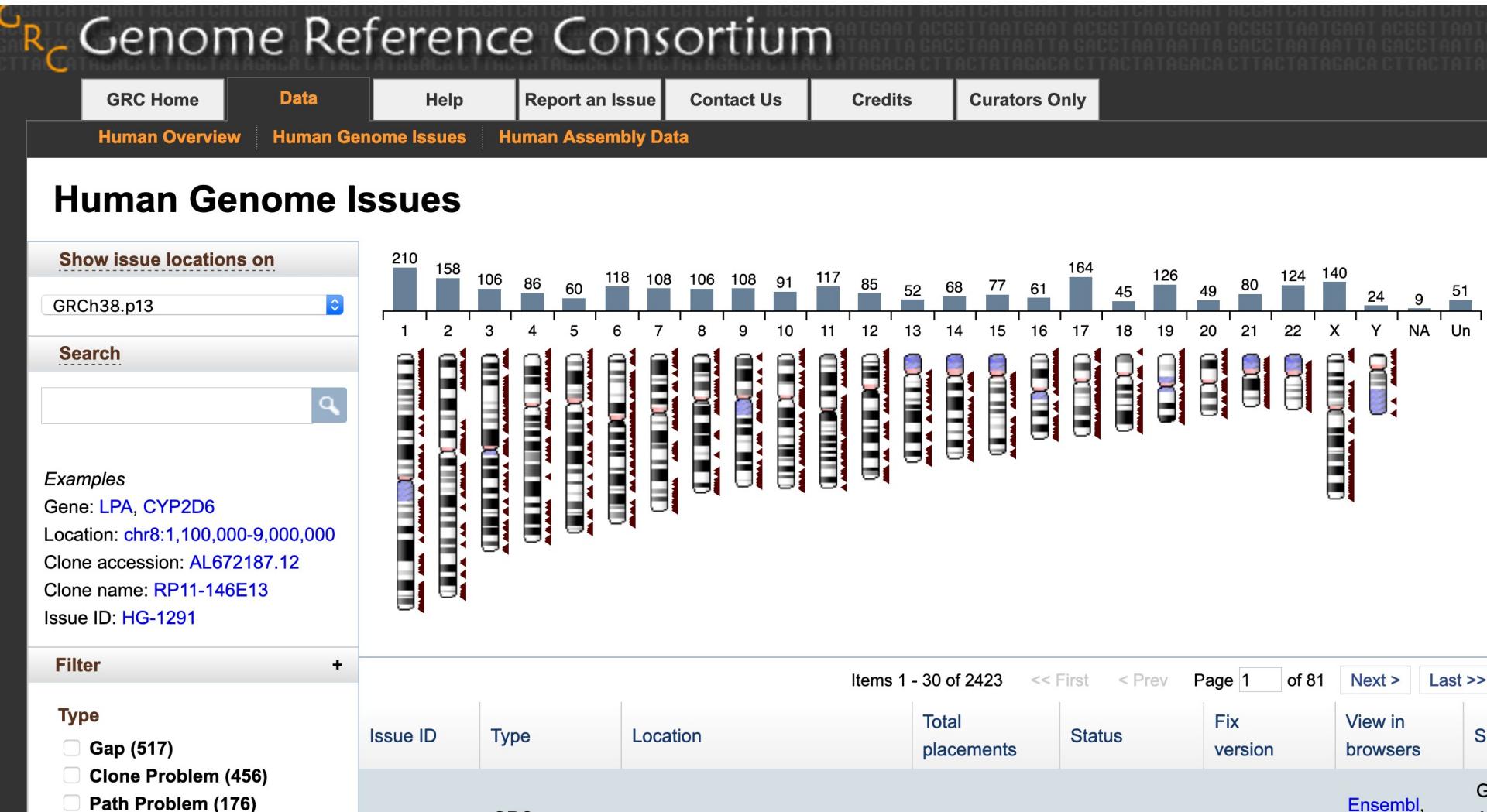
Counts

Chromosome lengths are calculated by summing the length of the placed scaffolds and estimated gaps.

Chromosome	Total length (bp)	GenBank accession	RefSeq accession
1	248,956,422	CM000663.2	NC_000001.11
2	242,193,529	CM000664.2	NC_000002.12
3	198,295,559	CM000665.2	NC_000003.12
4	190,214,555	CM000666.2	NC_000004.12
5	181,538,259	CM000667.2	NC_000005.10
6	170,805,979	CM000668.2	NC_000006.12
7	159,345,973	CM000669.2	NC_000007.14
8	145,138,636	CM000670.2	NC_000008.11
9	138,394,717	CM000671.2	NC_000009.12

A few issues remain to be resolved:

<https://www.ncbi.nlm.nih.gov/grc/human/issues>



# Human Genome Issues

Show issue locations on

GRCh38.p13

Search

Examples

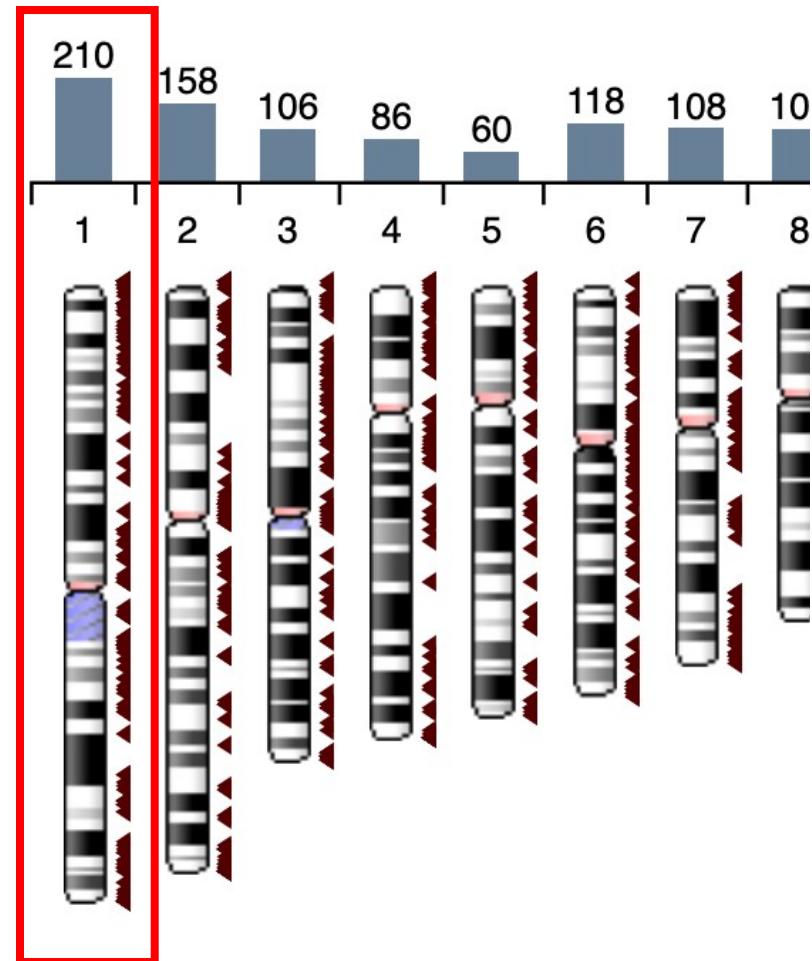
Gene: [LPA](#), [CYP2D6](#)

Location: [chr8:1,100,000-9,000,000](#)

Clone accession: [AL672187.12](#)

Clone name: [RP11-146E13](#)

Issue ID: [HG-1291](#)



Click on one of the chromosomes

# Human Genome Issues

## Show issue locations on

GRCh38.p13

## Search



## Examples

Gene: [LPA, CYP2D6](#)

Location: [chr8:1,100,000-9,000,000](#)

Clone accession: [AL672187.12](#)

Clone name: [RP11-146E13](#)

Issue ID: [HG-1291](#)

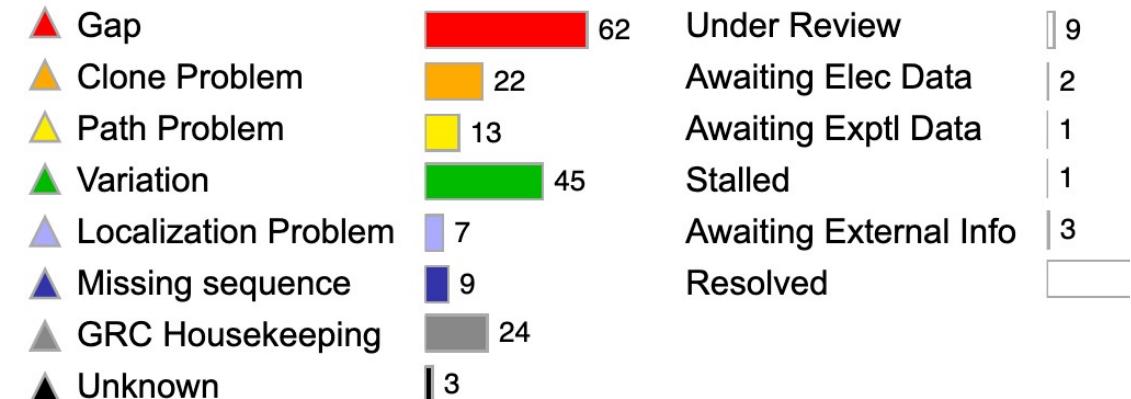
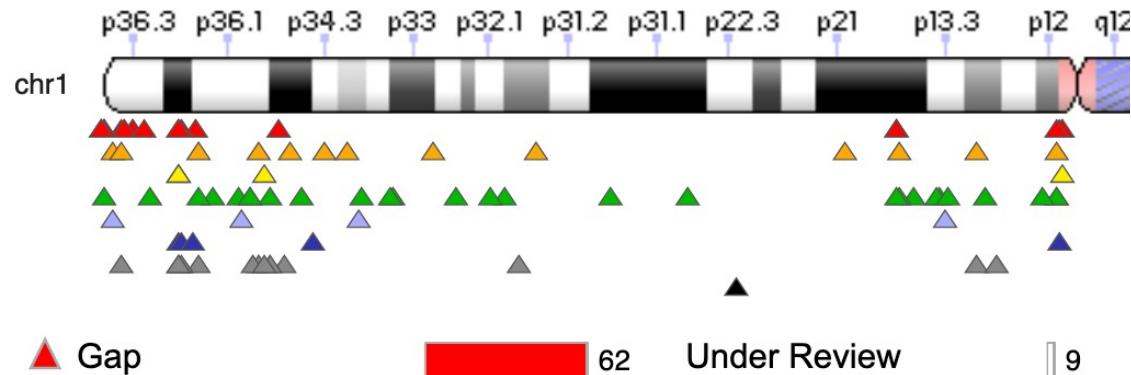
## Filter



### Type

- Gap (70)**
- Clone Problem (24)**
- Path Problem (14)**
- Variation (45)**
- Localization Problem (15)**

[More...](#)



Items 1 - 30 o

Issue ID	Type	Location	Total
----------	------	----------	-------

HG-2581 Variation chr1:19,218,148-19,370,277 1

What about this issue?

# Human Genome Issue HG-2581

Summary:	The Reference does not represent the coding allele for GeneID: 246181 ( <a href="#">AKR7L</a> ).
Description:	The Reference does not represent the coding allele for GeneID: 246181 ( <a href="#">AKR7L</a> ).
Status:	<a href="#">Resolved</a> (GRC Resolved- No Change)
Type:	<a href="#">Variation</a>
Last updated:	2020-10-07
Affects version:	GRCh38
Fix version:	<a href="#">GRCh39</a>
Resolution:	The mismatch needed to represent a coding allele for AKR7L (rs190747734) falls below allele frequency of .05, making it a rare allele.



## Patches and alternate loci

No patches or alts are associated with HG-2581.

## Find a gene: TFIIB, CFTR, GCDH, or ACE2

<http://genome.ucsc.edu/cgi-bin/hgGateway>

- Choose the assembly you want (use the latest release, HG38)
- Type “TFIIB”, “CFTR”, “GCDH” or “ACE2” in search term bar
- Following the search, click on the first or second line
- For GCDH:
  - It's on chromosome 19, left arm
  - spans nearly 8,787 bp
  - 11 exons
  - disease association (turn on OMIM genes track in “Phenotype and literature” bar, and hit refresh button)

Tracks: choose to visualize annotations to the gene

Example: OMIM, are there disease associations?

OMIM links: description of gene and its disease linkage

# What are the mutations in GCDH associated with glutaricaciduria in Pennsylvania Amish?

- Click on dark green OMIM bar (once the “OMIM genes” track has been turned on)
- Choose the OMIM link 608801
- Under Molecular Genetics heading, "...a single mutation was found as the cause of glutaric acidemia in the Old Order Amish of Lancaster County, Pennsylvania (**A421V**; 608801.0002), Biery et al. (1996).
- <https://www.youtube.com/watch?v=N2ox8g4uQqc&feature=youtu.be>

## What am I looking at? What does it mean?

- Each track has its own description and options that can be changed
- The top 'track' is the Gencode Track
- For a description of how a track works:
  - Example: for Gencode, go to the "Genes and Gene Predictions" bar below
  - Expand it by clicking '+' if it isn't already
  - 'Gencode v36' should be in 'pack' mode
  - Click on the **Gencode v36** link
- Note that non-coding and splice variants are shown by default

# How many versions of the genome are there?

- Human to human variation in sequence

- The 1000 genomes initiative

- <http://www.internationalgenome.org/1000-genomes-browsers>

- Personal genome project (voluntary)

- <http://www.personalgenomes.org/>

- Variation within a person

- Accumulation of mutations with age?

- Mutations associated with disease, e.g. cancer genomes  
(Cancer Genome Atlas:

- <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>

- )

# Some applications of human genome sequences

*Personalized genomic medicine, family planning*

- Will I get a disease? What should I do if so
- Will my children suffer genetic disorders

*Medical research: genetic basis for disease and effective treatments*

*Genealogy and human lineages: how have human populations evolved and migrated?*

# Want your genome sequenced for science?

<http://www.personalgenomes.org/>

The mission of the Personal Genome Project is to encourage the development of personal genomics technology and practices that:

- are effective, informative, and responsible
- yield identifiable and improvable benefits at manageable levels of risk
- are broadly available for the good of the general public

Family member wanting to get genome sequenced for science: what if another family member prefers genetic privacy?

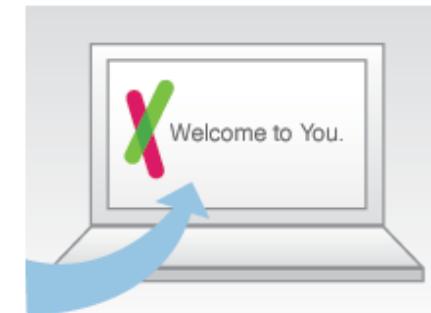
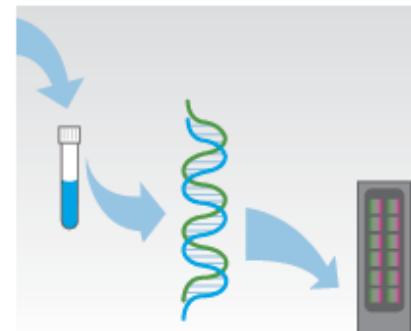
- Participants in the Personal Genome project

<https://www.youtube.com/watch?v=mVZI7NBgcWM>

# How to purchase your “personal genome”

Here's what you do:

PGS®



1. Order a kit from our [online store](#).

2. Register your kit, spit into the tube, and send it to the lab.

3. Our CLIA-certified lab analyzes your DNA in 6-8 weeks.

4. Log in and start exploring your genome.

What you get: a characterization of your “SNPs”, or “single nucleotide polymorphisms”. SNP chips are used.

There are differences in human genomes from one person to another, which can give information about ancestry. (Some of these changes correlate with disease states)

The SNP information can difficult to interpret. Effects of many mutations vary as a function of context.  
Also, the genome service will want to sell your information

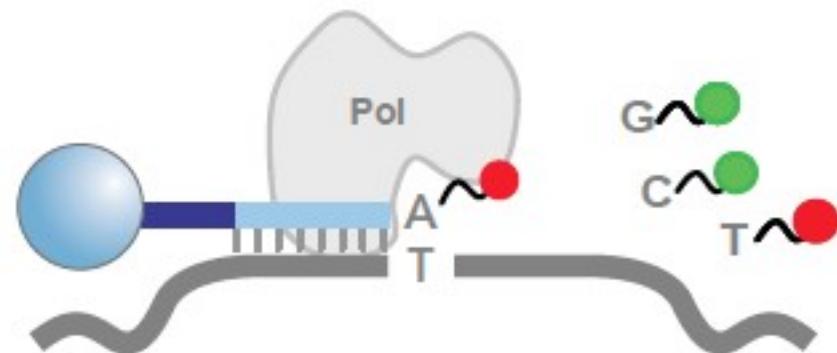
# How SNPs are detected

- 1) DNA is purified, then amplified (whole genome amplification)
- 2) DNA is fragmented and hybridized to probe DNA on array
- 3) Hybridized probe is extended with fluorophore-containing nucleotide



**Step 1. Selectivity**

Hybridization of unlabeled DNA fragment to 50mer probe on array



**Step 2. Specificity**

Enzymatic single base extension with labeled nucleotide

# Diagnostics and personal genome services

DIRECT-TO-CONSUMER GENETIC TESTS:  
“Misleading Test Results Are Further Complicated by Deceptive Marketing and Other Questionable Practices”

US Govt. Accountability Office (GAO) report (2010)

## Contradictory Risk Predictions for Prostate Cancer and Hypertension

Gender	Age	Condition	Company 1	Company 2	Company 3	Company 4
 Male	48	Prostate cancer	Average	Average	Below average	Above average
		Hypertension	Average	Below average	Above average	Not tested

Source: GAO.

# Personal genome services and the FDA

Prior to 2010, several companies marketed “personal genome services” (no doctor’s order required) for learning disease susceptibility

In 2010, the FDA notified 17 personal genome service companies that their services are essentially medical devices, and thus require review and approval

Since these tests are unlikely to accurately predict disease risk, most companies folded

23 and Me held out until 2013, when it received a warning letter from the FDA, and switched to “ancestry genetic report”

In 2017: 23 and me received approval to notify customers of genetic disease risks for 10 conditions

<http://www.latimes.com/business/la-fi-23andme-reports-20170414-htmlstory.html>

# Direct-to-consumer genetic tests and the FDA

1<sup>st</sup> FDA approval (2015): Bloom's Syndrome carrier test (23&Me)

- Carrier screening tests are medical devices, and classified as “Class II”, since higher risk, requiring greater regulatory controls to ensure device safety and effectiveness (example: condoms are Class II devices)
- Test doesn't require a licensed practitioner, but must include:
  - explanation of what results might mean to prospective parents
  - instructions for accessing a board-certified clinical molecular geneticist or equivalent
- Additional tests have been approved since then:
  - <https://www.fda.gov/medical-devices/vitro-diagnostics/direct-consumer-tests>
  - <http://www.latimes.com/business/la-fi-23andme-reports-20170414-htmlstory.html>

# What DNA tests can and can't tell you about ancestry

- <https://www.vox.com/videos/2019/4/16/18410869/dna-genetic-ancestry-tests>

# Genetic privacy: The Human Genome Project Ethical, Legal, and Social Issues (ELSI)

[http://www.ornl.gov/sci/techresources/Human\\_Genome/elsi/elsi.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/elsi/elsi.shtml)

and <https://www.genome.gov/Funded-Programs-Projects/ELSI-Research-Program-ethical-legal-social-implications>

In May 2008: GINA (Federal Genetic Information Non-discrimination Act) passed

(<http://www.ginahelp.org/GINAhelp.pdf> )

- Health insurance companies may not treat people differently based on genetic code
- Employers **cannot**
  - demand genetic tests
  - discriminate against who they hire or how much they pay on the basis of genetic information
  - disclose genetic information in their possession except under specific and specially controlled circumstances.

## Some issues with GINA

- Some kinds of insurance are not included in GINA, including disability, life, or long-term care insurances
- May clash with established state policies
- Doesn't specify regulations for “Personal Genome Services”

Two articles published in 2019 (PDFs on D2L)

How is the law responding to issues surrounding genetic testing?

And

How can human genome information be studied and used equitably?

## The Belmont Report (1979)

- Ethical Principles and Guidelines for the Protection of Human Subjects of Research
- Inspired in part by the ethical violations reported in the Tuskegee Syphilis Study  
(<https://www.cdc.gov/tuskegee/timeline.htm>)
  - Syphilis was left untreated in group of black men, to study progression of the disease
  - The men were willing participants, but were not informed of the study or its purpose
- Three fundamental ethical principles were defined in response to prevent further ethical failures

# Belmont Report Principles

- 1) The principle of Respect for Persons acknowledges the dignity and autonomy of individuals, and requires that people with diminished autonomy be provided special protection. **This principle requires that subjects give informed consent to participation in research**
- 2) The principle of Beneficence requires us to protect individuals by maximizing anticipated benefits and minimizing possible harms
- 3) The principle of Justice requires that we treat subjects fairly

So, research on human subjects requires adherence to these guiding principles:

*Consent, beneficence, and fair treatment*

These guidelines are clearly relevant when considering how human DNA sequences should be used in the clinic, in research, and elsewhere

# A study of human lineages: a 90-year old lock of hair from an indigenous Australian man yields complete genome sequence

- Finding: indigenous Australians are descendants of the first humans to leave Africa (other Asian populations came from a second migration)
- approval was given by representatives of indigenous group from the region where man would have lived
- What about other indigenous individuals?
- Other proposed studies have been severely restricted by indigenous Australians (ie. Consent not given)
- Archaeological specimen: is consent required? For how long?
- How must human body parts and specimens held in museums (like mummified remains) be treated?
- <https://www.nature.com/news/2011/110928/full/477522a.html>
- (SEE ALSO: Genome injustice 2019, posted on D2L)

# The human genome: applications and implications

- 1) What is the human genome? History of the human genome sequencing project, & where we are now
- 2) Where is the human genome and how is it annotated?  
UCSC browser and 'tracks'
- 3) Versions of the genome
- 4) Ethics and the genome

# Introduction to basic bioinformatics

- 1) Online databases
- 2) Making biological sense of DNA sequences:  
finding and predicting function of protein  
coding genes
- 3) Using NCBI
- 4) What is BLAST?
- 5) Using BLAST for sequence analysis

[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

## Guide to readings on bioinformatics:

1) 19 MC4 Bioinformatics

- Intro
- The UCSC Genome Browser
- Algorithms, portals and methods
- BLAST and ClustalW
- Motif finding

2) Margaret Dayhoff, pioneer of bioinformatics (2019)

3) The beginners guide to genome annotation (2012)

4) BLOSUM 62: How this homology search algorithm works.

5) 2019-nCoV: sequencing, and two perspectives articles

6) Web sites (referred to in the notes)

# Bioinformatics: storage and analysis of biological information

- Nucleotide and protein sequence
- Macromolecular structures
- Gene expression patterns
- Biochemical pathways
- Evolutionary relationships
- Sorting/visualizing large data sets

# Bioinformatics databases: repositories for biological information

## Primary sequence databases:

- NCBI/Genbank
- DNA Databank of Japan (DDBJ)
- European Molecular Biology Laboratory (EMBL)

## Annotated protein sequence databases:

- SWISS-PROT (most accurate annotation: structures, functions, protein families, with references)
- TrEMBL (most current, but not fully annotated)

## Protein structure: The Protein Databank (pdb.org)

Many other databases exist, e.g ENCODE, UCSC genomes, etc.

<https://academic.oup.com/nar/issue/49/D1>

# Genome sequencing projects

JGI: the Joint Genome Initiative at the US Dept. of Energy (DOE)

<https://genome.jgi.doe.gov/portal/>

# The genome annotation pipeline

1. Make sure the genome assembly is ready. N50 scaffold length should be at least the median size of a gene.
2. Find repeat sequences and mask them (mark as repeats)
3. Gene prediction:
  - Align known proteins and ESTs (Expressed Sequence Tags) to the sequence to identify exons. Then identify splice sites
  - Also use *ab initio* gene identification software
4. Gene identification:
  - Automated annotation: multiple gene finders are run, and the consensus is used
  - Alignment data can be used to improve results
  - Manual curation (evidence-based decisions)
5. Assess the quality of annotation
  - Number of protein domains
  - Agreement of annotation with RNA info (EST, seq)
6. Visualize, share, and *update* the annotation

# A genome annotation is continuously updated as new experiments are done

Many predicted genes have unknown functions

Many genome features are difficult to annotate

- Regulatory regions: promoters, DNA binding sites
- Non-coding RNAs
- Transposons
- Pseudogenes

“Like parenthood, annotation responsibilities do not end with birth. Incorrect and incomplete annotations poison every experiment that makes use of them.”

# Genome annotation in the news: nCoV-2, the virus causing COVID-19

## Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding

Roujian Lu\*, Xiang Zhao\*, Juan Li\*, Peihua Niu\*, Bo Yang\*, Honglong Wu\*, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, Yuhai Bi, Xuejun Ma, Faxian Zhan, Liang Wang, Tao Hu, Hong Zhou, Zhenhong Hu, Weimin Zhou, Li Zhao, Jing Chen, Yao Meng, Ji Wang, Yang Lin, Jianying Yuan, Zhihao Xie, Jinmin Ma, William J Liu, Dayan Wang, Wenbo Xu, Edward C Holmes, George F Gao, Guizhen Wu¶, Weijun Chen¶, Weifeng Shi¶, Wenjie Tan¶

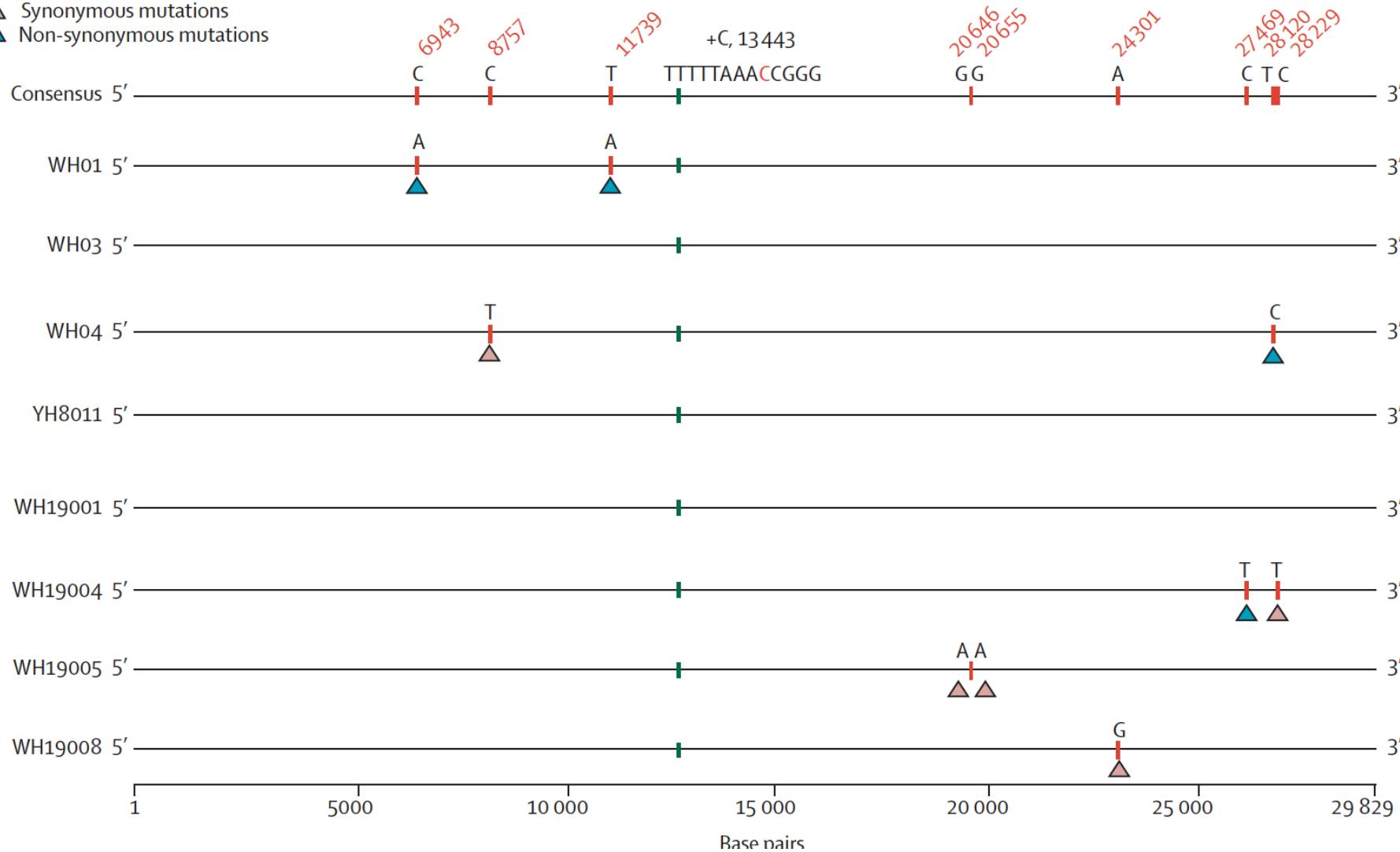
www.thelancet.com Vol 395 February 22, 2020

- Sequences of samples from 9 patients, several of whom had onset of symptoms in late December 2019
- Sequences were made freely available in January 2020

# 2019 nCoV-2 sequences: nearly identical (99.98%)

A

- ▲ Synonymous mutations
- ▲ Non-synonymous mutations



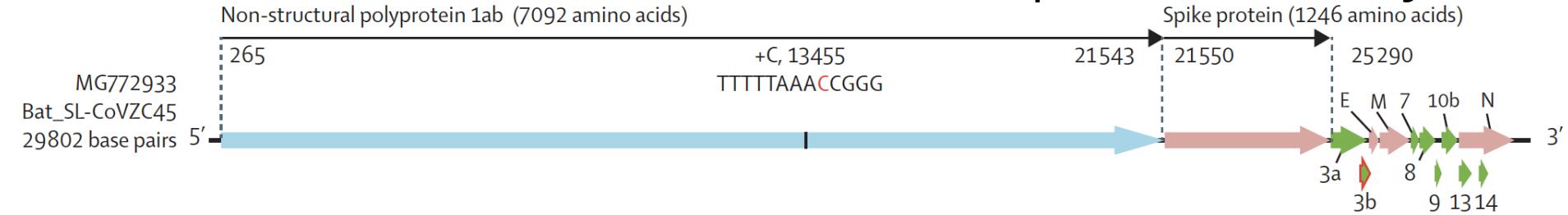
- therefore the virus has only made a recent entry into human populations

The viral genes were identified and the sequences compared with other coronaviral genomes

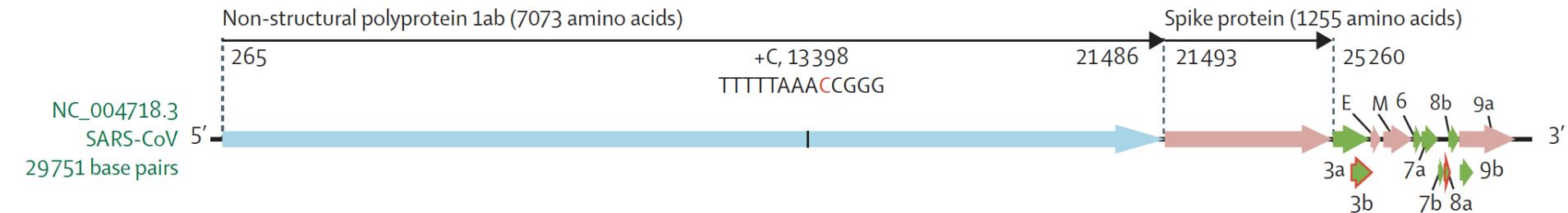
## 2019-nCoV genetic map



## Bat SARS-like Beta coronavirus (88% sequence similarity)

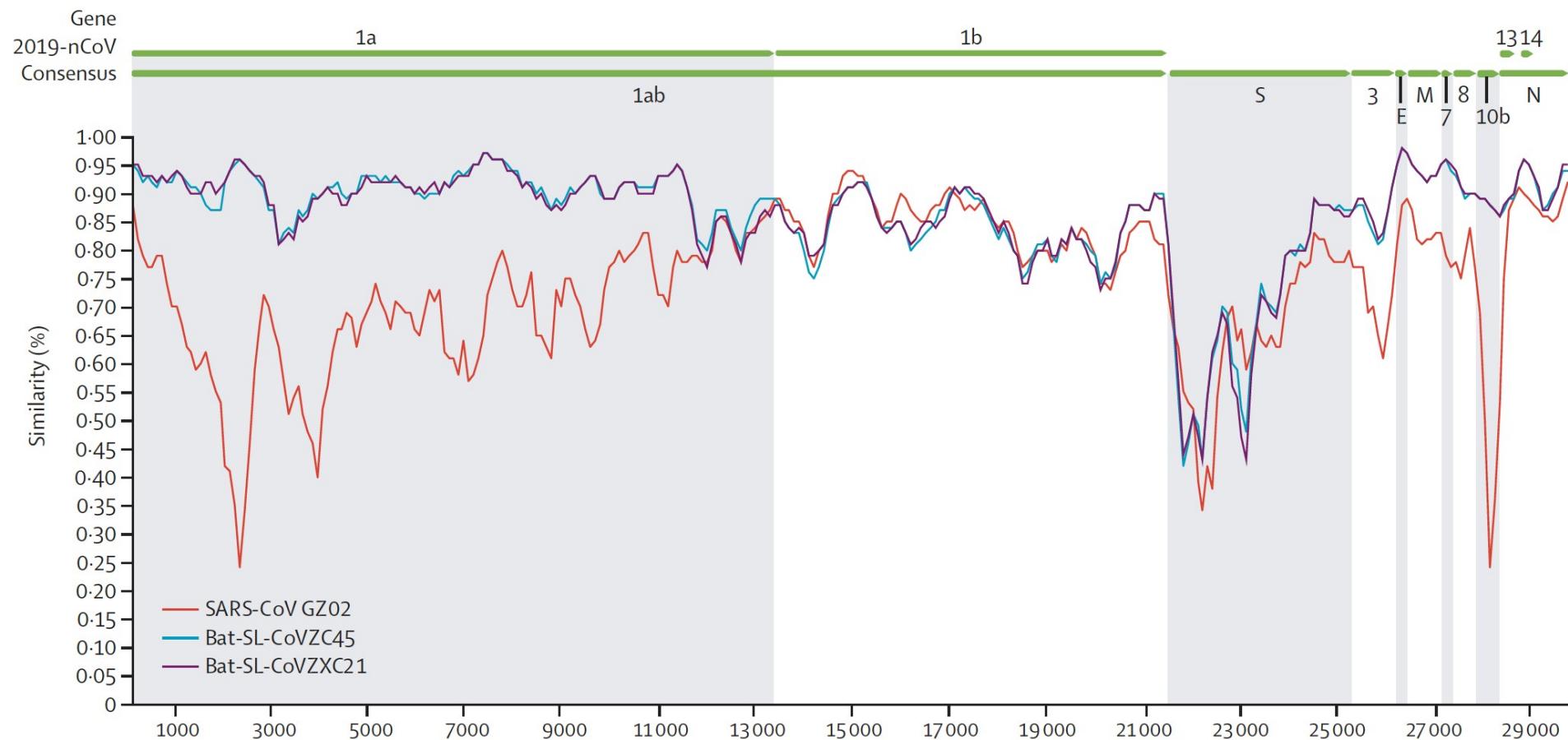


## SARS-CoV (79% sequence similarity)



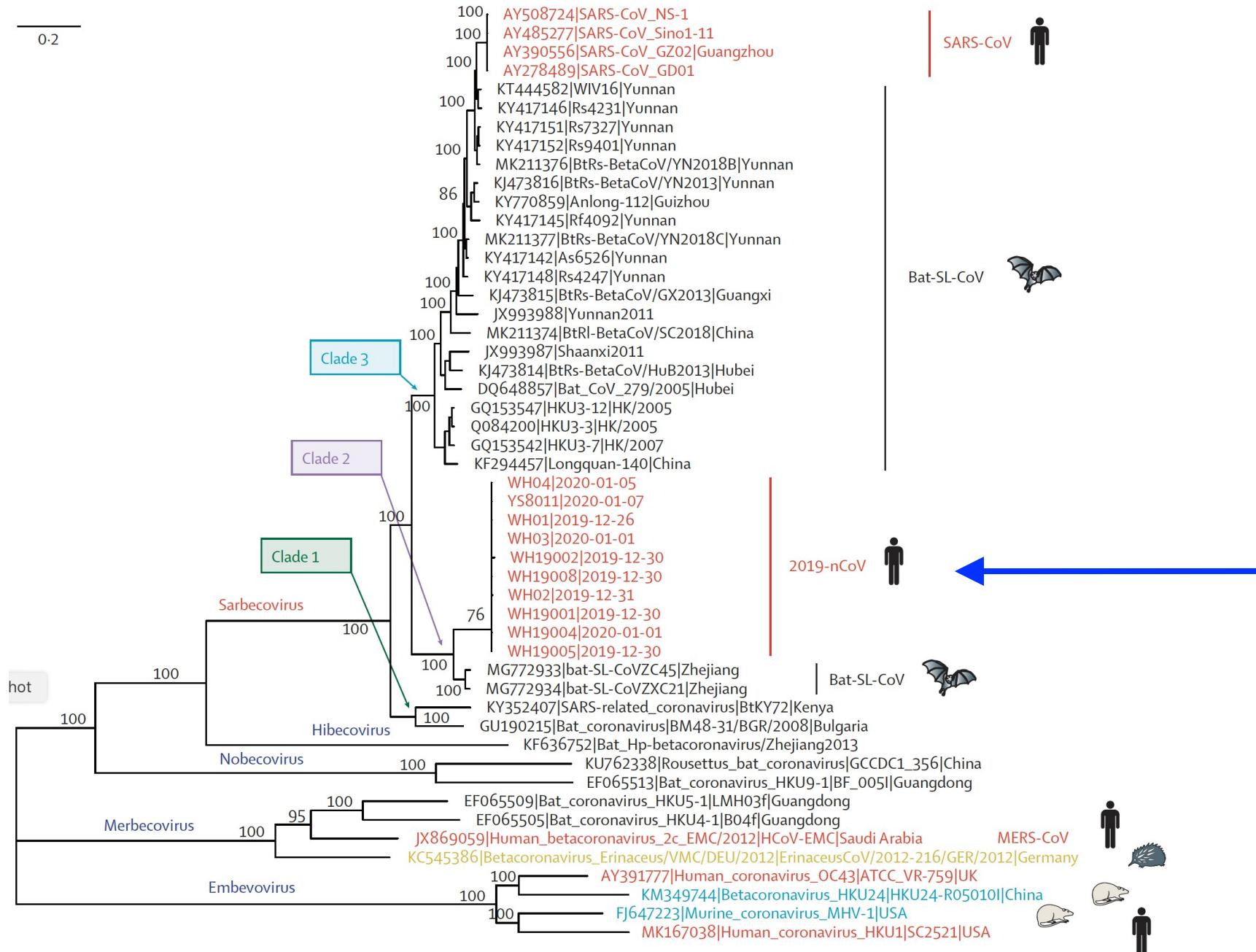
# Sequence similarity of related viruses to 2019-nCov

B

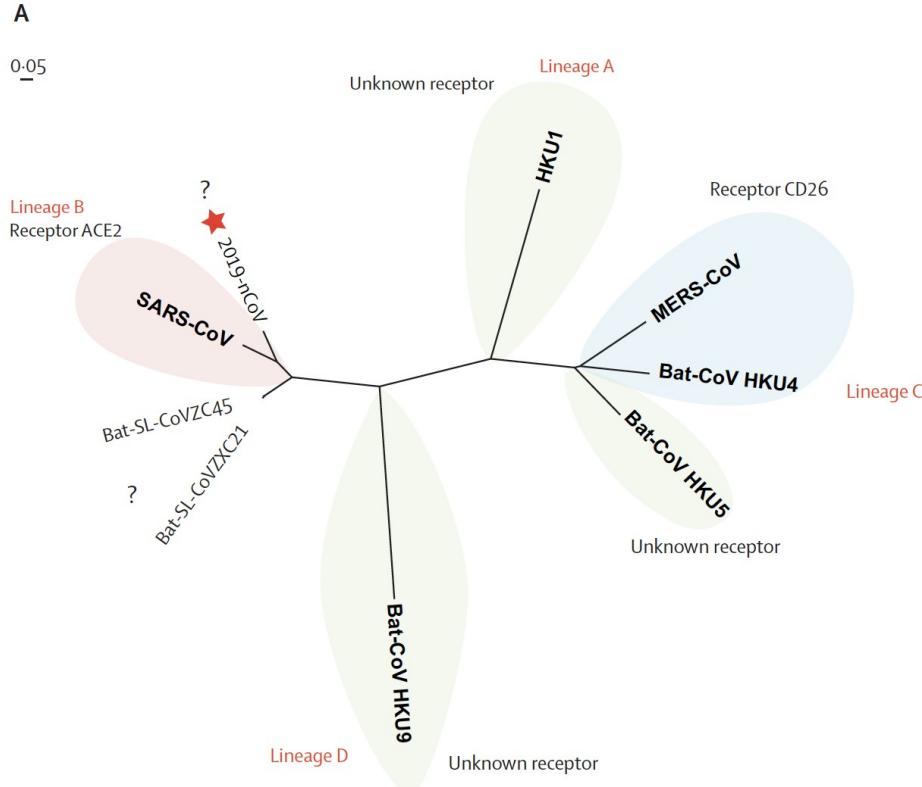


The lowest similarity with Bat CoV is in 'spike' protein

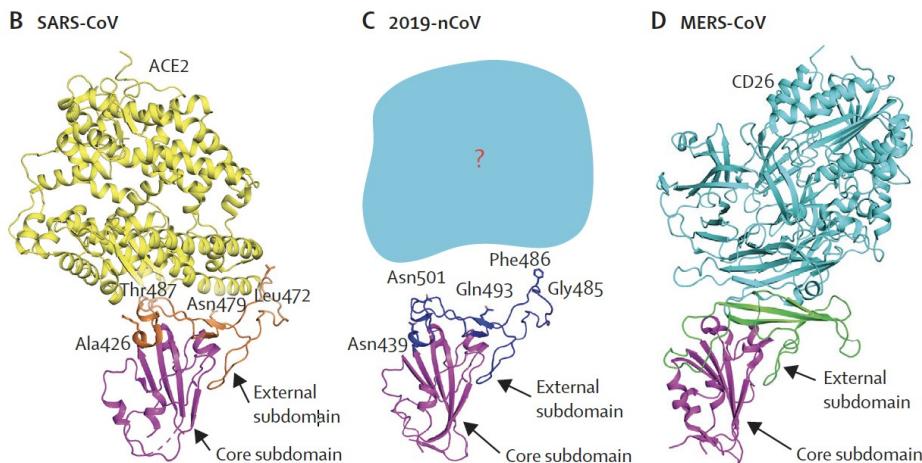
# 2019 nCoV-2 sequences are closest to bat CoV



# How does 2019-nCoV attach to human cells?



Receptor binding domain of spike protein closely related to SARS and bat CoV



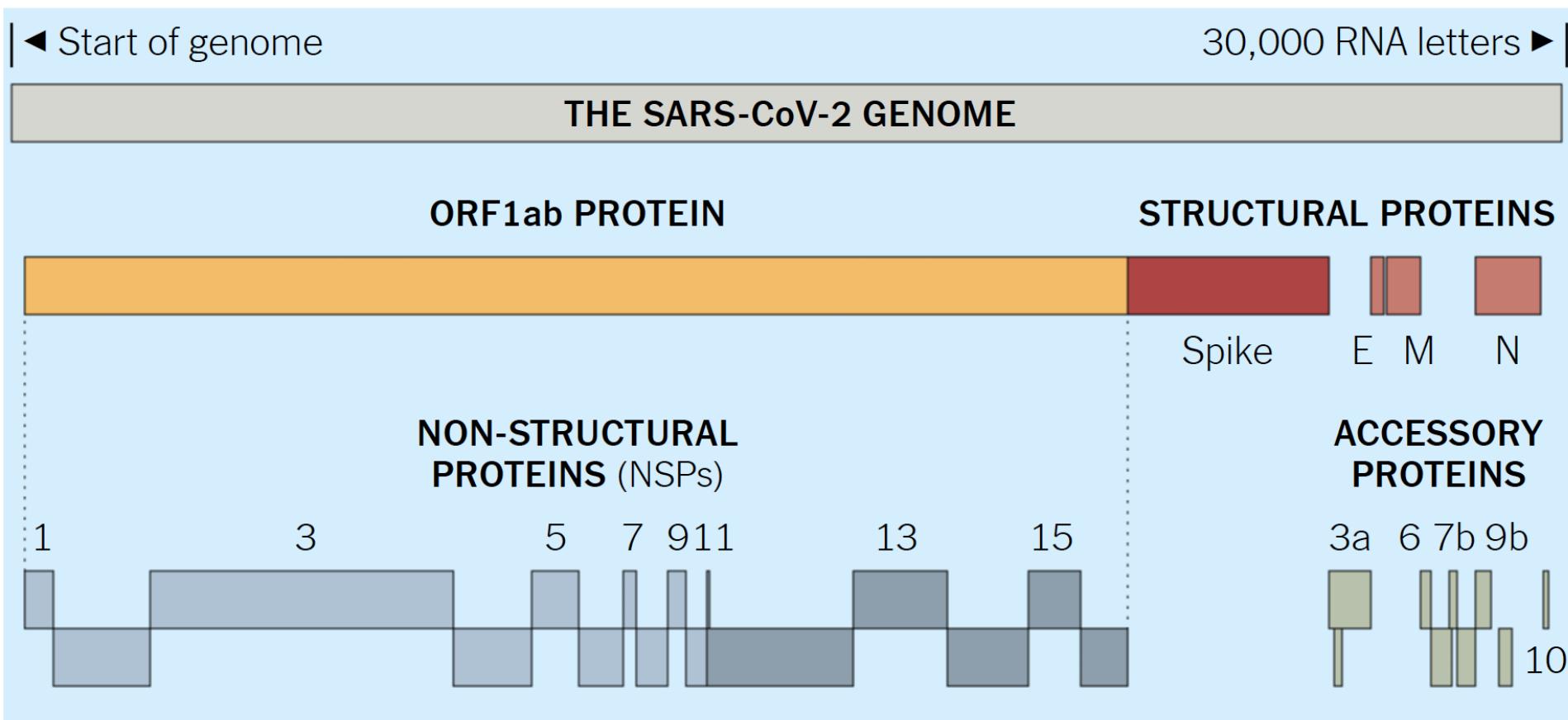
SARS CoV interacts with human ACE2 receptor, likely same for 2019-nCov

## Takeaways from the 2019-nCoV sequence and annotation

- Single strand RNA genome -- 29,829 bases in length
- Single introduction of the virus into humans, and then human to human spread
- Most similar to bat beta coronaviruses, but may not have come directly from bats, but instead from intermediary species
- The spike protein may interact with human ACE2 receptor protein (this is supported by recent structural work)

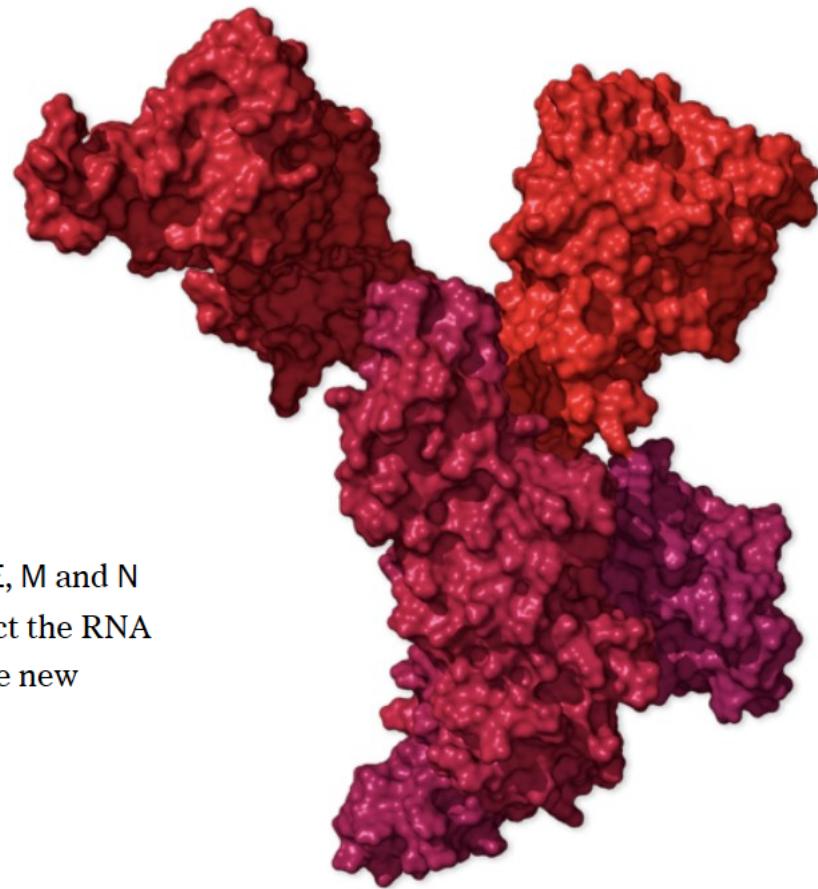
# An annotation of 2019-nCoV

'Bad news wrapped in protein' Corum & Zimmer 4/3/20



Each protein-coding sequence is discussed in order

# Protein structures can be predicted by comparison to closely related examples, as can functions

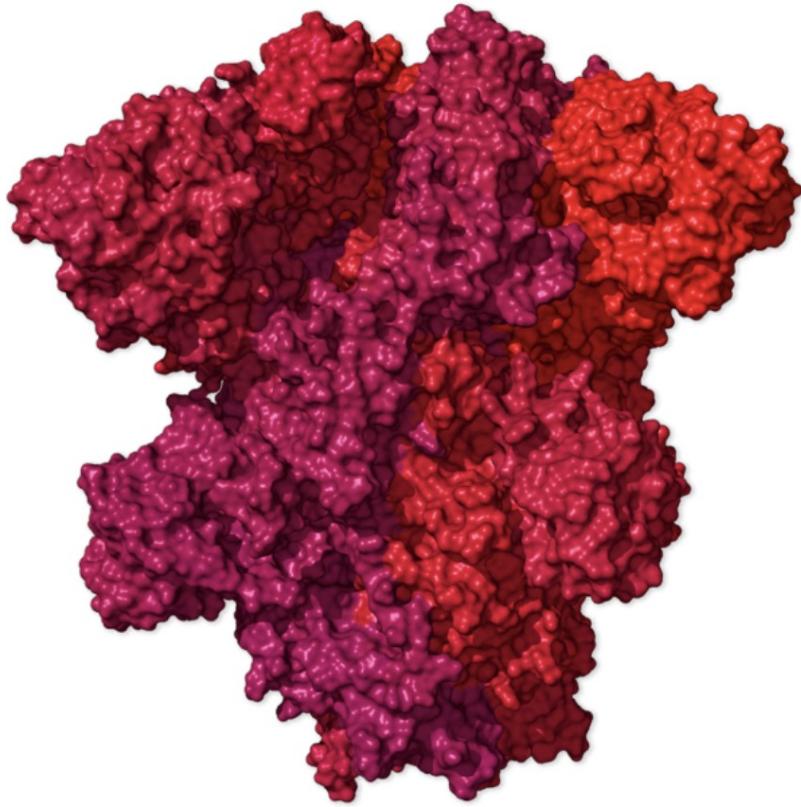


**Spike Protein · S**

The spike protein is one of four structural proteins — S, E, M and N — that form the outer layer of the coronavirus and protect the RNA inside. Structural proteins also help assemble and release new copies of the virus.

The S proteins form prominent spikes on the surface of the virus by arranging themselves in groups of three. These crownlike spikes give coronaviruses their name.

The spike protein is a target for the immune system, as well as for design of one class of antiviral therapies



Part of the spike can extend and attach to a protein called ACE2 (in yellow below), which appears on particular cells in the human airway. The virus can then invade the cell.

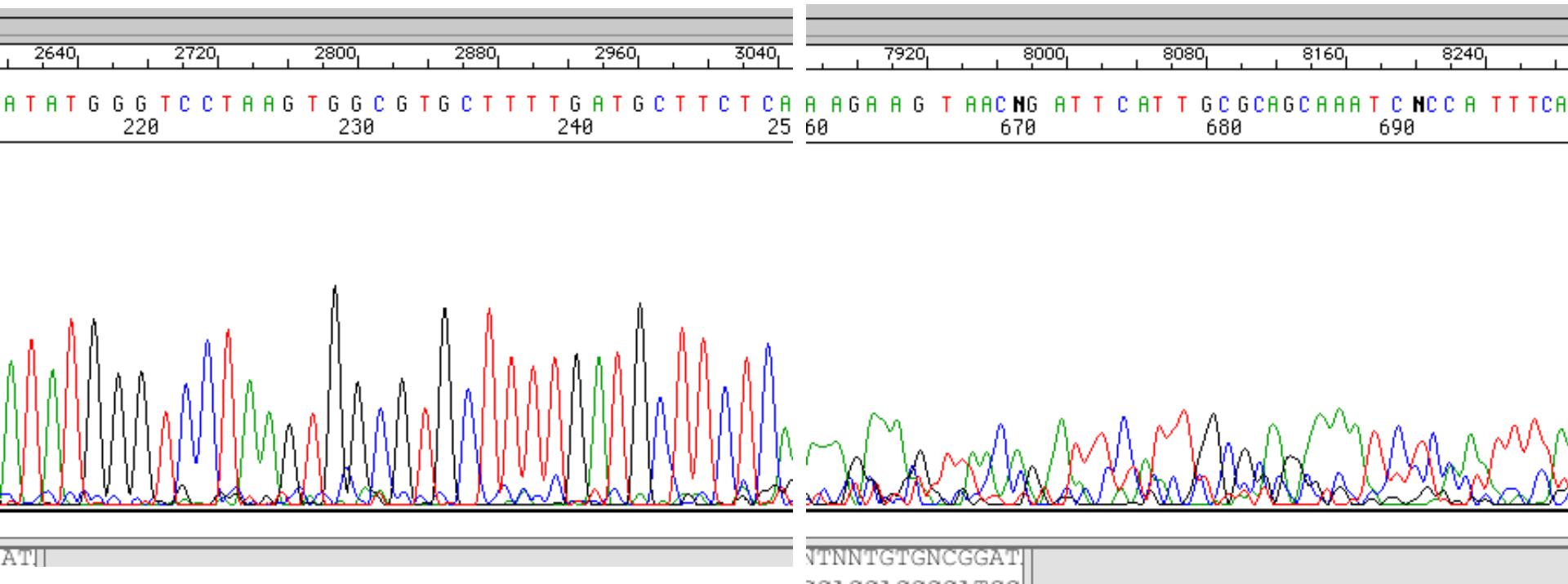
The spike protein changes conformation to interact with ACE2

# What's in a genome?

- 1) Genes
  - a) Protein-coding
    - Where are the open reading frames?
    - What are the ORFs most similar to? (What is the function/structure/evolution history?)
  - b) RNA
- 2) Non-genes
  - a) Regulation: promoters and factor-binding sites
  - b) Transactions: replication, repair, and segregation, DNA packaging (chromatin)

# Sequence output

## Raw data



## Computer calls

GNNTNNTGTGNCGGATACAATTCCCTCTAGAAATAATTTGTTAACCTTAAGAAGGAGATACATATGCACCAC  
CACCACCAACCATGGGTATGAATAAGCAAAAGGTTGCTCTGCTGTGAATCTGCGGAACCTATTATGATCCAGAAAG  
GGGGGAAATAGTCTGTGCCAAGTGCAGGTTATGTAATAGAAGAGAACATAATTGATATGGGTCTAACGTGGCGTGCTTTG  
ATGCTTCTCAAAGGGAACGCAGGTCTAGAAACTGGTGCACCAAGAAAGTATTCTTCTTCATGACAAGGGCTTCACTGCA  
ATTGGAATTGACAGATCGCTTCCGGATTAATGAGAGAGAACATGTACCGTTGAGGAAGTGGCANTCCANATTANGAGT  
TAGTGATGCAGCANANAGGAACCTAGCTTGCCTAAGTGAGTTGGATAGAATTNCTGCTCAGTAAAACCTCCNNGAC  
ATGTAGAGGAAGAAGCTGCAANGCTGNACANAGANGCAGNGNGANAGGGACTTATTNGANGCAGATCTATTGAGAGCGTT  
ATGGCGGCANGTGTACCTGCTTAGGTTATTAAAAGNTCCGGGACTCTGGATGAGATTGCTGATATTGCTAGAGC

# What does this sequence do?

atgttgtatttgtctgaagaaaataaatccgtatccactccttg  
ccctcctgataagattatcttgcagagaggggggagtaca  
tttgctctgaaactggagaagtttagaagataaaattatagat  
caagggccagagtggagggcttcacgccagaggagaagaaaa  
gagaagcagagtggagggcttaaacaataactattcacgata  
gggttatccactcttataactggaaagataaggatgctatg  
ggaagaactttagaccctaagagaagacttgaggcattgagatg  
gagaaagtggcaaattaga

Perhaps it encodes a protein...

Does the DNA encode a protein?  
Find an open reading frame (ORF) using “ ORF Finder”

<https://www.ncbi.nlm.nih.gov/orffinder/>

- a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database
- Identifies all open reading frames using the standard or alternative genetic codes
- Deduced amino acid sequence can be searched against other sequence databases, e.g. using the WWW BLAST server

Look up the orfs using NCBI ORF finder:

# ORF identification: things to consider

- The identification of ORFs catches most, but not all protein coding genes
- Not all genes initiate with ATG, e.g. in certain microbes (e.g. archaea)
- What is the shortest possible length of a real ORF? 100 amino acids is the typical boundary, but:
  - There are many ORFs of 100-150 codons that don't encode proteins
  - There are some ORFs of less than 100 codons that do encode proteins
- In eukaryotes, identifying the full protein coding region is complicated by the presence of introns

# What is the function of the ORF?

## Classical methods (slow, but reliable)

- mutate gene, observe phenotype for clues to function  
*(genetics)*
- purify protein product, test activity *in vitro*  
*(biochemistry)*

## Similarity of ORF to other genes

- if a gene has been previously studied, you want to know right away!
- gene sequences that have high sequence identity often have the same or similar functions

# Homology of proteins

Homology: similarity of biological structure, physiology, development, and evolution, based on common ancestry

Homologous proteins: statistically similar sequence *may* indicate similar function

# Alignment of sequences

The principle: two homologous sequences derived from the same ancestral sequence will have at least some identical (similar) amino acid residues that allow them to be aligned

Alignment quality is judged by **three** things:

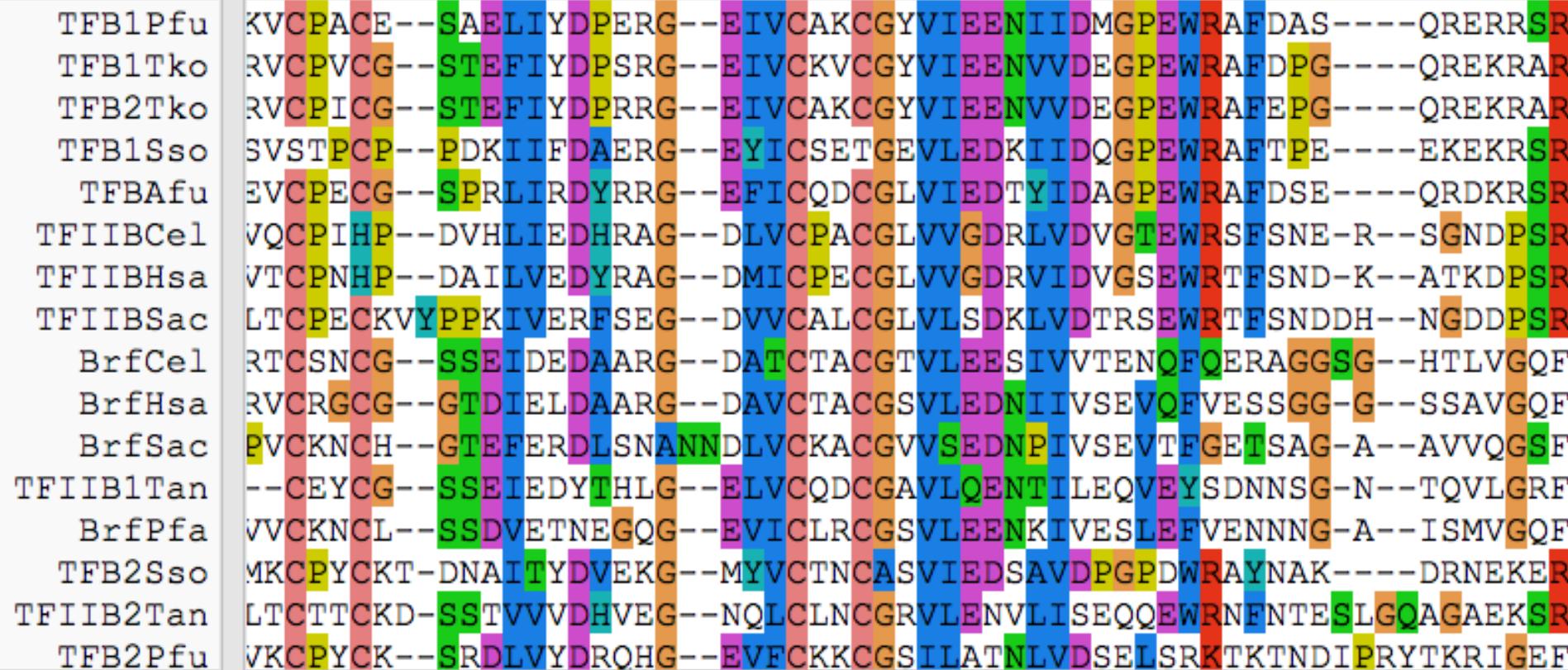
- 1) Percent identity: fraction of identical amino acids as a measure of structural/functional similarity
- 2) Similarity score: amino acids that have similar physical/chemical properties are more likely to substitute for each other in important functional regions, therefore contribute to alignment quality
- 3) Gaps in similar/homologous sequences are infrequent, and are cause penalty scores in alignment

# Alignment of specified sequences: archaeal TFB and eukaryotic TFIIB and Brf. Similar sequence, and similar structure/function

ClustalX 2.1

Multiple Alignment Mode

Font: 18



30

40

50

60

70

80

# NCBI: National Center for Biotechnology Information

NCBI home page --Go to [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) for the following (and much, much more)

PubMed, PubMed Central: search tool for scientific literature--search by author, subject, title words, etc.

BLAST: Basic Local Alignment Search Tool

Nucleotide, Genome, Gene, Protein: databases for each

OMIM: Online Mendelian Inheritance in Man

Bookshelf: many online textbooks available

GEO: Gene Expression Omnibus, for analysis of microarray and related data

PubChem: Information of biological activities of small molecules

Guide to NCBI: see first entry at this link

<https://academic.oup.com/nar/issue/49/D1>

## Using NCBI: Education and Tutorials pages

“ The Handbook” :

<https://www.ncbi.nlm.nih.gov/books/NBK143764/>

Video guides and tutorials:

<https://www.ncbi.nlm.nih.gov/home/learn/>

Other training and tutorials:

<https://www.ncbi.nlm.nih.gov/guide/training-tutorials/>

Help manual:

<https://www.ncbi.nlm.nih.gov/books/NBK3831/>

BLAST guide:

[ftp://ftp.ncbi.nih.gov/pub/factsheets/HowTo\\_BLASTGuide.pdf](ftp://ftp.ncbi.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf)

**Note:** Lots of other non-NCBI servers exist for bioinformatics tools,  
see index at:

<https://academic.oup.com/nar/issue/49/D1>

# What does BLAST do?

- 1) Searches the chosen sequence database and identifies sequences with similarity to test sequence
- 2) Ranks similar sequences by degree of homology (E value)
- 3) Illustrates alignment between test sequence and similar sequences

# Programs available for BLAST searches

## Protein sequence

**blastp**--compares an amino acid query sequence against a protein sequence database

**tblastn**--compares a protein query sequence against a nucleotide sequence database translated in all reading frames

## DNA sequence

**blastn**--compares a nucleotide query sequence against a nucleotide sequence database

**blastx**--compares a nucleotide query sequence translated in all reading frames against a protein sequence database

**tblastx**--compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

# How a protein BLAST search works

A query protein sequence is first converted into overlapping segments (words)

Synonyms for each query word are located in a sequence database, with “ scores” for each, based on a Dayhoff matrix (e.g. BLOSUM 62: BLOcks of amino acid **S**Ubstitution **M**atrix)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4	-1	-2	-2	0	-1	-1	0	-2	-3	-2	-1	-1	-2	-1	1	0	-3	-2	0
Arg	5	-1	0	-2	-3	1	-4	-2	-3	-3	-2	-1	-1	-2	-1	-2	-1	-4	-3	-1
Asn	-2	-1	0	6	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
Asp	-2	-2	-2	1	6	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
Cys	0	-3	-3	-3	-3	9	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
Gln	-1	1	0	0	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
Glu	-1	0	0	2	-4	2	5	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
Gly	0	-2	0	-1	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
His	-2	0	1	-1	-3	0	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	2	-1	-1	-1	-1	-1	-1	-1	-1
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	0	0	0	0	0	0	0
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-3	-1	-1	-1	-1	-1	-1	-1	-1
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-1	-1	-1	-1	-1	-1	-1
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	-2	11	7
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	4
Val	0	-3	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-1	-2	-2	0	-3

Synonym values over a T, or threshold value, are analyzed by extending out from the words, and a similarity/identity/gap score (S) is generated

# Margaret Dayhoff, founder of bioinformatics



- **1948:** PhD in quantum chemistry
- **1961** and onward: worked as **evolutionary biologist**, and developed the first substitution matrices used to judge the significance of alignments between protein sequences

These are the PAM matrices (Position Accepted Mutation), available today for your BLAST searches

- Invented the single letter codes for amino acids (very important for computation)
- **1965:** Initiated first collection of protein sequences, "Atlas of Protein Sequences and Structure"
- **1966:** with Richard Eck, created first computationally derived phylogenetic tree reconstruction
- **1971:** Created first computer database for protein sequences (the Protein Information Resource)
- **1980:** made largest nucleotide sequence database freely available by telephone network

<https://www.whatisbiotechnology.org/index.php/people/summary/Dayhoff>

<https://www.smithsonianmag.com/science-nature/how-margaret-dayhoff-helped-bring-computing-scientific-research-180971904/>

# Where did the BLOSUM62 alignment score matrix come from?

Sean R Eddy

NATURE BIOTECHNOLOGY VOLUME 22 NUMBER 8 AUGUST 2004

*"...details in BLOSUM62 that may seem counterintuitive at first glance. For instance, tryptophan (W/W) pairs score +11, while leucine (L/L) pairs only score +4; why shouldn't all identities get the same score? The rarer the amino acid is, the more surprising it would be to see two of them align together by chance. In the homologous alignment data that BLOSUM62 was trained on, leucine/leucine (L/L) pairs were in fact more common than tryptophan/trypophan (W/W) pairs ( $p_{LL} = 0.0371$ ,  $p_{WW} = 0.0065$ ), but tryptophan is a much rarer amino acid ( $f_L = 0.099$ ,  $f_W = 0.013$ ). Run those numbers (with BLOSUM62's original  $\lambda = 0.347$ ) and you get +3.8 for L/L and +10.5 for W/W, which were rounded to +4 and +11."*

# BLAST scores

In any given alignment:

- matches (identical or similar) RAISE score
- mismatches LOWER score
- gaps LOWER score

Three criteria are used in the display of the highest scoring ('best') sequences:

- 1) percent identity
- 2) similarity score
- 3) **E-value**--probability that two sequences will have the similarity they have by chance (lower numbers mean a higher probability of evolutionary homology, and so a higher probability of similar function)

## What is the E-value? (<https://youtu.be/nO0wJgZRZJs>)

The E value (also called Expect value) represents **the chance that the similarity is random and therefore insignificant.**

...the E value describes the random background noise that exists for matches between sequences. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

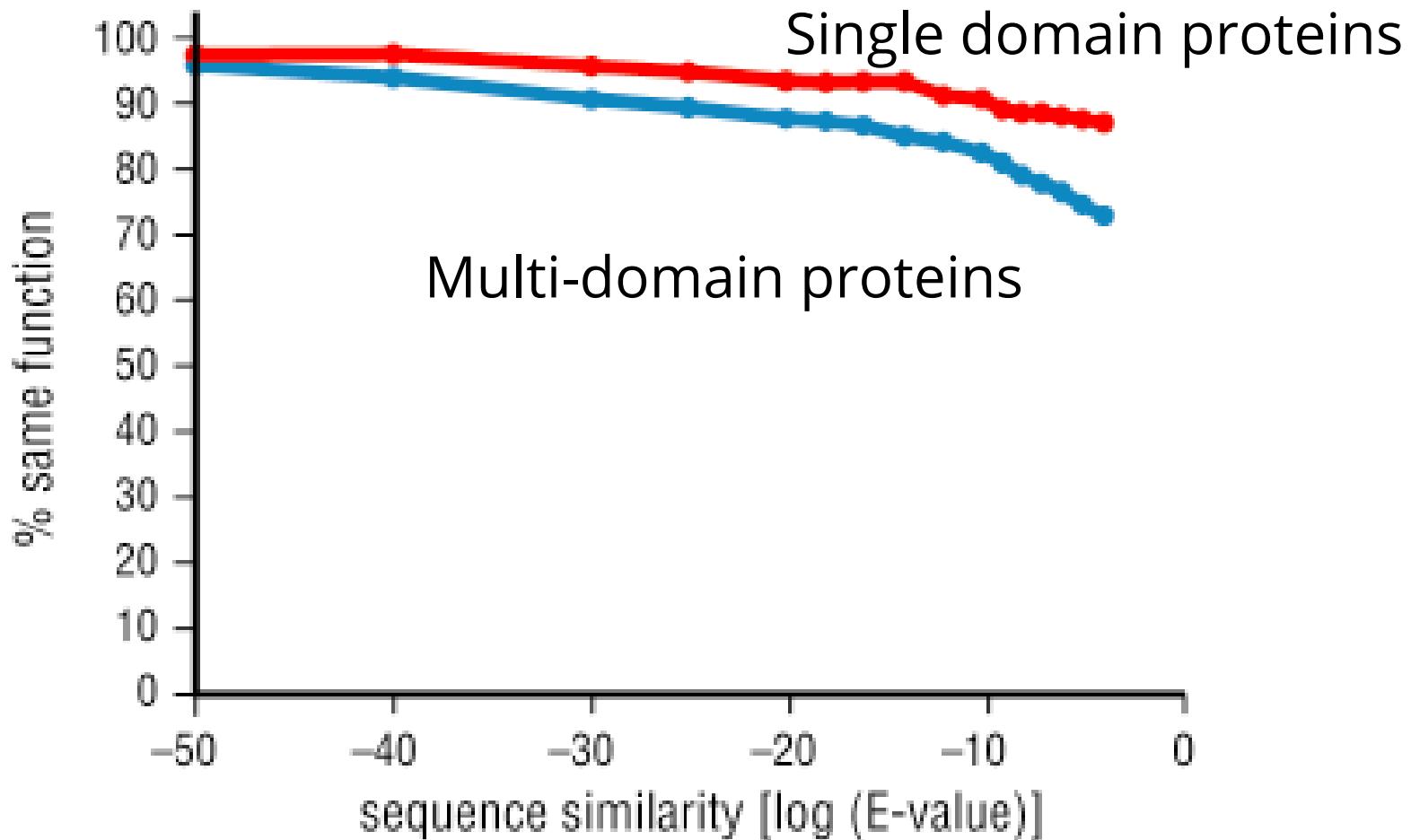
You can change the Expect value threshold on most main BLAST search pages. **When the Expect value is increased from the default value of 10, a larger list with more low-scoring hits will be listed.**

## E values (continued)

From the BLAST tutorial:

Although hits with E values much higher than 0.1 are unlikely to reflect true sequence relatives, it can be useful to examine hits with lower significance (E values between 0.1 and 10) for short regions of similarity. In the absence of longer similarities, these short regions may allow the tentative assignment of biochemical activities to the ORF in question. The significance of any such regions must be assessed on a case by case basis.

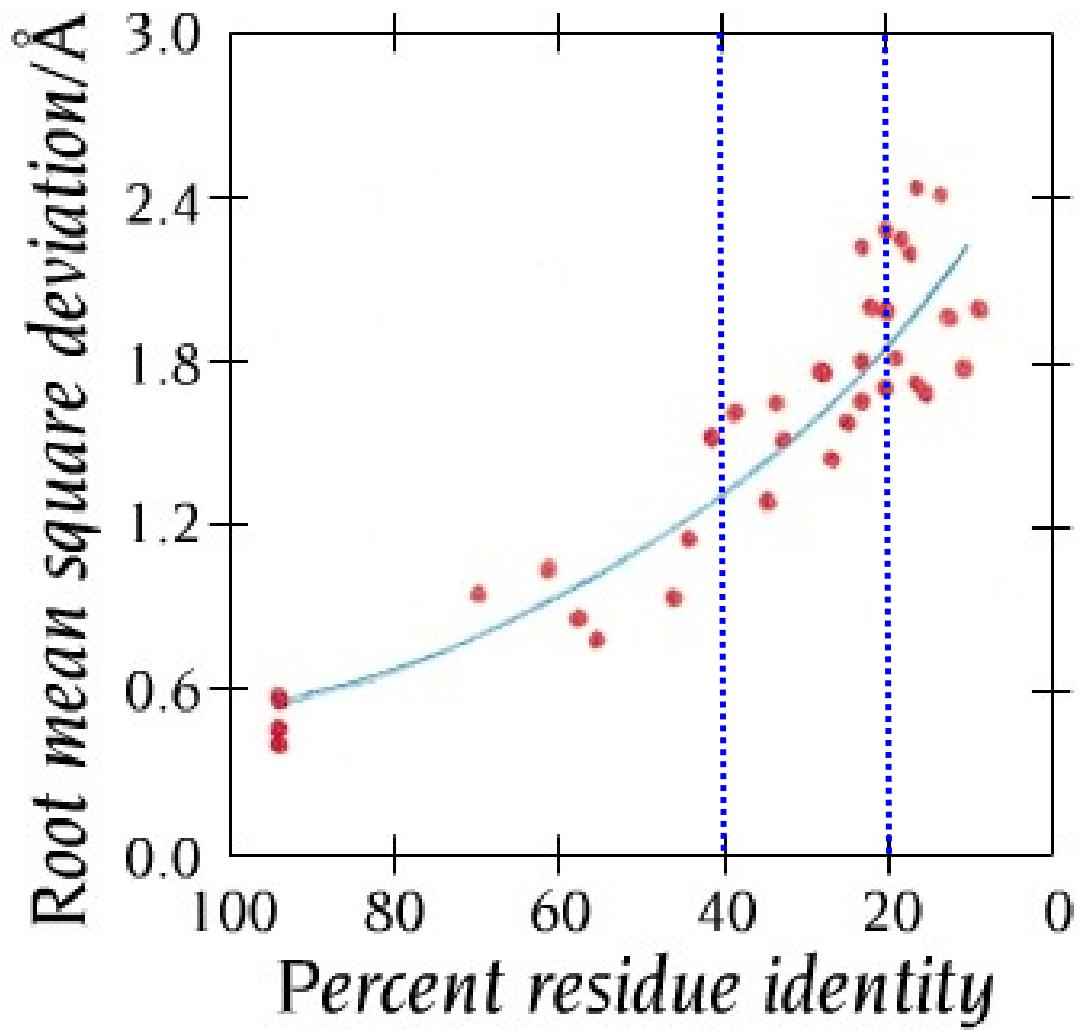
## Relationship between E-value and function



E value greater than  $10^{-10}$ , could be a similar structure but may have a different function

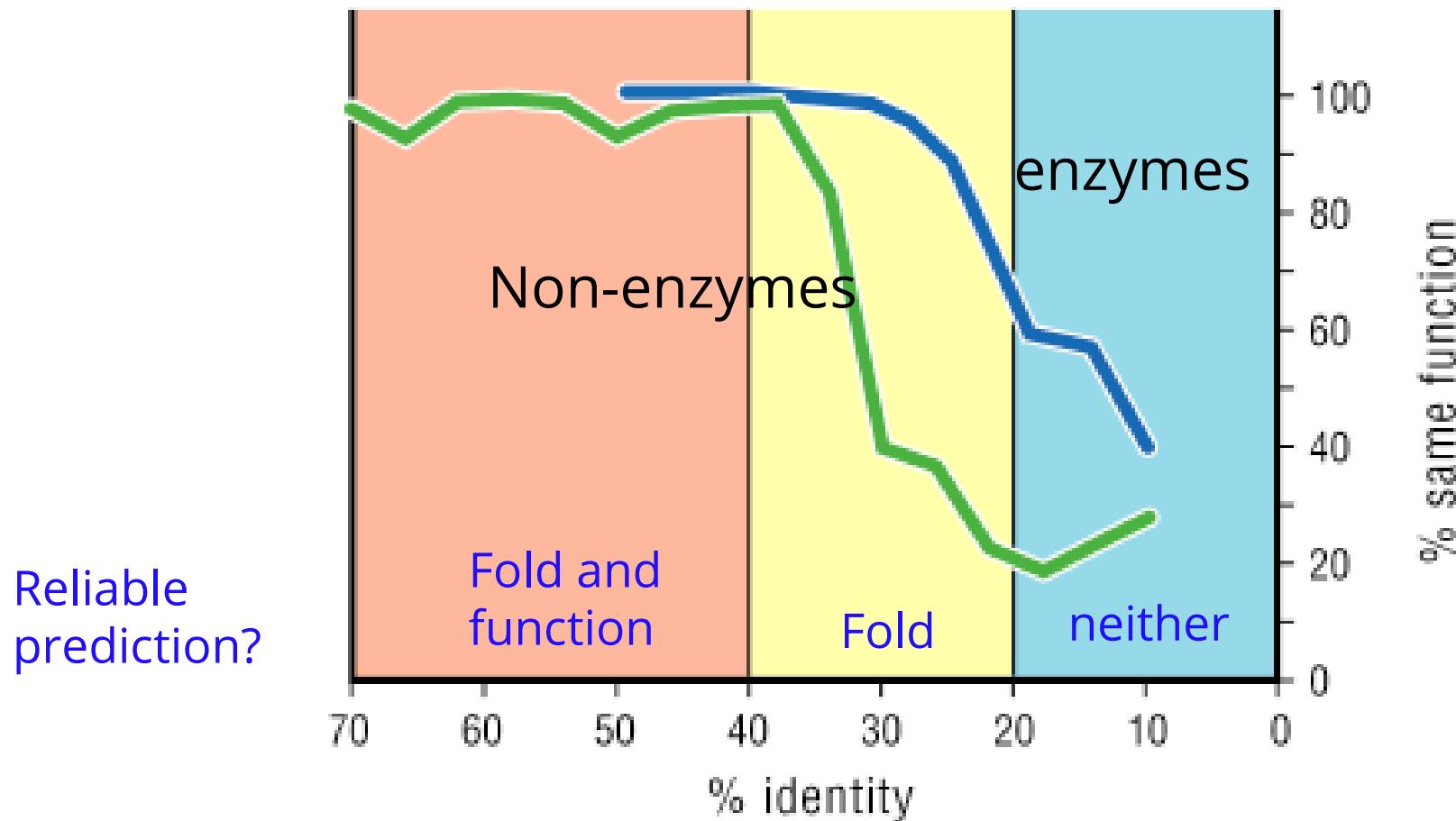
# Similar sequence = similar structure

Deviation in main chain atoms in protein cores increases as percent identity of protein sequence decreases



(Plot calculated from the known structures of 32 pairs of homologous proteins)

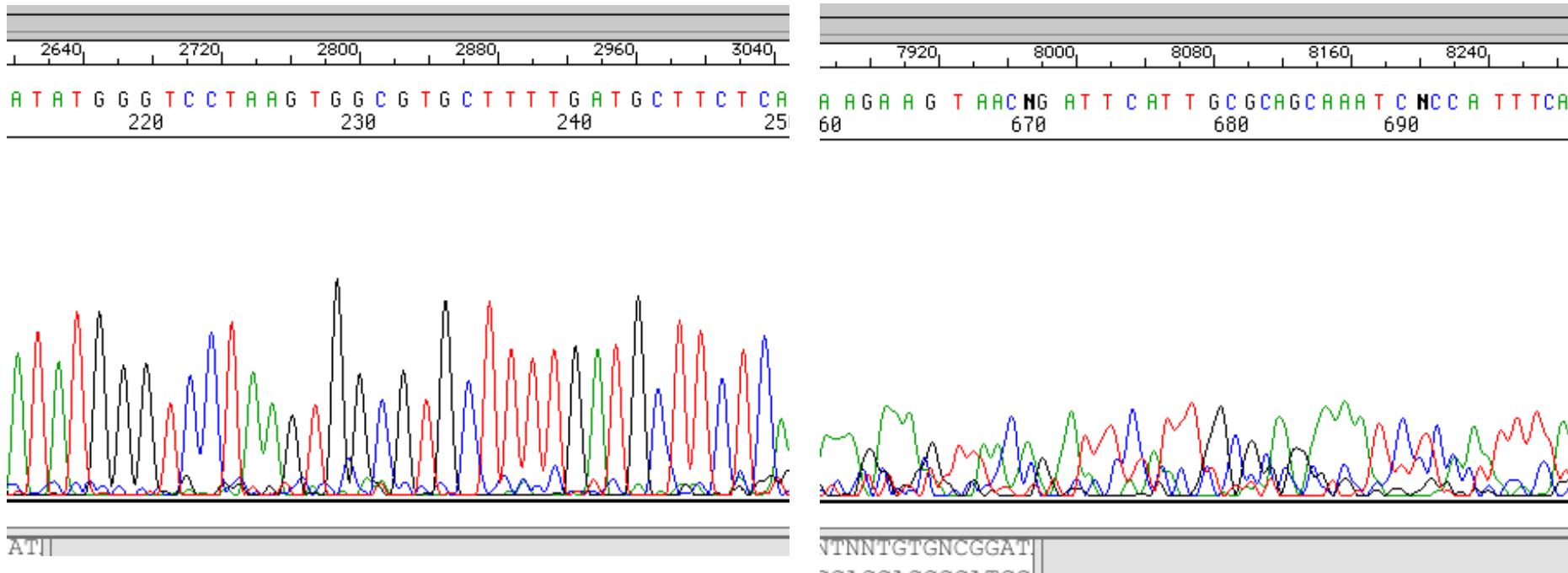
# High sequence similarity correlates with functional similarity



40-20% identity: fold can be predicted by similarity but precise function cannot be predicted (the 40% rule)

# Biological function of a new sequence? BLAST.....

## Raw data



Computer calls GNNTNNNTGTGNCGGATACAATTCCCCCTCTAGAAATAATT

TTGTTAACCTTAAGAAGGGAGATACATATGCACCAACCAC

CACCACCCACCCATGGGTATGAATAAGCAAAAGGTTGTCCTGCTTGAAATCTGCGGAACCTTTATGATCCAGAAAG

GGGGGAAATAGTCTGTGCCAAGTGCGGTTATGTAATAGAAGAGAACATAATTGATATGGGTCTAAGTGGCGTGCTTTG

ATGCTTCTCAAAGGGAACGCAGGTCTAGAAACTGGTGCACCAGAAAGTATTCTC

TTCATGACAAGGGGCTTCAACTGCA

ATTGGAATTGACAGATCGCTTCCGGATTAATGAGAGAGAAGATGTACCGTTGAGGAAGTGGCANTCCANATTANGAGT

TAGTGATGCAGCANANAGGAACCTAGCTTGCCTAAGTGAGTTGGATAGAATTNCTGCTCAGTTAAACTTCCNNGAC

ATGTAGAGGAAGAAGCTGCAANGCTGNACANAGANGCAGNGNGANAGGGACTTATTNGANGCAGATCTATTGAGAGCGTT

ATGGCGGCANGTGTACCTGCTTAGGTTATTAAAAGNTCCCAGGGACTCTGGATGAGATTGCTGATATTGCTAGAGC

Find the open reading frame(s) and translate

MKCPYCKSRDLVYDRQHGEVFCKKCGSILATNLVDSELSRKT  
KTNDIPRYTKRIGEFTREKIYRLRKWQKKISSERNLVLAMSE  
LRRLSGMLKLPKYVEEEAAYLYREAAKRGLTRRIPIETTVAA  
CIYATCRLFKVPRTLNEIASYSKTEKKEIMKAFRVIVRNLNL  
TPKMILLARPTDYVDKFADELELSERVRRRTVDILRRANEEGI  
TSGKNPLSLVAAALYIASLLEGERRSQKEIARVTGVSEMTVR  
NRYKELA

# Query Sequence in FASTA Format

Amino acid sequence of a protein, in FASTA format:

```
>ribosomal protein L7/L12 [Thiomicrospira crunogena XCL-2]
MAITKDDILEAVANMSVMEVVELVEAMEEKFGVSAAAVAVAGPAGDAGAA
GEEQTEFDVVLTGAGDNKVAAIKAVRGATGLGLKEAKSAVESAPFTLKEG
VSKEEAETLANELKEAGIEVEVK
```

Nucleotide sequence of a gene, in FASTA format:

```
>gi|118139508:333094-333465 Thiomicrospira crunogena XCL-2
ATGGCAATTACAAAAGACGATATTTAGAACAGCAGTTGCTAACATGTCAGTAATGGAAGTTGT
TGAACTTGTTGAAGCAATGGAAGAGAAGTTGGTGTTCAGCAGCAGTTGCAGGTTGCAG
GTCCTGCAGGTGATGCTGGCGCTGCTGGTGAAGAACAAACAGAGTTGACGTTGTCTTGA
GGTGTGGTGACAACAAAGTTGCAGCAATCAAAGCCGTTCGTGGCGCAACTGGTCTTGGGCT
TAAAGAACGAAAAGTGCAGTTGAAAGTGCACCATTACGCTTAAAGAGGGTGTCTAAAG
AAGAACGAGAAACTCTTGCAAATGAGCTTAAAGAACAGGTATTGAAGTCGAAGTTAAATAA
```

# The FASTA format

”description line” (not read as sequence data)

- Begins with >
- Ends with a hard return

Sequence data  
(amino acid in this  
case)

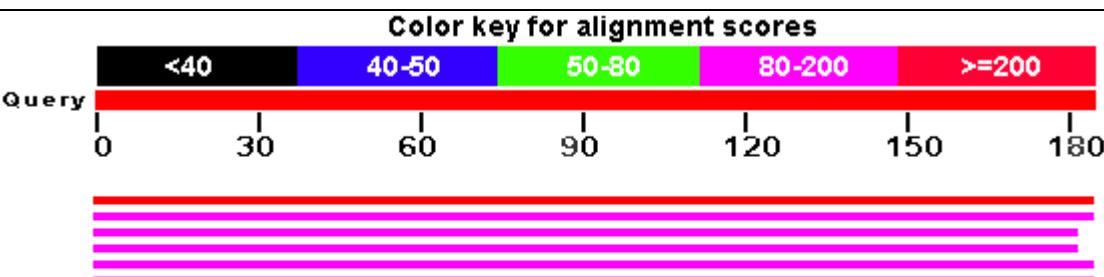
```
> ribosomal proteinL7/L12
MAITKDDILEAVANMSVMEVVELVEA
MEEKFGVSAAAVAVAGPAGDAGAA
GEEQTEFDVVLTGAGDNKVAAIKAVR
GATGLGLKEAKSAVESAPFTLKEG
VSKKEEAETLANELKEAGIEVEVK
```

# NCBI BLAST Interface (blastp: Proteins)

The screenshot shows the NCBI BLAST interface for protein searches (blastp). The top navigation bar includes links for Home, Recent Results, Saved Strategies, and Help, along with My NCBI and Sign In/Register options. The main search area has tabs for blastn, blastp (which is selected), blastx, tblastn, and tblastx. A large text input field is labeled "Enter Query Sequence". Below it, there's a "Query subrange" section with "From" and "To" fields. A note says "BLASTP programs search protein databases using a protein query." A yellow box highlights the "Or, upload" link and the "Job Title" input field, with the text "(Paste FASTA format sequence here)" overlaid. There's also a link to "Enter a descriptive title for your BLAST search". A checkbox for "Align two or more sequences" is present. The "Choose Search Set" section allows selecting a database (Non-redundant protein sequences (nr)), organism (via a dropdown menu or text input), and exclude options (models, uncultured sequences). An Entrez query input field is also available.

# NCBI BLAST Results Page:

## Potential homologs retrieved from database



### Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
NP_440048.1	potential FMN-protein [Synechocystis sp. PCC 6803] >sp P727	379	379	100%	1e-103
YP_001864295.1	flavin reductase domain-containing protein [Nostoc punctiforme]	199	199	100%	2e-49
YP_321888.1	flavin reductase-like, FMN-binding [Anabaena variabilis ATCC	198	198	98%	3e-49
NP_488484.1	flavoprotein [Nostoc sp. PCC 7120] >sp Q8YNW7.1 DFA4_ANA	197	197	98%	6e-49
CAO89562.1	dfa4 [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630850.1	flavoprotein [Nodularia spumigena CCY9414] >gb EAW44518.	193	193	100%	6e-48

>[ref|YP\\_002482587.1|](#) G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]

[gb|ACL44226.1|](#) G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]  
Length=585

[GENE ID: 7287783 Cyan7425 1859](#) | flavin reductase domain protein FMN-binding  
[Cyanothecae sp. PCC 7425]

Score = 176 bits (446), Expect = 1e-42, Method: Compositional matrix adjust.  
Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query 1 SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVWITQTGRH----- 52  
+G++FA+ L+ K+QR RQ+ E Q+DRT+QAVGRIIGS+ V+T + H

Sbjct 393 AGSDFAQVLKKAKKQRSPRSQPILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP 452

Query 53 -----QGILTSWVSQASFPTPPGIMLAIPGEFDAYGLAGQNKAFVLFNLLQEGRSVRRHFDH 107  
+L SWVSQASF PPG+ +A+ E A GL AFVLFN+L+EG ++RRHF

Sbjct 453 QLEVPTAMLVSWVSQASFNPPLGTIALAKE-RAEGLDHSGDAFVLNVLKEMNLLRRHFSK 511

Query 108 QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVHQ 167  
P G++ F+ L +NGC +L + LAYLEC VQS GDH L+YATV G+VHQ

Sbjct 512 SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATWNNGKVLQ 569

Query 168 PNGITAIRHRKSGGQY 183  
P G TA++HRKSG QY

Sbjct 570 PTGTTAVQHRKSGNQY 585

BLAST page: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

-- go to “protein blast”

-- Program: blastp

-- Search set: leave blank, or choose taxonomic group (example: archaea) <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>

-- choose algorithm: blastp (default)

-- PSI-, PHI-, and DELTA-BLAST are variants that can find homologs with relatively low sequence identity (distant homologs),  
DELTA- is most modern and seems to work best (esp. cross-domain homology)

-- click “ BLAST” button

## View Report:

- 1) Distribution of hits: query sequence followed color-coded positions ‘hit’ sequences that gave alignments
- 2) Sequences producing significant alignments (sorted by Max score)
  - Accession number (this takes you to the sequence that yielded the hit: gene or contig)
  - Description: gene name, organism
  - Max score, total score, query coverage
  - E-value

## DELTA-BLAST:

- Domain Enhanced Lookup Time Accelerated BLAST
- The algorithm constructs a specific PSSM (position-specific scoring matrix) using the results of a Conserved Domain Database (CDD) search
- A PSSM differs from typical PAM and BLOSUM matrices in which scores are position independent
- The constructed PSSM is used to search sequence database

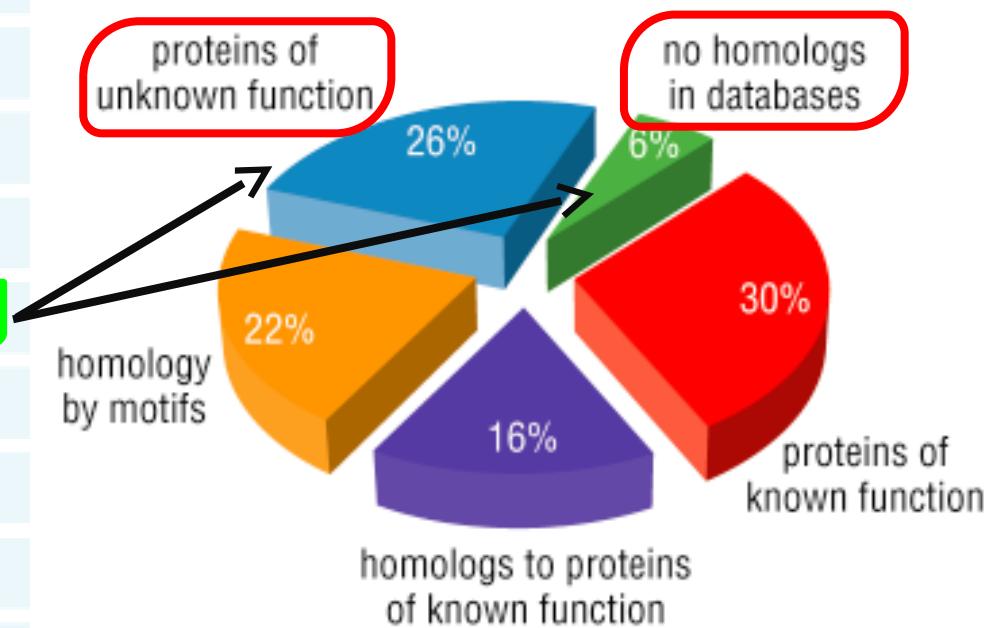
# BLAST is good, but not perfect

1) High homology? *function is inferred* but not known

2) Many coding proteins  
*cannot be assigned function*  
based on homology

Genome Sizes of Representative Organisms

Organism	Genome size (base pairs)	Number of genes
<i>Mycoplasma genitalium</i>	$45.8 \times 10^5$	483
<i>Methanococcus jannaschii</i>	$1.6 \times 10^6$	1,783
<i>Escherichia coli</i>	$4.6 \times 10^6$	4,377
<i>Pseudomonas aeruginosa</i>	$6.3 \times 10^6$	5,570
<i>Saccharomyces cerevisiae</i>	$1.2 \times 10^7$	6,282
<i>Caenorhabditis elegans</i>	$1.0 \times 10^8$	19,820
<i>Drosophila melanogaster</i>	$1.8 \times 10^8$	13,601
<i>Arabidopsis thaliana</i>	$1.2 \times 10^8$	25,498
<i>Homo sapiens</i>	$3.3 \times 10^9$	-30,000 (?)



# Bioinformatics: making sense of biological sequence

- New DNA sequences are analyzed for ORFs (Open Reading Frames: protein)
- Any DNA or protein sequence can then be compared to all other sequences in databases, and similar sequences identified
- There is much more -- a huge number of bioinformatics tools are available

## Midterm exam: Friday, April 30

Exam covers topics through the end of this discussion of '-Omics'

See quizzes 1-5, and assignments 1 and 2 for examples of the kinds of questions to expect

# Parallel measurements of gene expression/activity: RNA and protein “ -omics”

## 1. The ‘transcriptome’

- Detecting expression of many genes: arrays of inverted northern blots
- RNA-seq: ‘next gen’ sequencing provides an alternative to arrays

## 2. The ‘proteome’

- ID of all proteins in a mixture: 2-D gels and mass spectrometry
- ID of DNA-binding protein locations: Chromatin Immunoprecipitation
- Protein activity measurements

## Guide to readings:

- 1) 21 MC4 ChIP. Introduction to Chromatin Immunoprecipitation for DNA binding proteins
- 2) RNA seq 2009. The RNA-seq approach to transcriptome analysis.
- 3) Chip seq 2010. Chromatin immunoprecipitation followed by sequencing.
- 4) Mass spec proteomics 2007. Using mass spectrometry for identifying proteins in a complex mixture.
- 5) The ENCODE project: Encyclopedia of human DNA elements
- 6) Mass data false positives 2012. Big data sets can lead to false conclusions

# What is an -ome?

The totality of \_\_\_\_.

Genome (all the genes)

Transcriptome (all the RNAs)

Proteome (all the proteins)

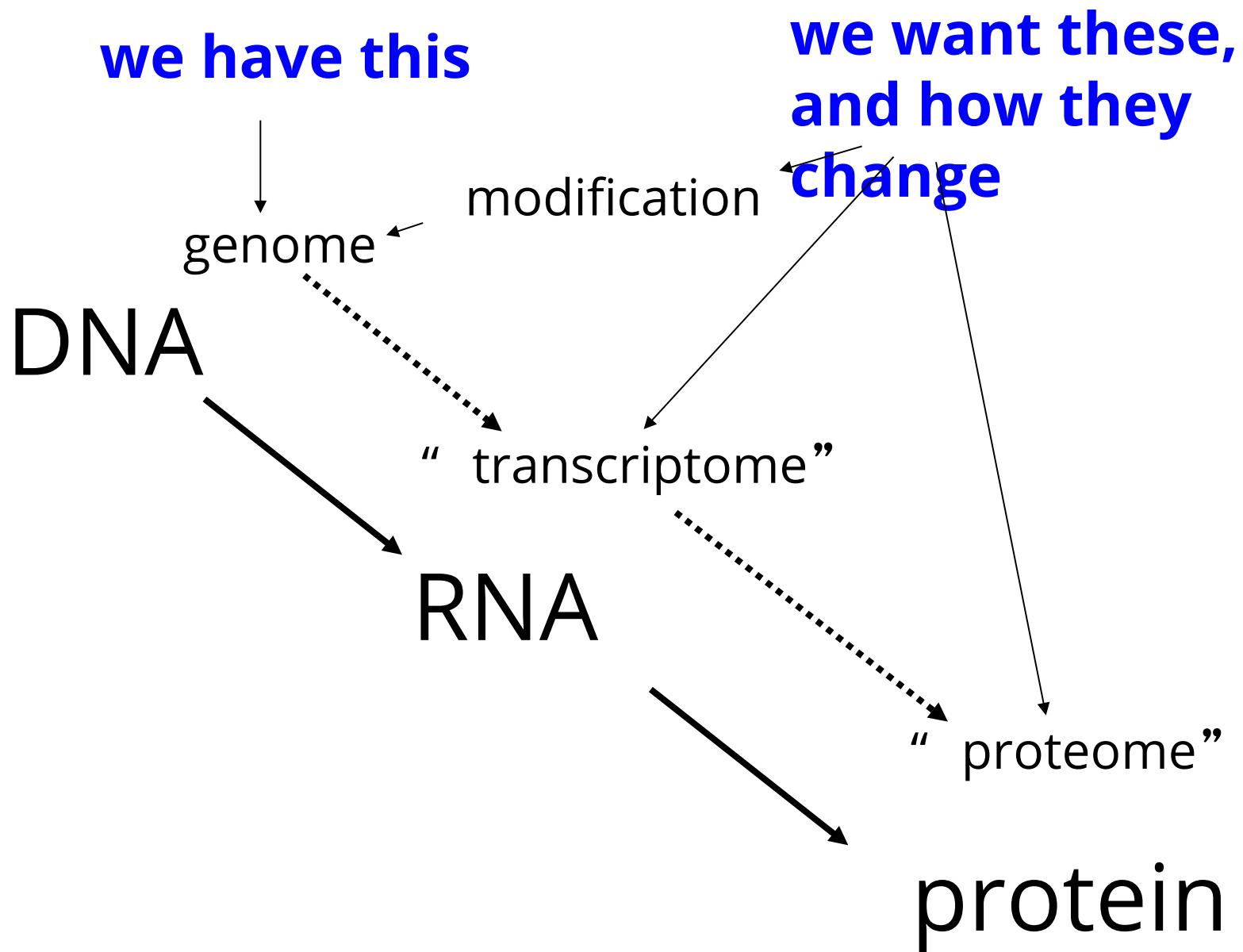
Methylome (all the sites of DNA methylation, epigenetic modifications)

Omeome (all the -omes)

<http://www.genomicglossaries.com/content/omes.asp>

## Why study the -omes?

To understand a complex system, we need to know what all the parts are, and what they are doing



# Detection of mRNA transcripts

- Northern Blot – immobilize mRNA on membrane, detect specific sequence by hybridization with **one labeled probe**--requires a separate blotting for each probe
- DNA microarray – immobilize many probes (thousands) in an ordered array, hybridize (base pair) with **labelled DNA**
- RNA-seq – isolate RNA, reverse transcribe to make DNA, ‘next-gen’ sequencing

# The value of DNA microarray/RNA-seq for studying gene expression

- 1) Measure the levels of all RNA transcripts at same time
- 2) RNA abundance usually determines the level of gene expression – a lot of gene control occurs at the level of transcription
- 3) Changes in transcription patterns correlate with changing environment – overarching patterns can be detected by microarray/RNA-seq, and may suggest new biological mechanisms

# DNA microarray: an array of probes

Identify protein coding genes (from open reading frames in the genome), then...

- PCR each gene, attach each PCR product to a solid support in a specific order

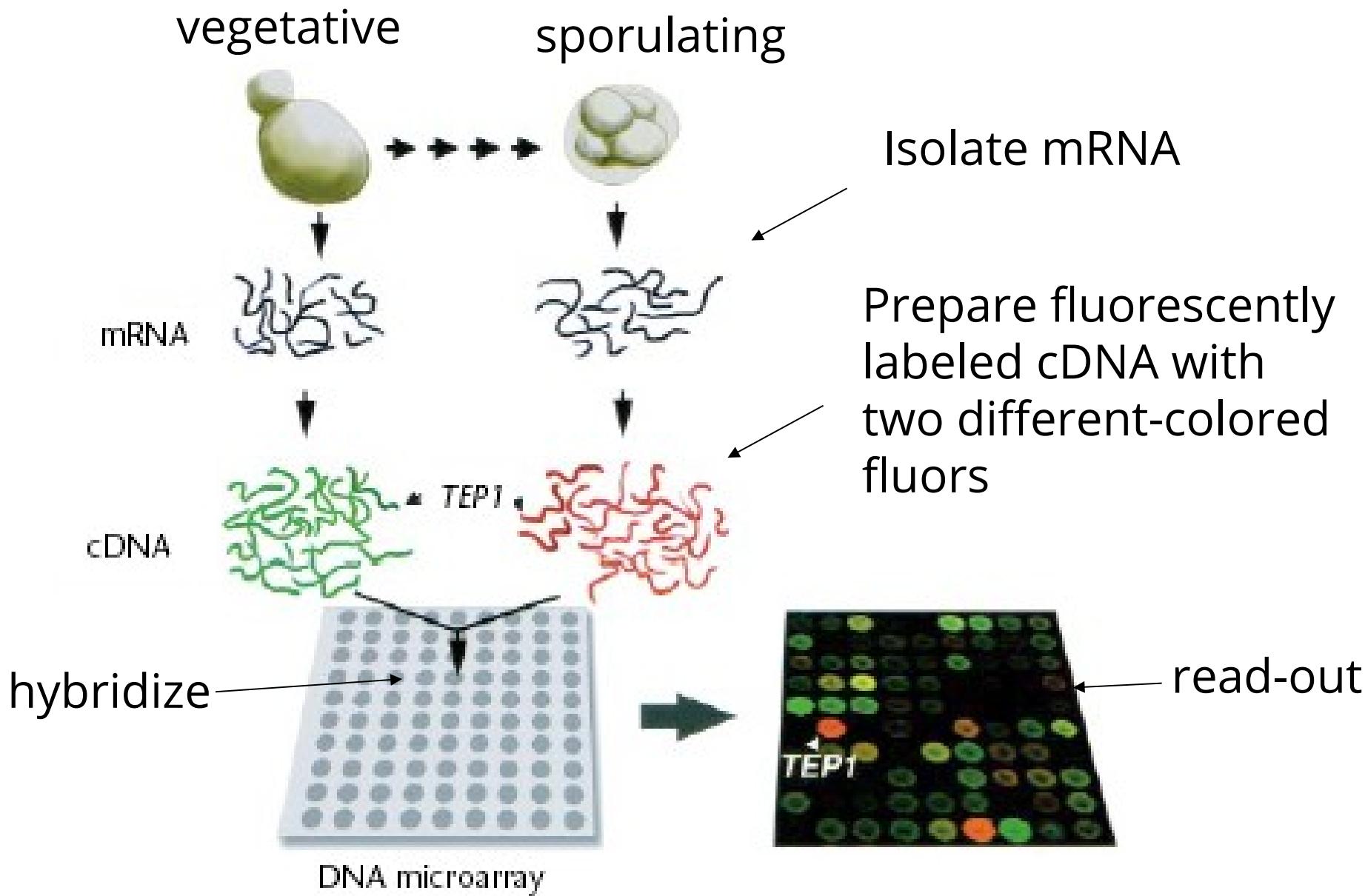
or

- Chemically synthesize gene-specific oligonucleotide probes directly on microchip

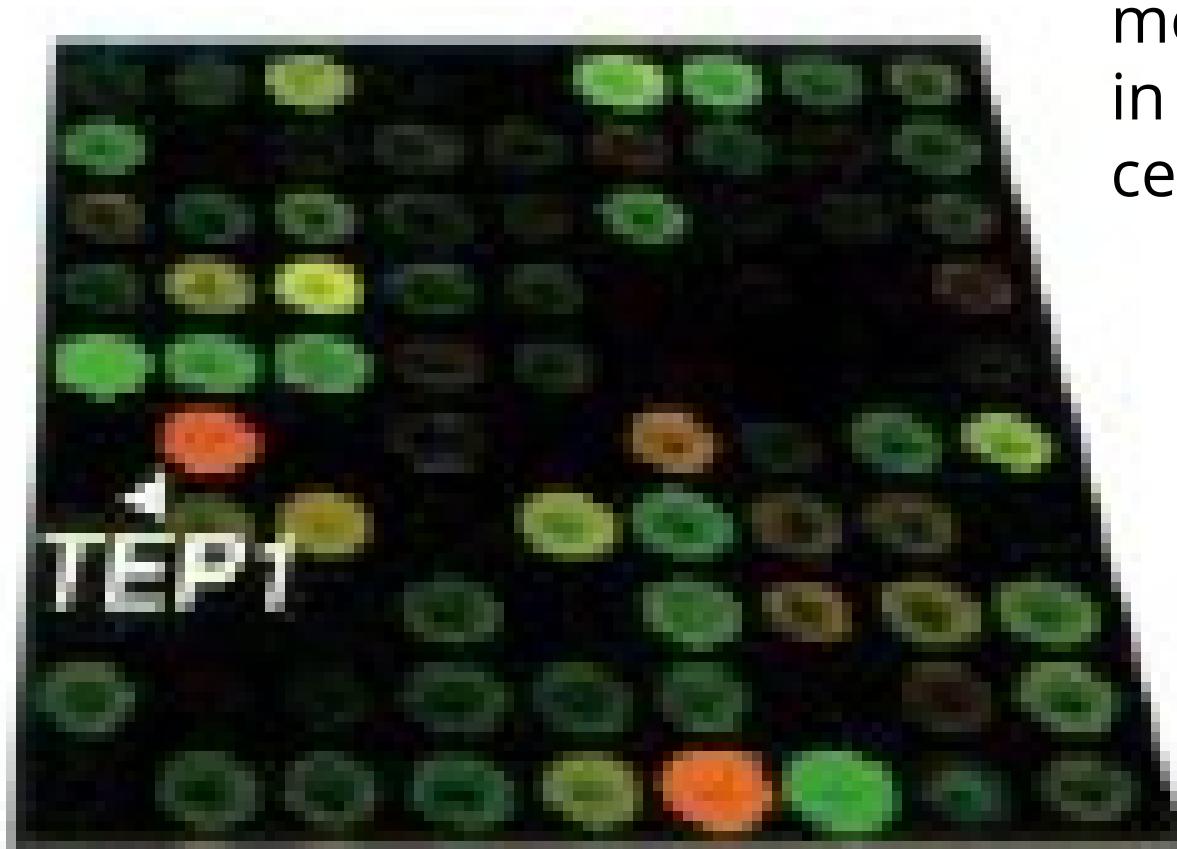
then

- Hybridize **labelled** RNA to the chip. More hyb. signal at a specific spot means more of that RNA

# Microarray: genes up-regulated during meiosis?



# Example microarray data



**Green:** mRNA  
more abundant  
in vegetative  
cells

**Yellow:** equivalent mRNA abundance in vegetative and sporulating cells

**Red:** mRNA more abundant in sporulating cells

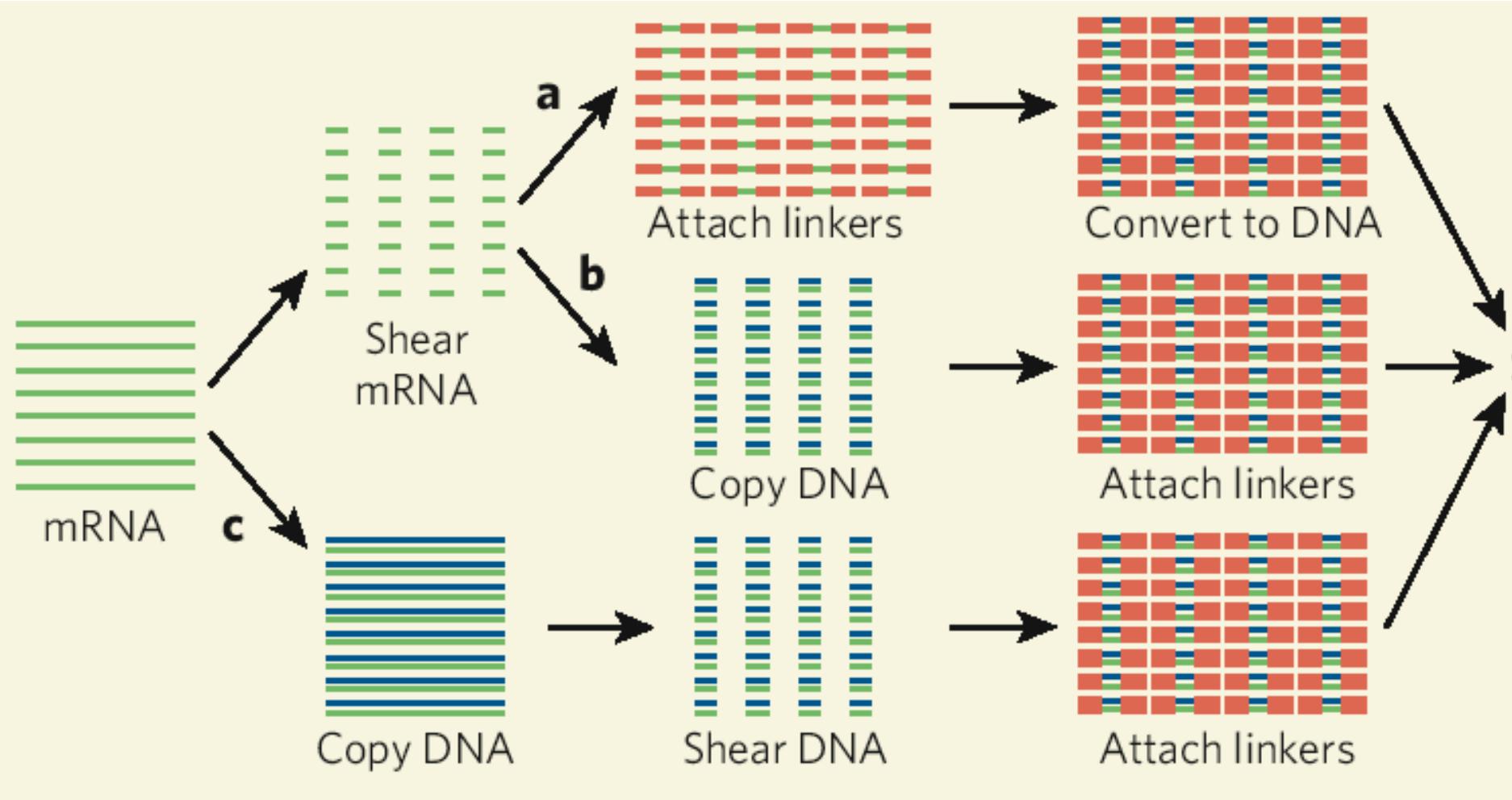
# RNA-seq: sequence cDNA directly

- 1) Isolate RNA
- 2) Make cDNA
- 3) Sequence the cDNA with **next generation** methods
- 4) The more abundant the mRNA, the more sequence data

## Advantages:

- Quantitative, high sensitivity, very low background
- No need to make an array
- Direct identification of the RNA being made
  - Gives info about splicing variants, 5' and 3' ends
  - (no need for hybridization, which doesn't give RNA sequence information)
- Sequencing costs continue to drop

# RNA-seq: which RNAs are expressed, and how much?

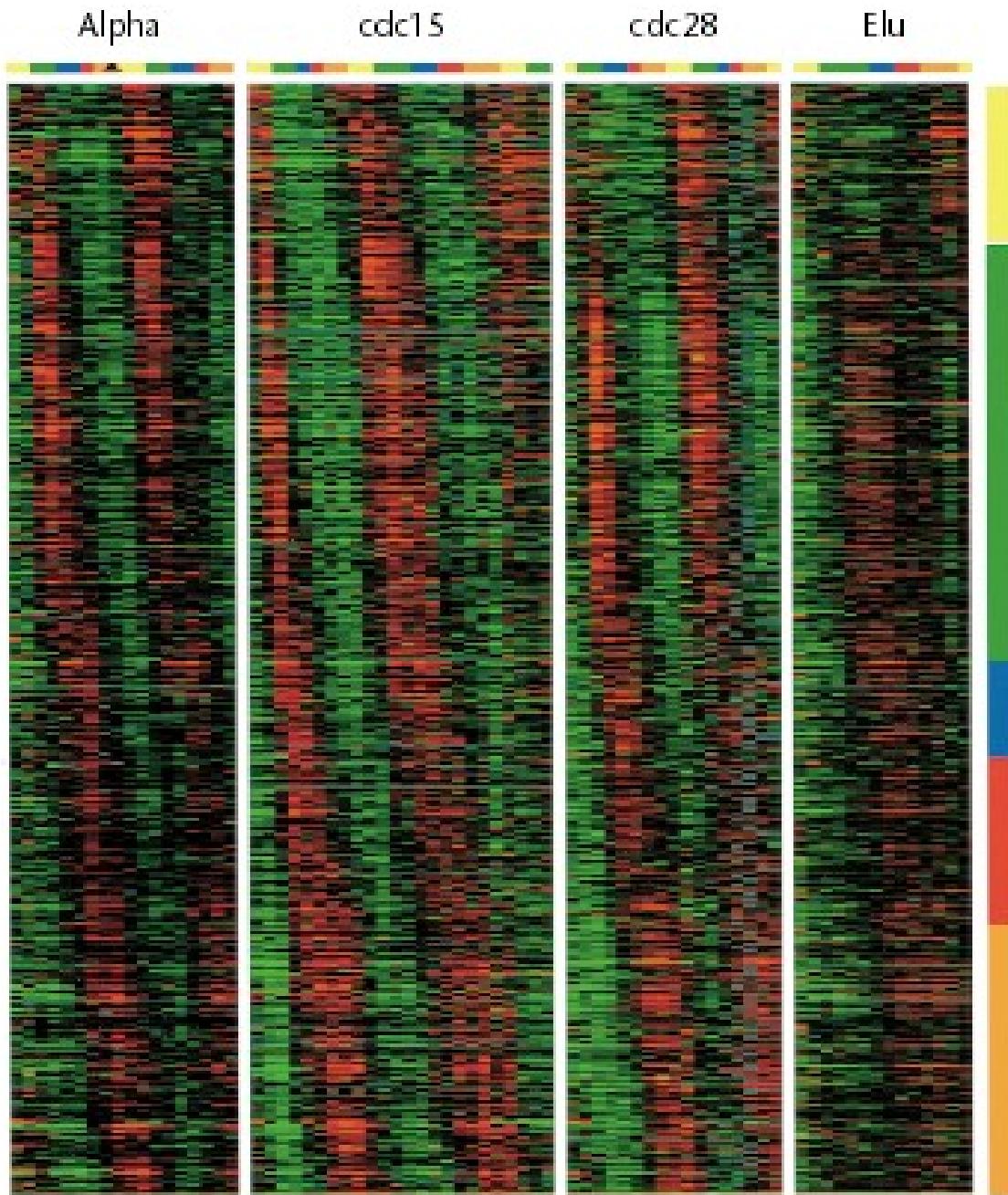


# What to do with data for 1000s of genes?

- 1)Organize data by clustering to see if patterns emerge
- 2)Display data graphically to assist in understanding and hypothesis generation

Each individual gene that is cell-cycle regulated

## Cell synchronization method



All cell cycle-regulated genes

M/G1

G1

S

G2

M

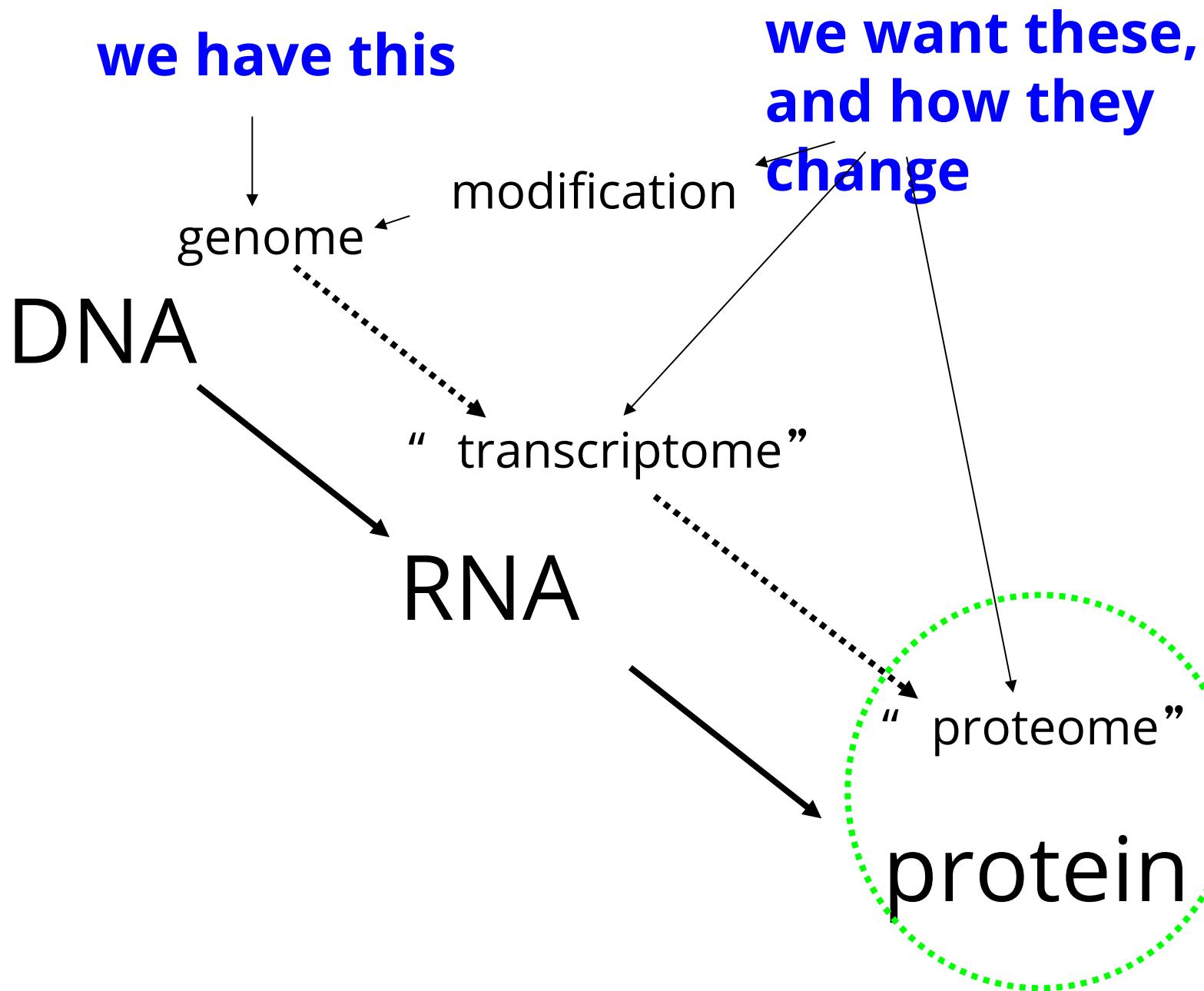
phase in which each gene was expressed at high levels

High mRNA levels



low mRNA levels





# Analysis of the proteome: “proteomics”

- Which proteins are present and when?
- What are the proteins doing?
  - What interacts with what?
    - Protein-DNA interactions (chromatin immunoprecipitation)
    - Protein-protein interactions
  - What is the function of each protein?

Phizicky et al. (2003) “Protein analysis on a proteomic scale” *Nature* **422**, p. 208-215

# How to detect protein expression

- Antibodies to specific proteins (those antibodies need to be available first)
- Specific label on protein *in vivo*, for example GFP to reveal expression in different tissues or subcellular locations
- Specific assay for activity: assuming you have a simple assay already designed

*Above methods arduous for whole proteome*

- Mass spectrometry for direct ID and quantitation

# Defining the human proteome by immunodetection

<http://www.proteinatlas.org/>

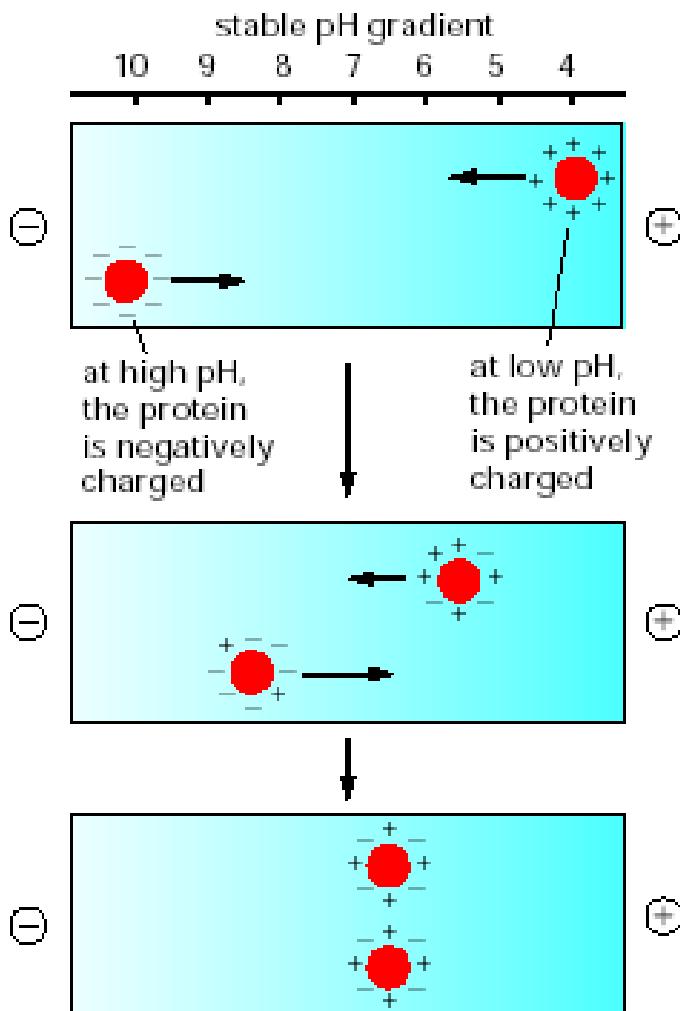
- ~26,000 antibodies to human proteins, targeting ~17,000 different proteins
- 44 major tissues and organs, >13 million tissue-based immunohistochemistry images
- Complemented with RNA-seq analysis of the tissues and organs (to confirm RNAs for detected proteins)
- Cellular proteomes:  
<https://www.proteinatlas.org/humanproteome/cell>
- Cancer proteomes:  
<https://www.proteinatlas.org/humanproteome/pathology>

# Simultaneous detection and identification of all (or most) proteins

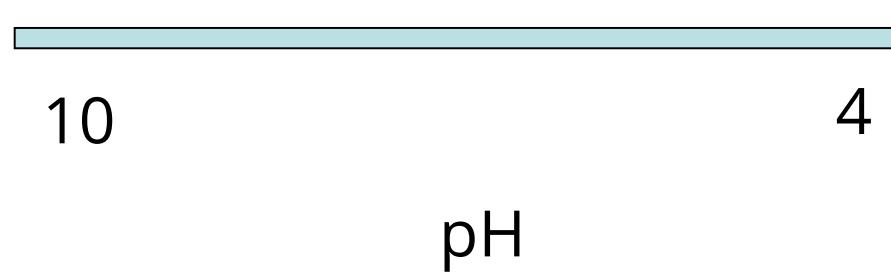
- 2D gel electrophoresis
  - Separate proteins in a given organism or tissue type by migration in gel electrophoresis
  - Identify protein (cut out of gel, sequence or mass-spec)
  - Pattern of spots like a barcode for hi-throughput studies
- Mass spectrometry
  - Separate individual proteins from cell by charge and mass, individual proteins can be identified (need genome sequence information for this)
- Microarray/seq analysis: identify all the DNA or RNA that is bound by a protein

# 2D gel electrophoresis

## 1) Separate proteins by isoelectric point



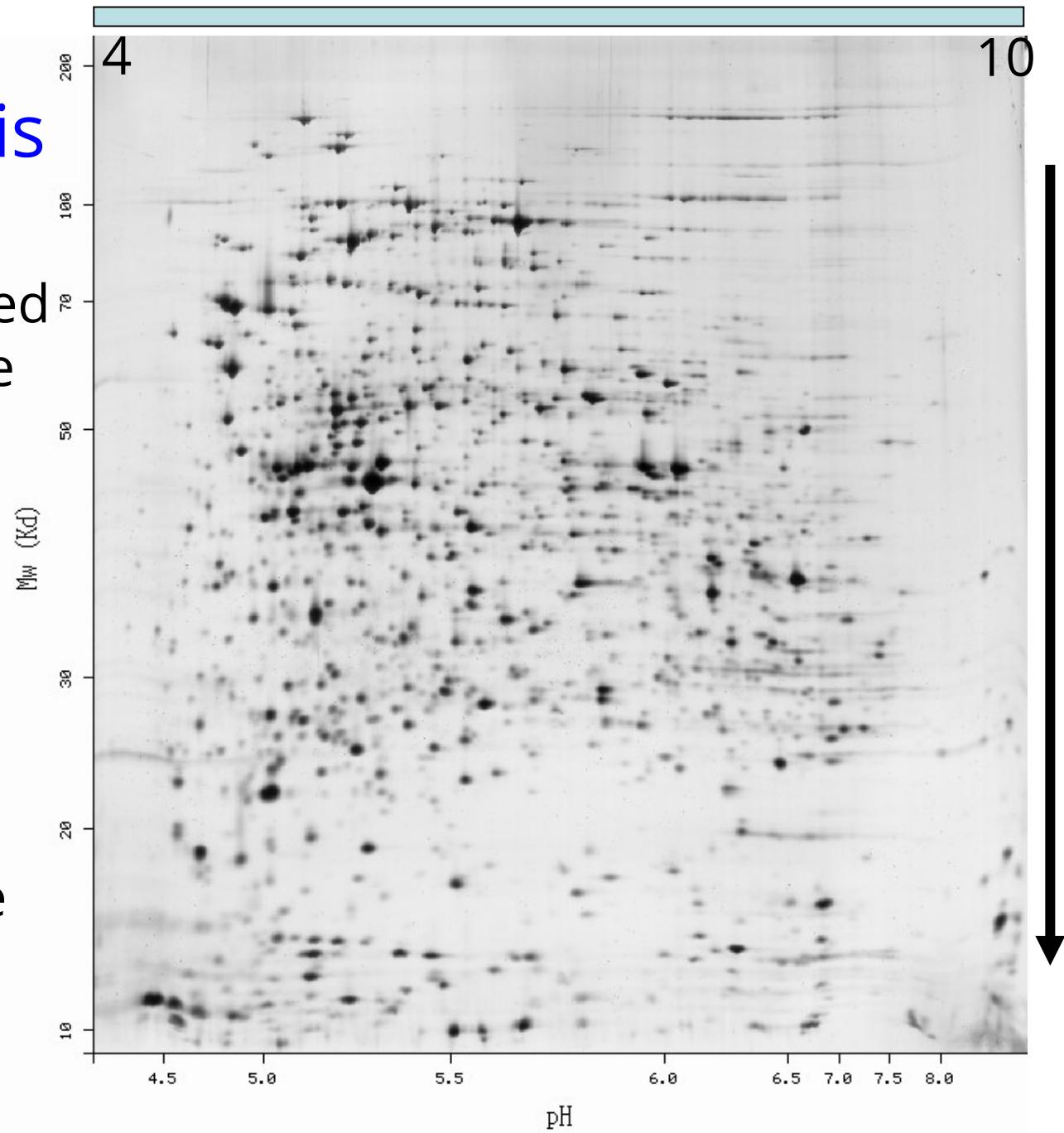
Use a long, narrow gel



The protein shown here has an isoelectric pH of 6.5.

# 2-D gel electrophoresis

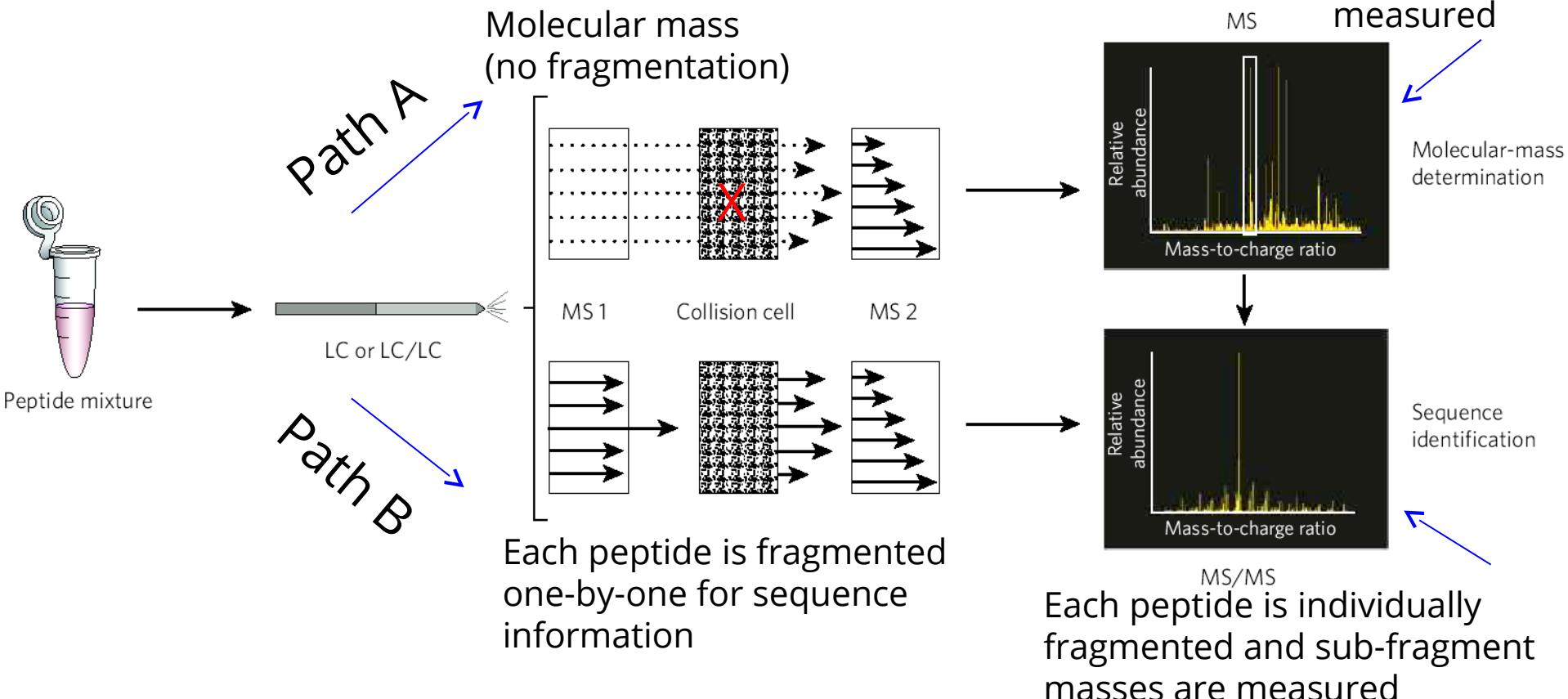
Lay gel containing  
isoelectrically focused  
protein on SDS page  
gel, separate on the  
basis of size



E.coli protein profile  
From swissprot  
database,  
[www.expasy.ch](http://www.expasy.ch)

# Mass spectrometry: identify the proteins from a complex mixture

Liquid chromatography followed by tandem mass spectrometry



From Cravatt *et al.* (2007) “ The biological impact of mass-spectrometry-based proteomics.” *Nature* **450**, p. 991.

# How protein function gets defined

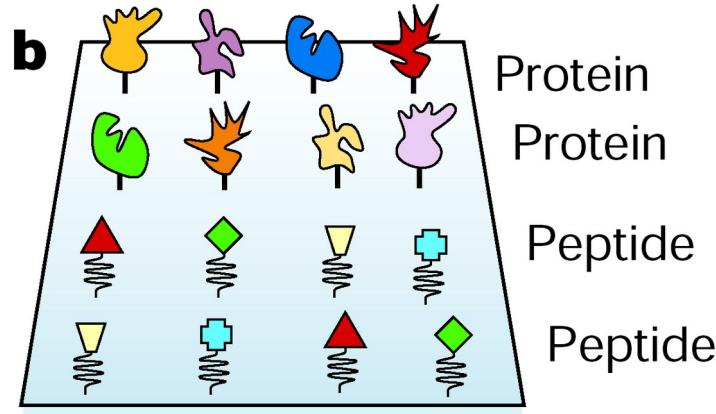
Classical methods define activity of protein, develop an assay for activity

- Biochemistry: use a specific assay to purify a protein or protein complex from a cell, find out the structure and function of the protein *in vitro*
- Genetics: find mutant versions of a protein that have altered or lost activity, observe the phenotype of the organism with that mutation, obtain additional mutant genes that may interact with protein of interest, etc.

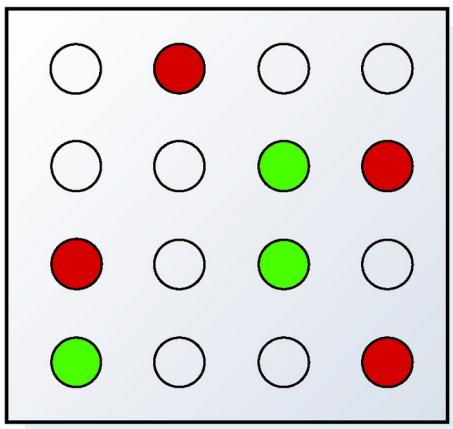
# Protein structure and function analysis from a proteomic approach

- Massively parallel screens for activity--protein arrays
- Protein-DNA interactions: identifying binding sites for DNA-binding proteins, study regulation of gene expression
- (Structural genomics: solve structures of as many open reading frame peptides as possible)

# Protein arrays for function



Protein probes  
Nucleic acid probes  
Drug probes  
Enzymes



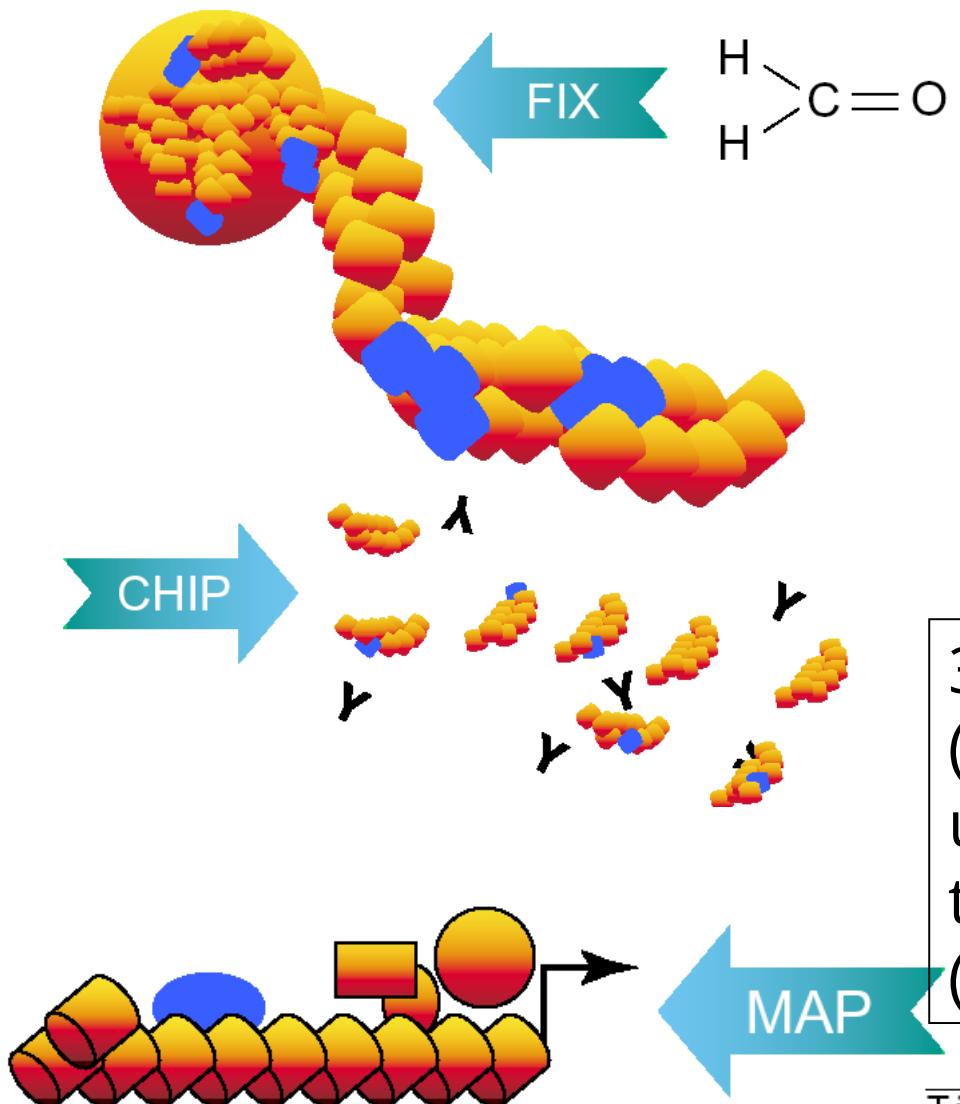
Protein binding properties  
Pathway building  
Drug discovery

Proteins immobilized,  
usually by virtue of a tag  
sequence (6 x histidine tag,  
biotin, etc.)

Probe all proteins  
at once for a  
specific activity

Structural diversity and  
complexity of proteins  
means **not all proteins**  
**are active** in this form

# "Chromatin ImmunoPrecipitation" (ChIP)



1) Grow cells, add formaldehyde to cross-link everything to everything (including DNA to protein)

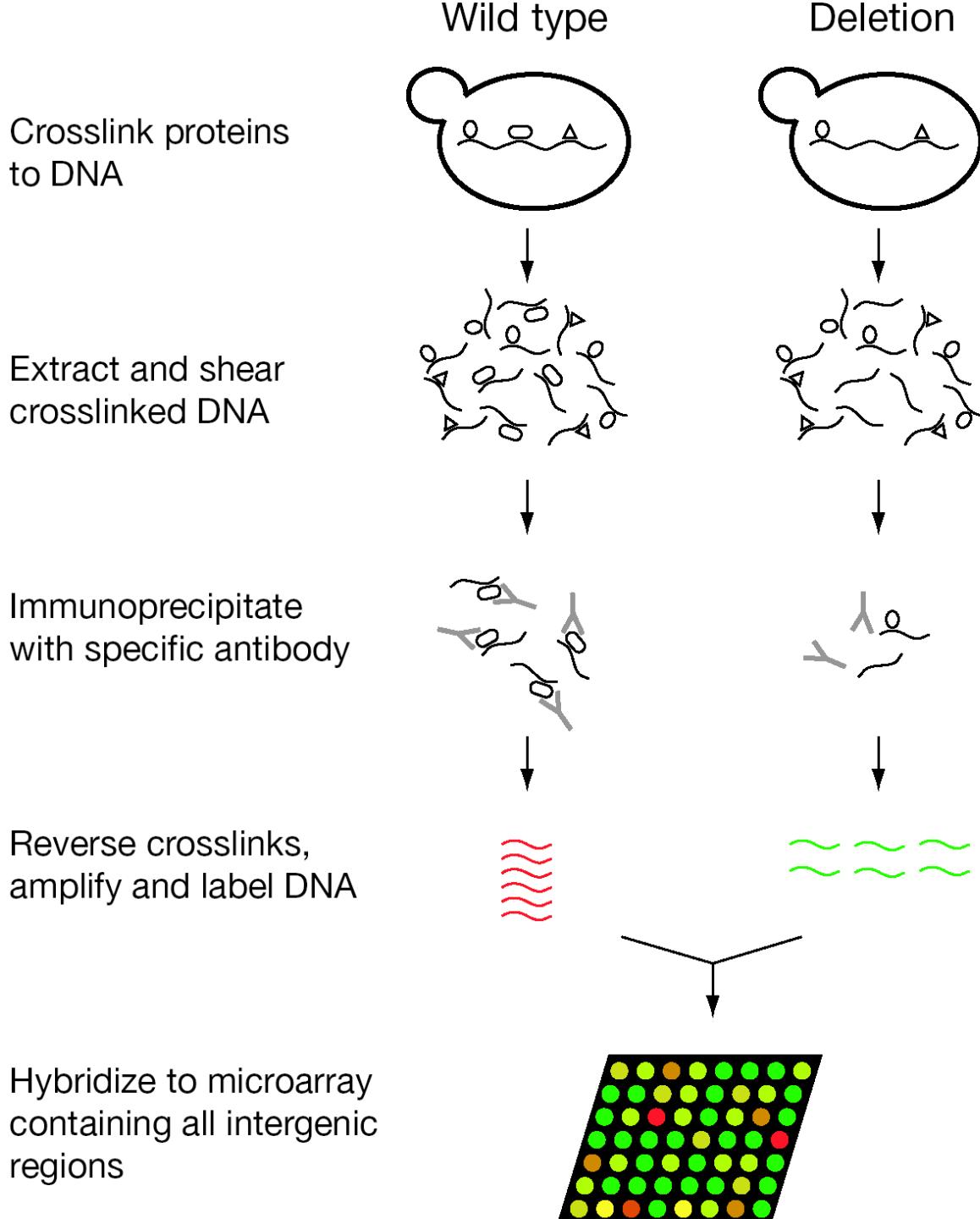
2) Lyse cells, break up DNA by shearing

3) Retrieve protein of interest (and the DNA it is bound to) using specific antibody to that protein (immunoprecipitation)

4) Determine presence of DNA by quantitative PCR

# Genome-wide ChIP

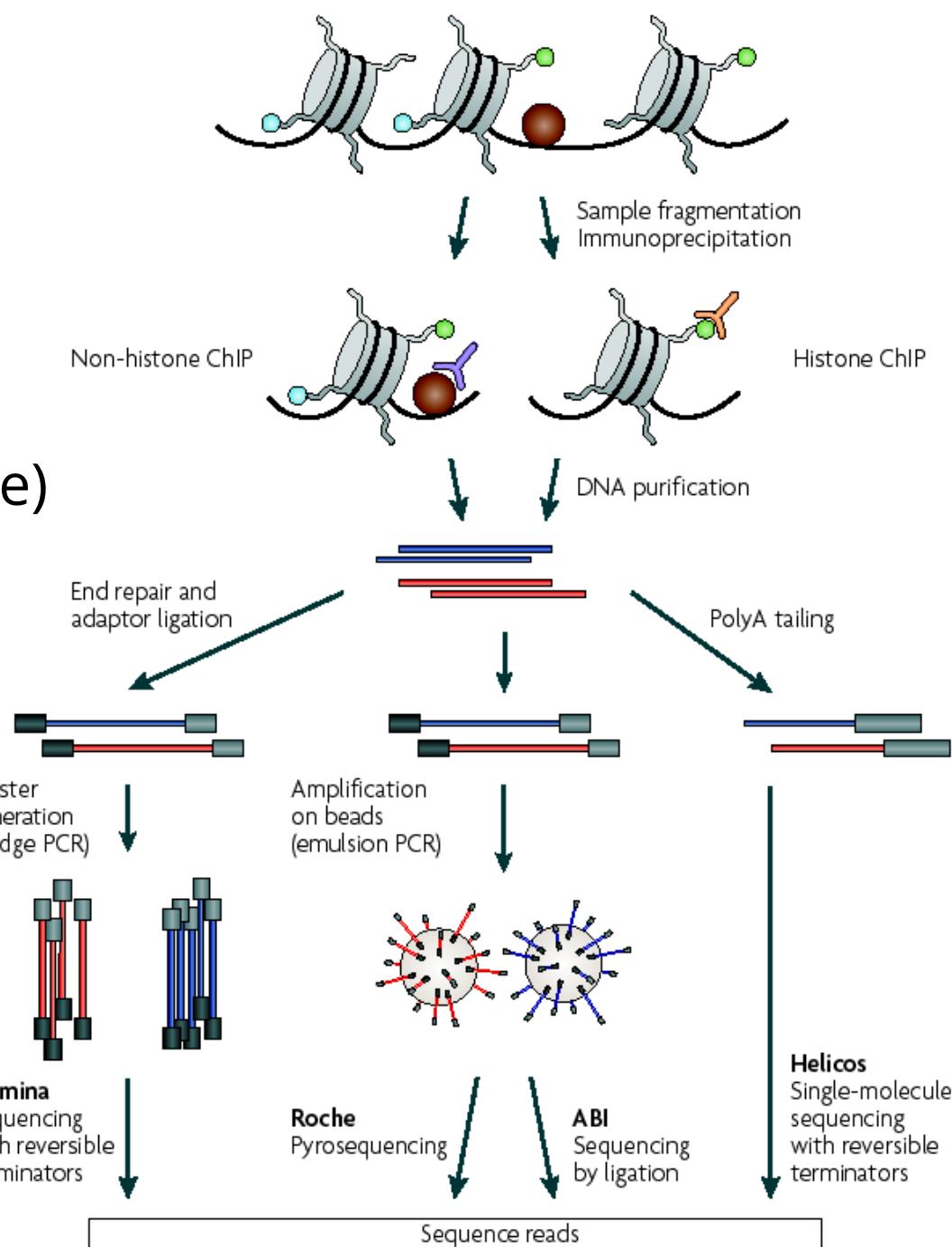
PCR, label with fluorescent dyes



# ChIP-seq

Sequence the IP DNA using  
“next generation”  
sequencing techniques

- High resolution (single base)
- Low noise
- Coverage not limited by array



## APPLICATIONS OF NEXT-GENERATION SEQUENCING

### ChIP-seq: advantages and challenges of a maturing technology

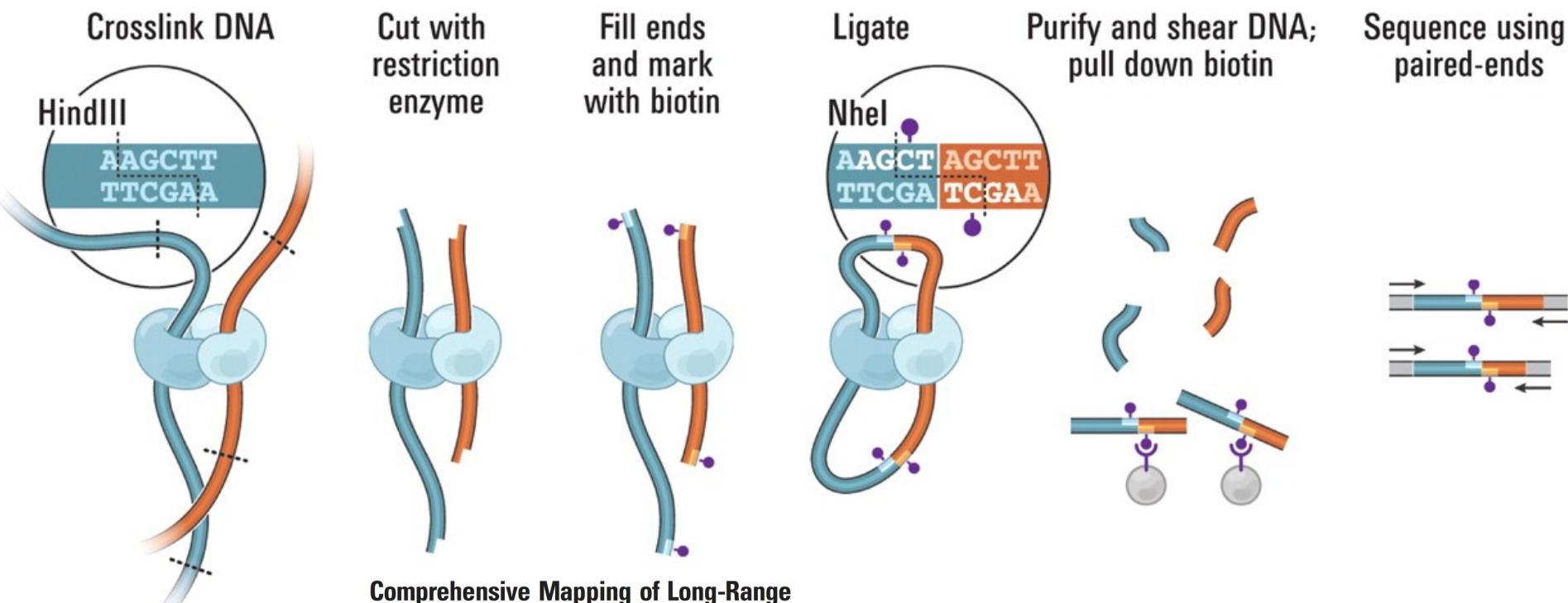
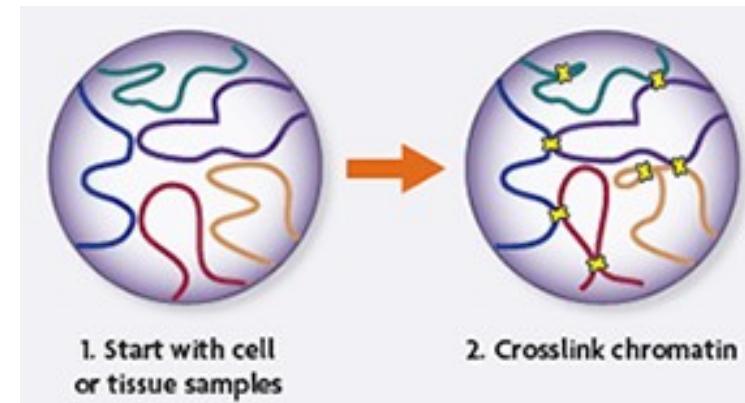
Peter J. Park

NATURE REVIEWS | GENETICS VOLUME 10 | OCTOBER 2009 | 669

Figure 1 | Overview of a ChIP-seq experiment. Using chromatin immunoprecipitation

# Chromatin conformation: which regions of DNA are close to each other

Hi-C: proximity based ligation, followed by massively parallel sequencing



## Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,<sup>1,2,3,4,\*</sup> Nykje L. van Berkum,<sup>5\*</sup> Louise Williams,<sup>3</sup> Maxim Imakaev,<sup>2</sup> Tobias Ragozy,<sup>6,7</sup> Agnes Telling,<sup>6,7</sup> Ido Amit,<sup>1</sup> Bryan R. Lajoie,<sup>3</sup> Peter J. Sabo,<sup>8</sup> Michael O. Dorschner,<sup>9</sup> Richard Sandstrom,<sup>8</sup> Bradley Bernstein,<sup>1,9</sup> M. A. Bender,<sup>10</sup> Mark Groudine,<sup>6,7</sup> Andreas Gnirke,<sup>3</sup> John Stamatoyannopoulos,<sup>8</sup> Leonid A. Mirny,<sup>2,11</sup> Eric S. Lander,<sup>1,12,13†</sup> Job Dekker,<sup>5‡</sup>

# The ENCODE project: putting it all together

ENCyclopedia Of functional DNA Elements in the human genome

<https://www.encodeproject.org/>

<http://genome.ucsc.edu/ENCODE/>

Systematic mapping of

- Regions that are transcribed
- Transcription factor binding sites in DNA
- RNA binding proteins
- Chromatin structure
- DNA modifications
- Histone modifications
- ...and much much more

In 2019: >14,000 datasets

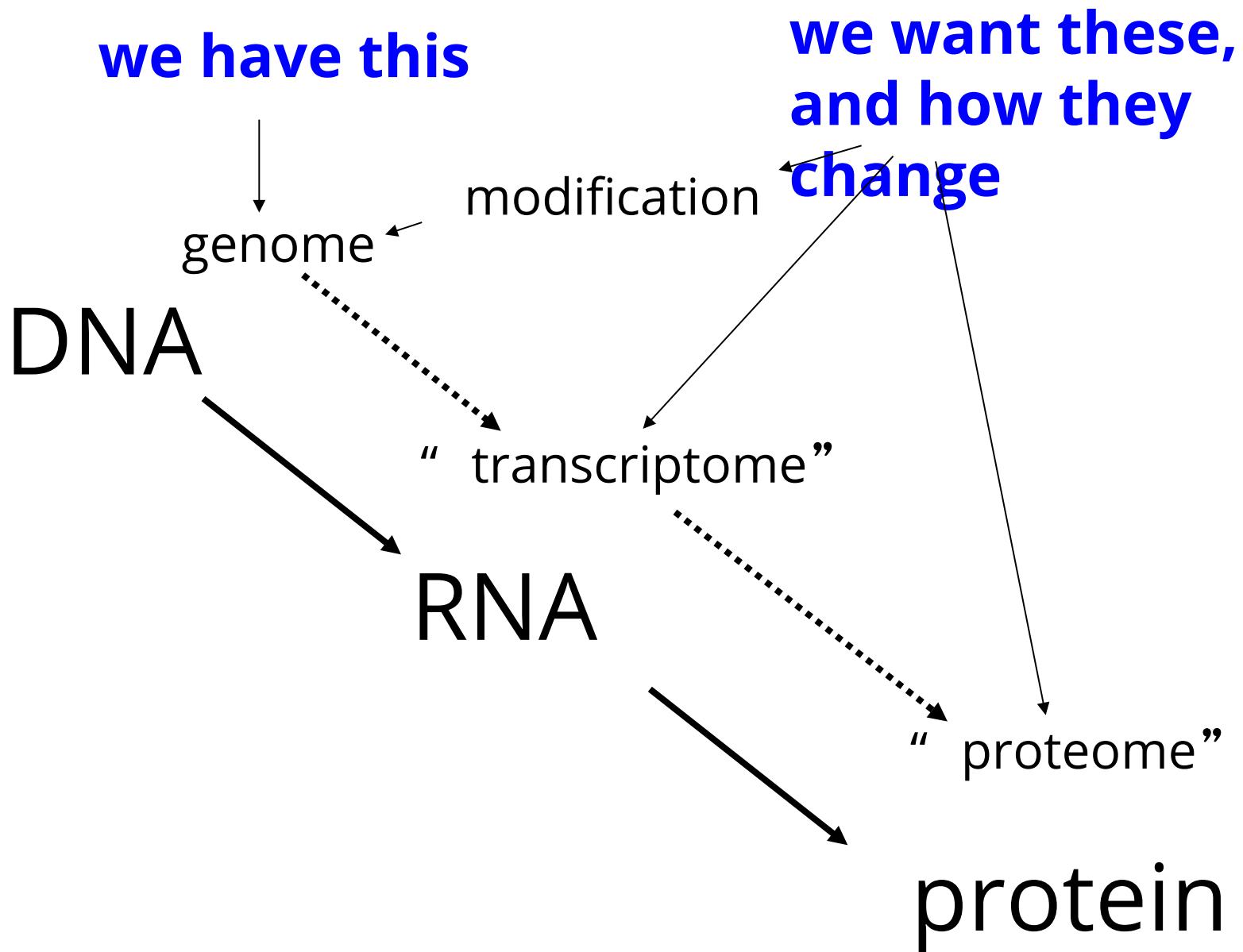
# The ENCODE project

- ~80% of the human genome participates in an RNA or chromatin associated event, in at least one cell type
- RNA synthesis correlates with chromatin and transcription factor binding: promoters account for most RNA variation
- More disease associated Single Nucleotide Polymorphisms (SNPs) are in non-coding functional elements than in protein coding genes, e.g. affecting transcription factor binding

Not all cell types have been assayed

Not all transcription factors have been assayed

So there is more to be done





# Face up to false positives

Macarthur (2012)  
*Nature* **487** p.427

Dealing with very large data sets can suggest erroneous conclusions

- 1) Large data sets mean unusual (statistically rare, insignificant) events crop up often. Statistical analysis helps to assign significance (or lack of it)
- 2) Error/system bias often occurs in high-throughput methods – so the novice gets burned

Stringent quality control and standards are essential!