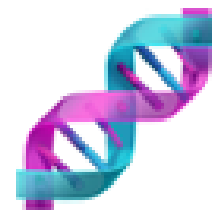


Sequencing and transcriptomics



Daniel Fürth

Assistant professor
SciLifeLab/Uppsala University

 furthlab.xyz

 [@furthlab](https://twitter.com/furthlab)



Learning goals

Course Syllabus:  [3MG022](#)

Knowledge and understanding

“Explain current techniques for research, diagnostics and treatment of genetic diseases including cancer”

Competence and skills

“independently analyse, process and formulate relevant scientific questions...”

Judgement and approach

“...ability to critically evaluate and appraise scientific research...”



Learning goals

Topics covered today:

-  History: from Sanger sequencing to single-cell methods
-  Transcriptomic analysis
-  Different types of RNAs and their functions

 Sanger sequencing  single-cell methods

Separation chemistry

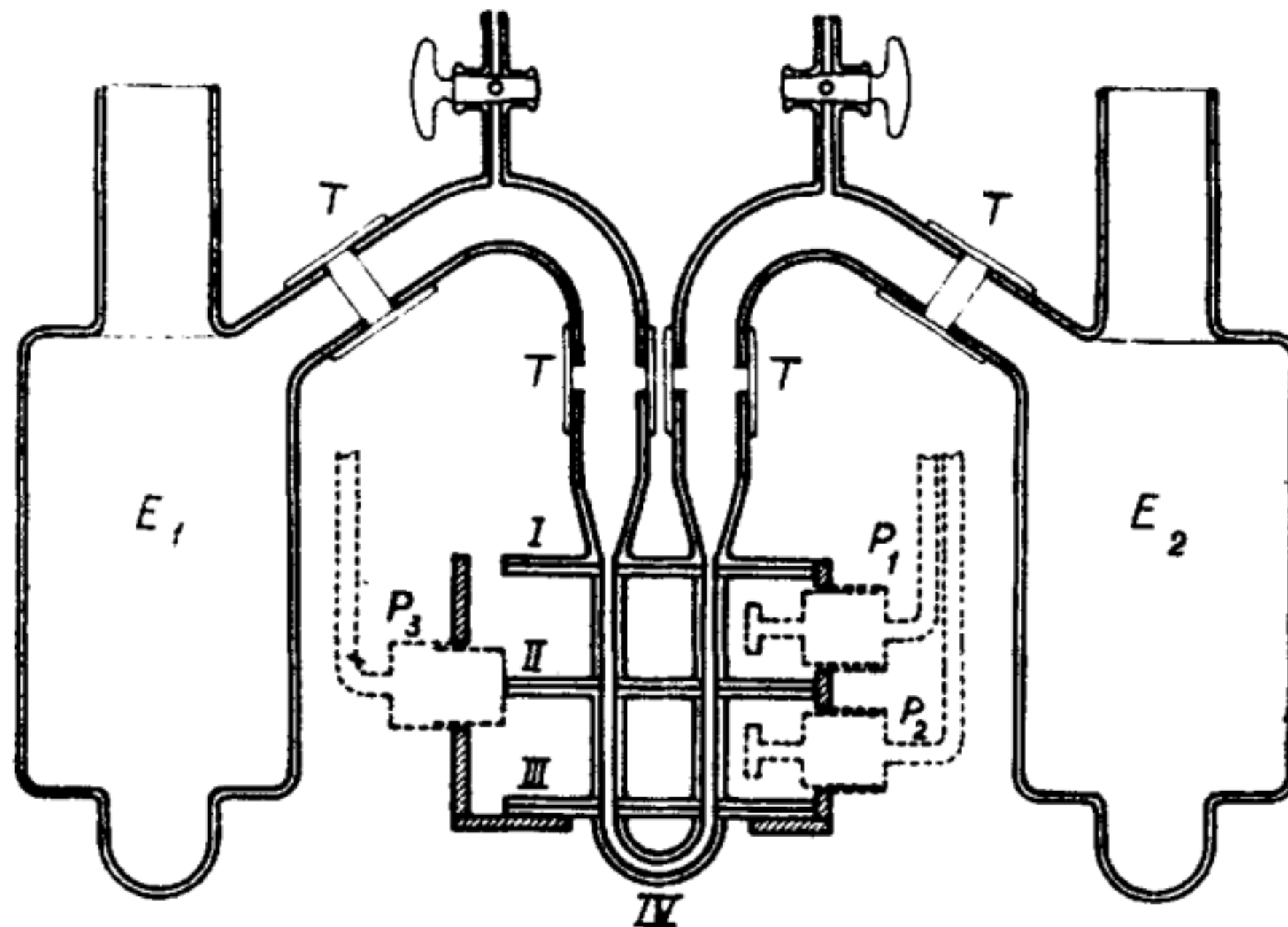


FIG. 3.—Apparatus for electrophoretic analysis.

E_1 , E_2 represent the electrode tubes (the larger type), T rubber tubings; I, II, III, IV, the four U-tube sections, of which II and III can be moved with the pneumatic arrangement P_1 , P_2 , P_3 .

Electrophoretic machine as designed by Tiselius

Separation chemistry

Polyacrylamide gel electrophoresis (PAGE)

- Length
- Charge
- Confirmation

Molecules can be run in their “*native state*” preserving higher-order structure.

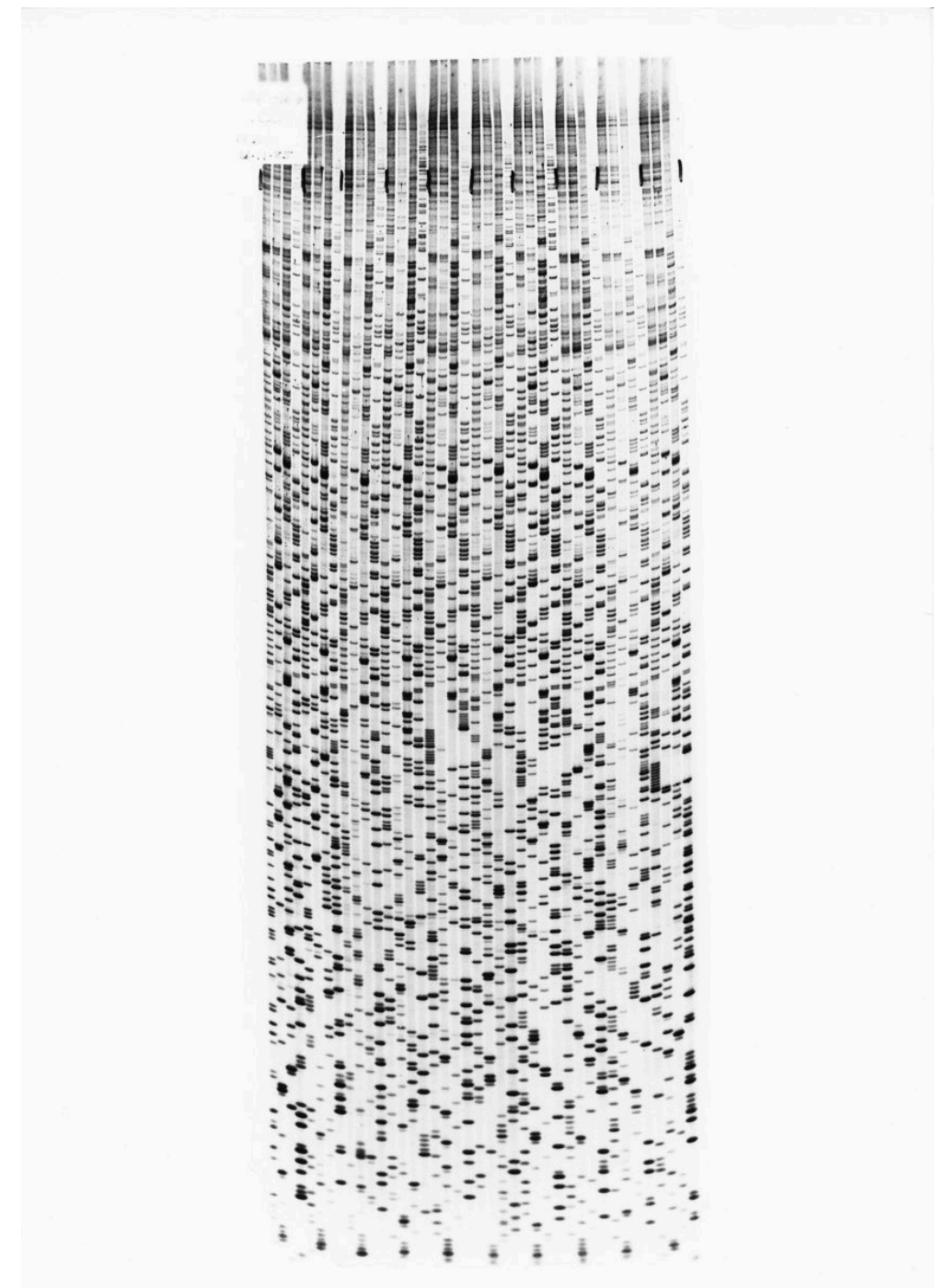
Or a chemical denaturant can be added to remove structure (urea for .

Native vs **non-native** PAGE.

Pore size controlled by % of acrylamide.

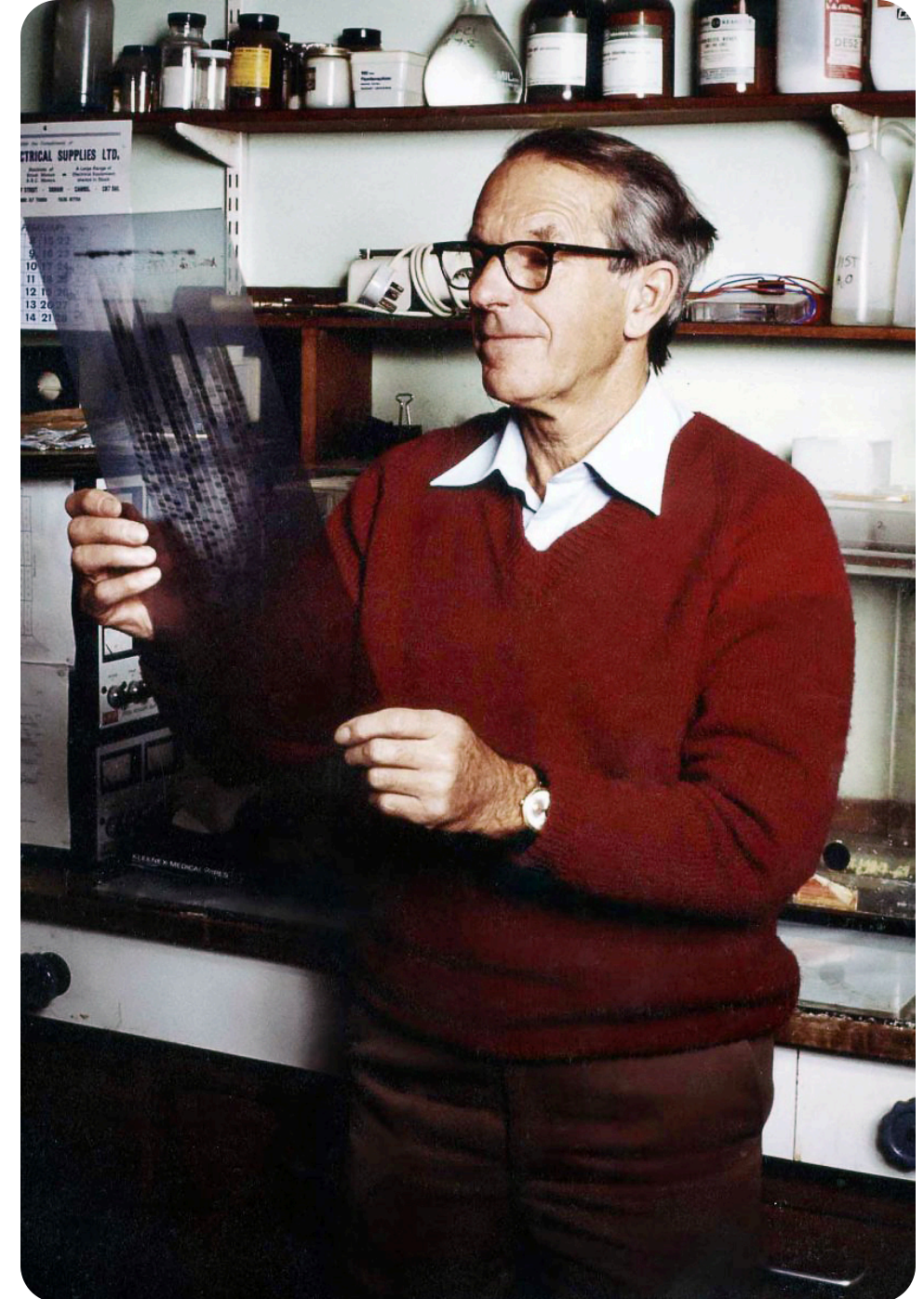
 **Resolution:**

100-1000 nt down to single-nucleotide.



Fredrick Sanger

- One of the few who received two Noble Prizes.
 -  1958
 - “structure of proteins, especially that of insulin”
 -  1980
 - “determination of base sequences in nucleic acids”



🇬🇧 Medical Research Council Laboratory of Molecular Biology

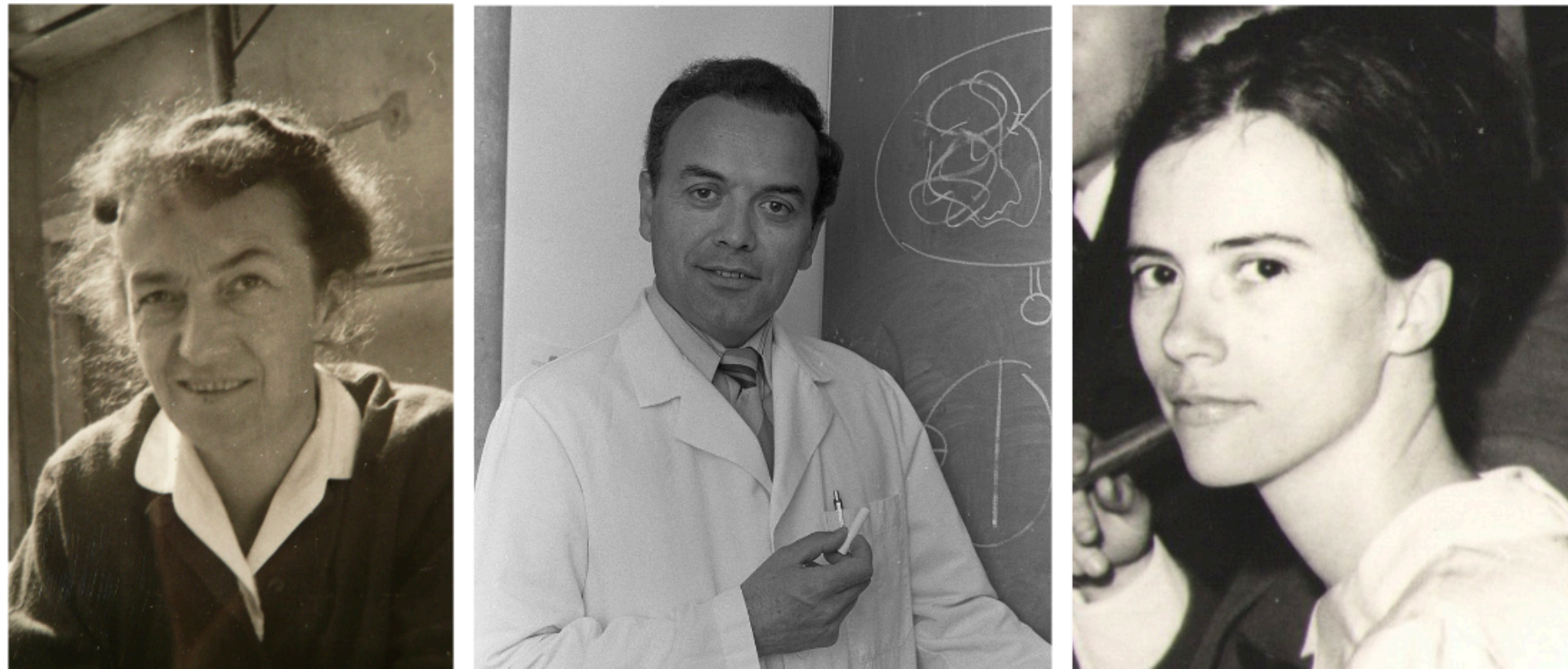


Left MRC Laboratory of Molecular Biology, 1962. **Right** Left to right: Hugh Huxley, 🏅 John Kendrew ('62), 🏅 Max Perutz ('62), 🏅 Francis Crick ('62), 🏅 **Fred Sanger ('58,'80)** and 🏅 Sydney Brenner (2002).

🔧🔧🔧 Restriction enzymes 🧬

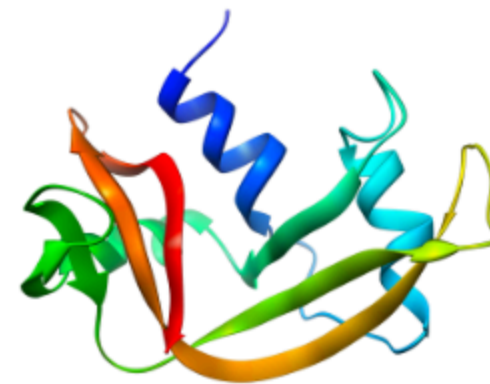
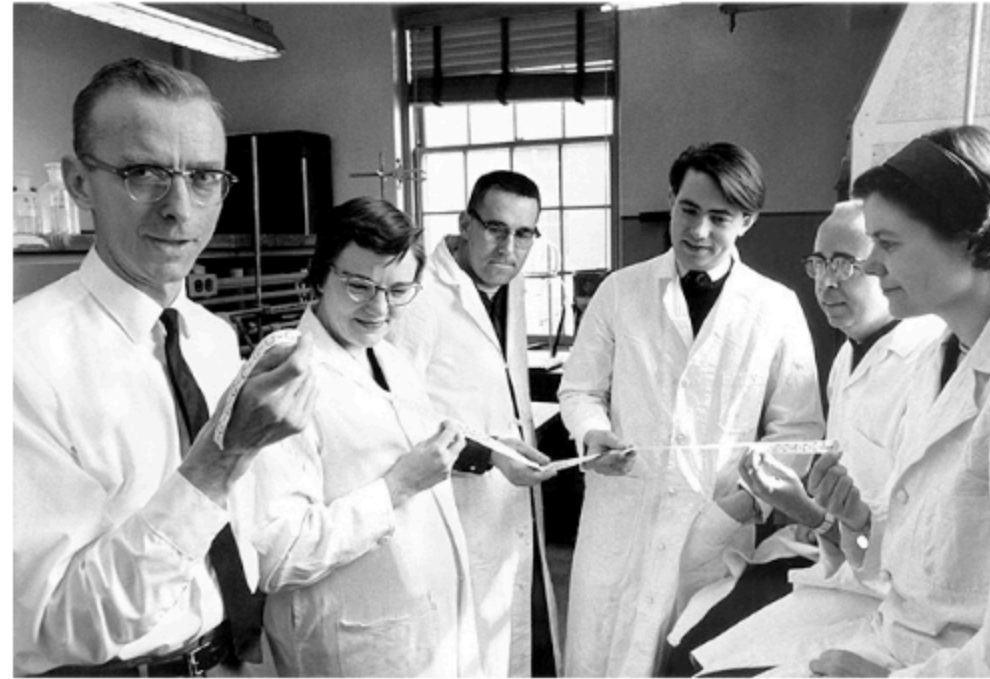
🇺🇸 Restriction and modification enzymes 🇨🇭

🏆 **Noble Prize:** 🇺🇸 Hamilton O. Smith, 🇺🇸 Daniel Nathans, and 🇨🇭 Werner Arber.

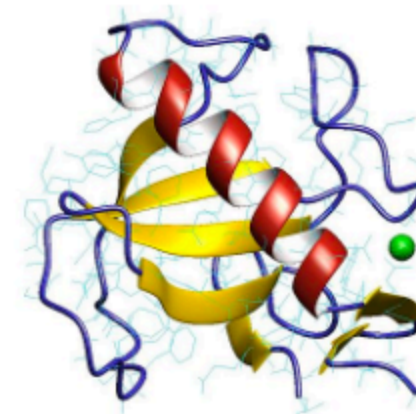


Left to right: 🇨🇭 Grete Kellenberger, 🇨🇭 Werner Arber, 🇨🇭 Daisy Dussoix

Structure of a Ribonucleic Acid



RNase A



RNase T1

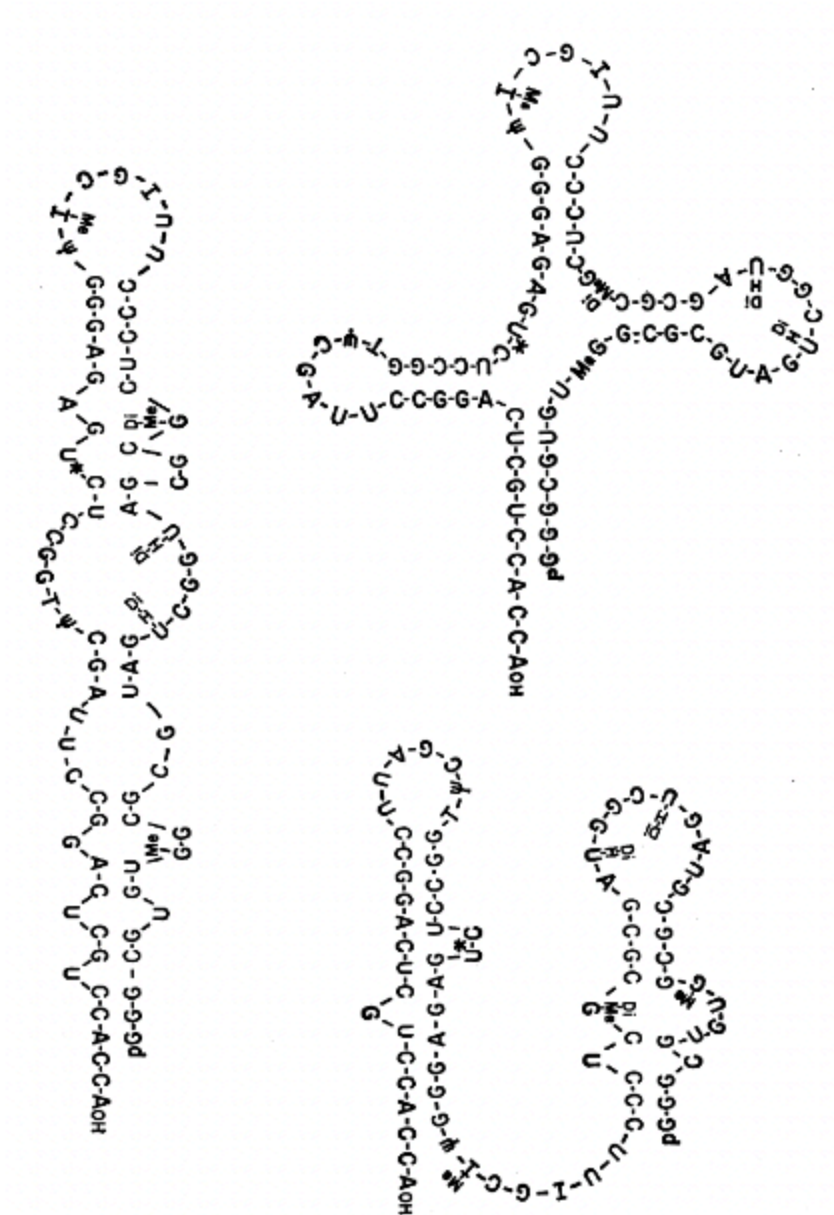


Fig. 2. Schematic representation of three conformations of the alanine RNA with short, double-stranded regions.

Structure of a Ribonucleic Acid

RNase A (Bovine pancreas ribonuclease):

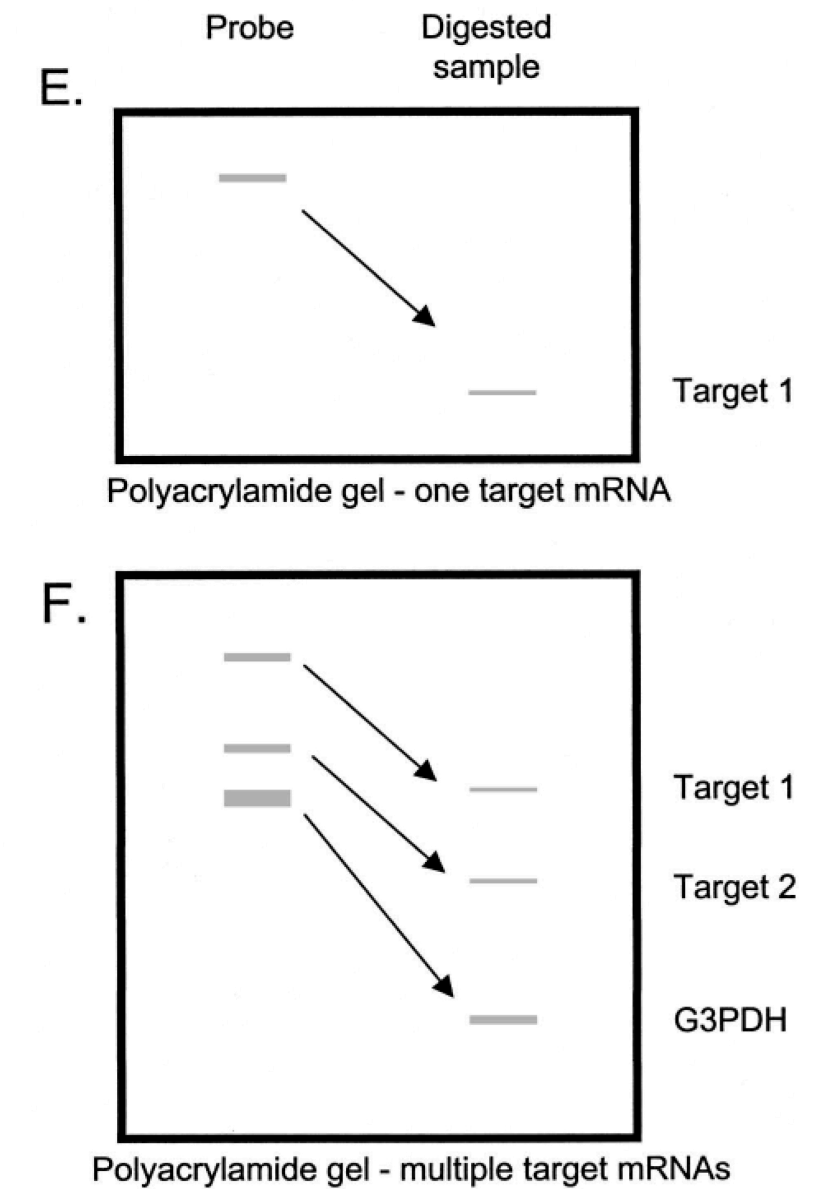
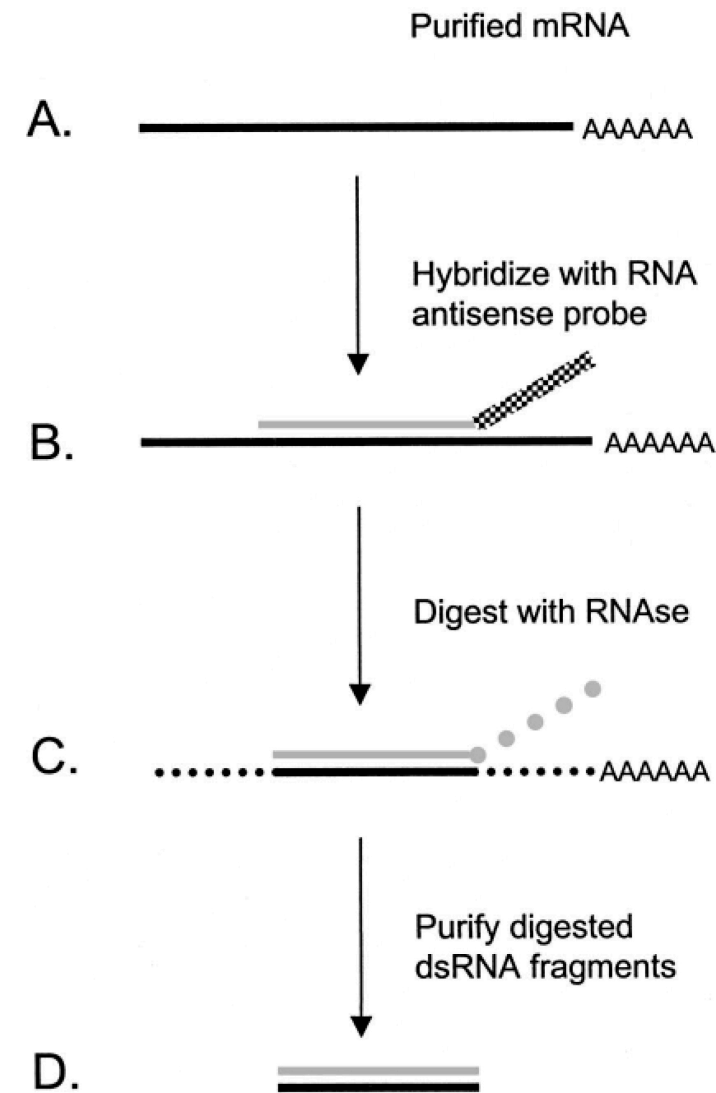
- Only active on RNA.
- Very sturdy, survives boiling etc 🔥.
- Cleaves 3'-end of unpaired C and U residues (pyrimidine).

RNase T1 (Taka-Diastase ribonuclease T1):

- Single-stranded (ssDNA or RNA) specific exonuclease
- Catalyzes the removal of nucleotides from linear single-stranded DNA or RNA in the 3' to 5' direction

Nuclease Protection Assay (NPA)

A proto sequencing method.



Two-dimension partition sequencing

A proto sequencing method.

- First seperate on *isoelectric point* (pI)
 - pH @ molecule has **no** net electric char
- Then seperate on *molecular weight*.

380

F. SANGER, G. G. BROWNLEE AND B. G. BARRELL

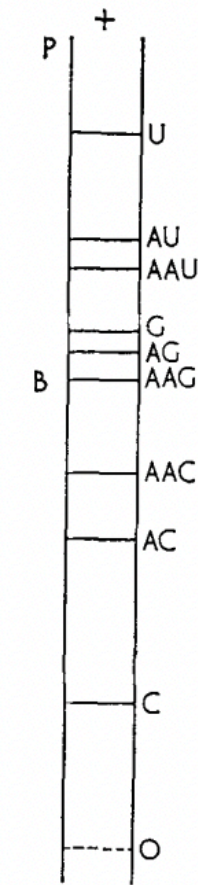
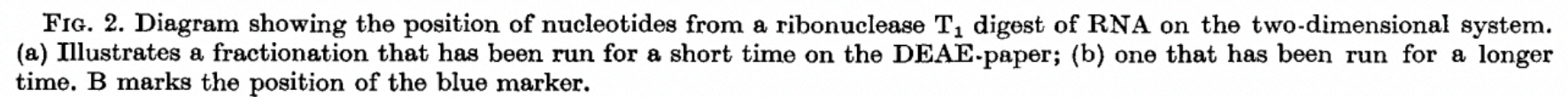


FIG. 1. Diagram showing the position of products from pancreatic digests of nucleotides from ribonuclease T_1 digests on ionophoresis at pH 3.5. B and P show the positions of the blue and pink markers, respectively.



Two-dimension partition sequencing

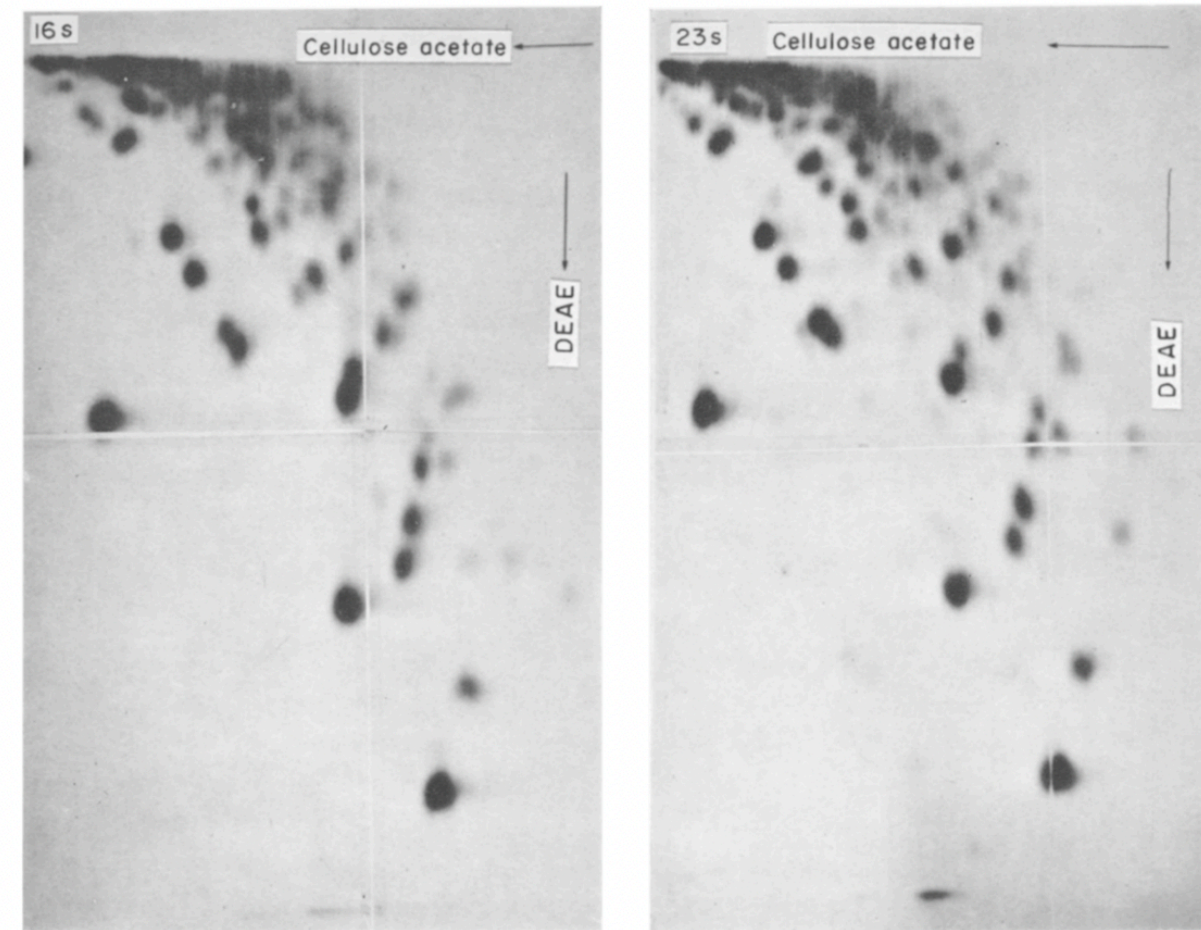


PLATE I

PLATE II

PLATE I. Radioautograph of a two-dimensional fractionation (cellulose acetate pH 3.5, 3 kv, 2 hr; DEAE-paper pH 1.9, 1.5 kv, 4 hr) of a ribonuclease T_1 digest of the 16 s component of ribosomal RNA.

PLATE II. Radioautograph of a two-dimensional fractionation (as in Plate I) of a ribonuclease T_1 digest of the 23 s component of ribosomal RNA.

dNTP incorporation has base specificity

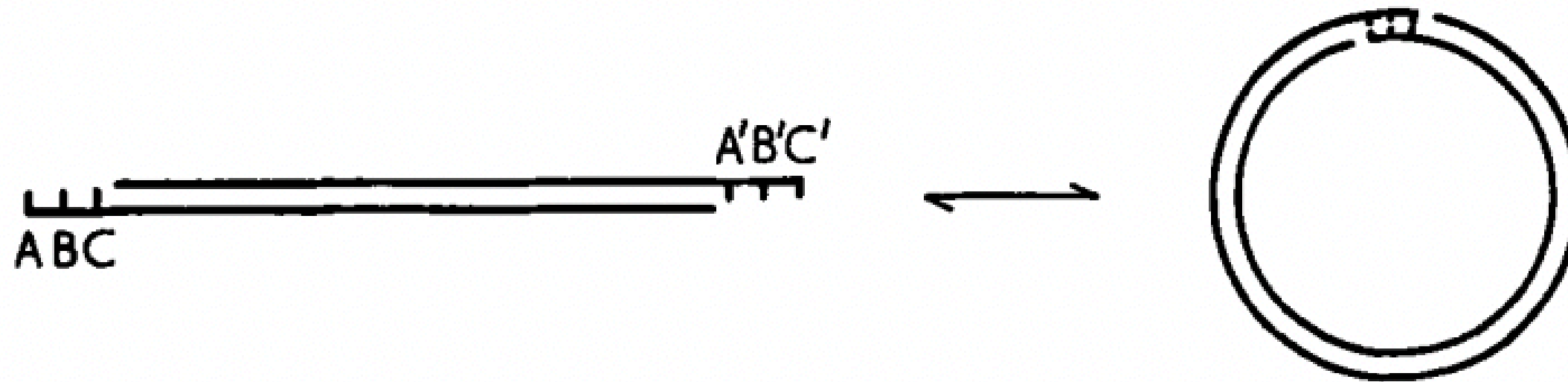


FIG. 1. Cohesion of λ DNA

A possible structure for the cohesive ends, in which the two ends are held together by base pairing between A and A', B and B', C and C'.

The birth of the “-”-*approach* of Sanger.

First sequence of a single gene

- Bacteriophage MS2
- Circular ssDNA genome.

82

NATURE VOL. 237 MAY 12 1972

Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein

W. MIN JOU, G. HAEGEMAN, M. YSEBAERT & W. FIERS

Laboratory of Molecular Biology and Laboratory of Physiological Chemistry, State University of Ghent, Belgium

By characterization of fragments, isolated from a nuclease digest of MS2 RNA, the entire nucleotide sequence of the coat gene was established. A "flower"-like model is proposed for the secondary structure. The genetic code makes use of 49 different codons to specify the sequence of the 129 amino-acids long coat polypeptide.

DETAILED knowledge of the structure of a viral gene should provide a firmer basis for explaining the mechanism, specificity and regulation of translation. Because the genes of RNA bacteriophages are part of the viral genome, their structures may also have evolved for optimal functioning in replication and encapsulation.

The RNA of the closely related phages f2, MS2 and R17, and especially the cistron coding for the coat protein, has been used extensively for *in vitro* study of protein synthesis¹. ³²P-labelled bacteriophage MS2 RNA was promising as a material for investigating the primary structure of this viral messenger². By using limited enzymatic digestion of the RNA at low temperatures and separating the products on polyacrylamide gels, we isolated RNA-fragments of limited length, suitable for direct sequence analysis³. Sanger *et al.*⁴ were the first to link nucleotide sequences with specific functions in translation, isolating a "hairpin", 57 nucleotides long, from the R17 coat protein cistron, from which they could deduce a set of codons used in the genetic message. Argetsinger-Steitz⁵ identified a hairpin containing the initiating AUG-codon, while Nichols⁶ determined the nucleotide sequence of another hairpin, which contained the coat termination signal; information on two more hairpin-regions was also reported⁷. Working with the phage MS2, we identified and characterized similar coat cistron hairpins⁸. These sequences were extended⁹, and we can now present the entire sequence of the coat protein gene.

Pure Fragments from the Coat Gene

All RNA fragments used for sequencing the coat protein cistron originated from ribonuclease T₁ hydrolysis of uniformly labelled ³²P-MS2 RNA, and subsequent separation on neutral polyacrylamide gels. Two sets of conditions were used: 1° hydrolysis with 1 U of enzyme per 20 µg of RNA for 30 min at 0° C and separation on a 12% gel; and 2° with 1 U per 100 to 200 µg of RNA for 30 min at 0° C, followed by separation

on a 6% gel. Further details and typical patterns of both separations have been published before^{10,11}.

As the amino-acid sequence of the coat polypeptide is known¹²⁻¹⁵ (albeit incorrectly as it turned out), a nucleotide sequence can be deduced for this gene, leaving open the many

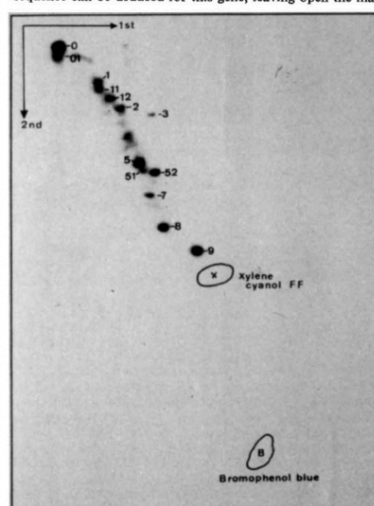


Fig. 1 Two-dimensional separation of band B4. The RNA of band B4 was extracted and subjected to two-dimensional gel electrophoresis. The arrows indicate the directions of the run: the first dimension was on an 8% gel in 0.025 M citric acid+6 M urea; subsequent separation was on a 16% gel in 0.04 M Tris-acetate, pH 8. The positions of the dye markers xylene cyanol FF (X) and bromophenol blue (B) are marked on the figure. Spots b1, b2 and b5 correspond to different regions of the coat protein cistron (Table 2). Spot b12 of this particular separation is a rarely observed fragment starting around the GGCGGA sequence (coding for amino-acids 13-14) and going till arrow 14 in Fig. 2.

© 1972 Nature Publishing Group

Sanger's Plus/Minus sequencing

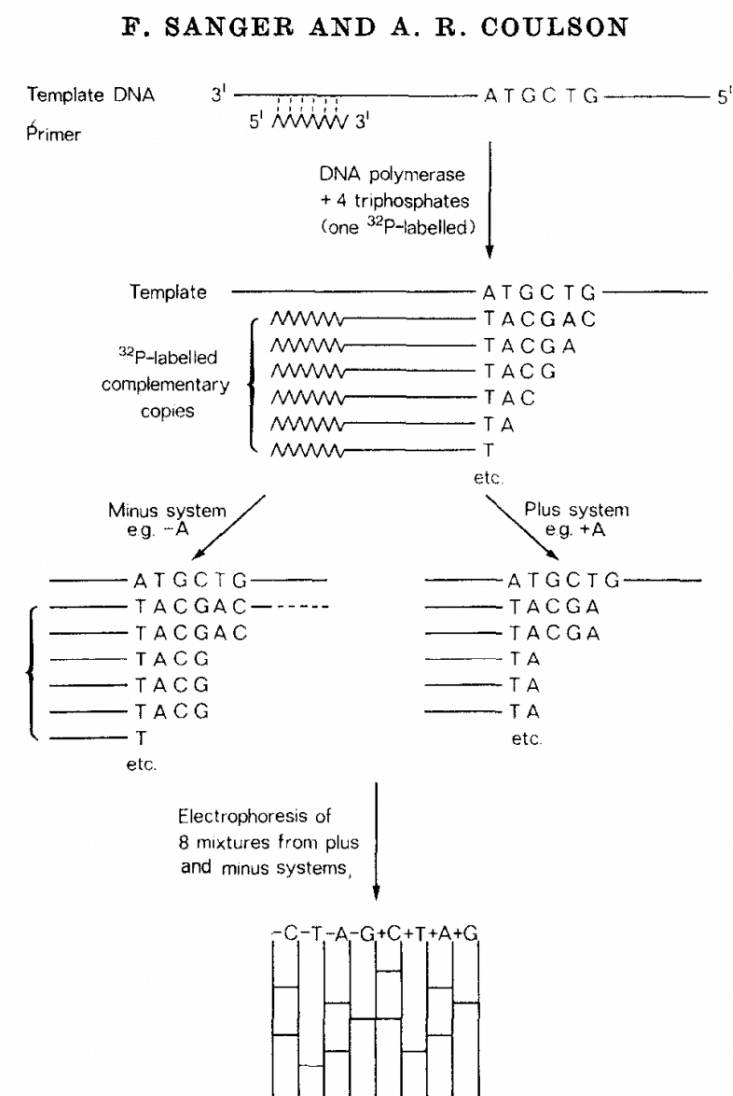


Fig. 1. The principle of the method.

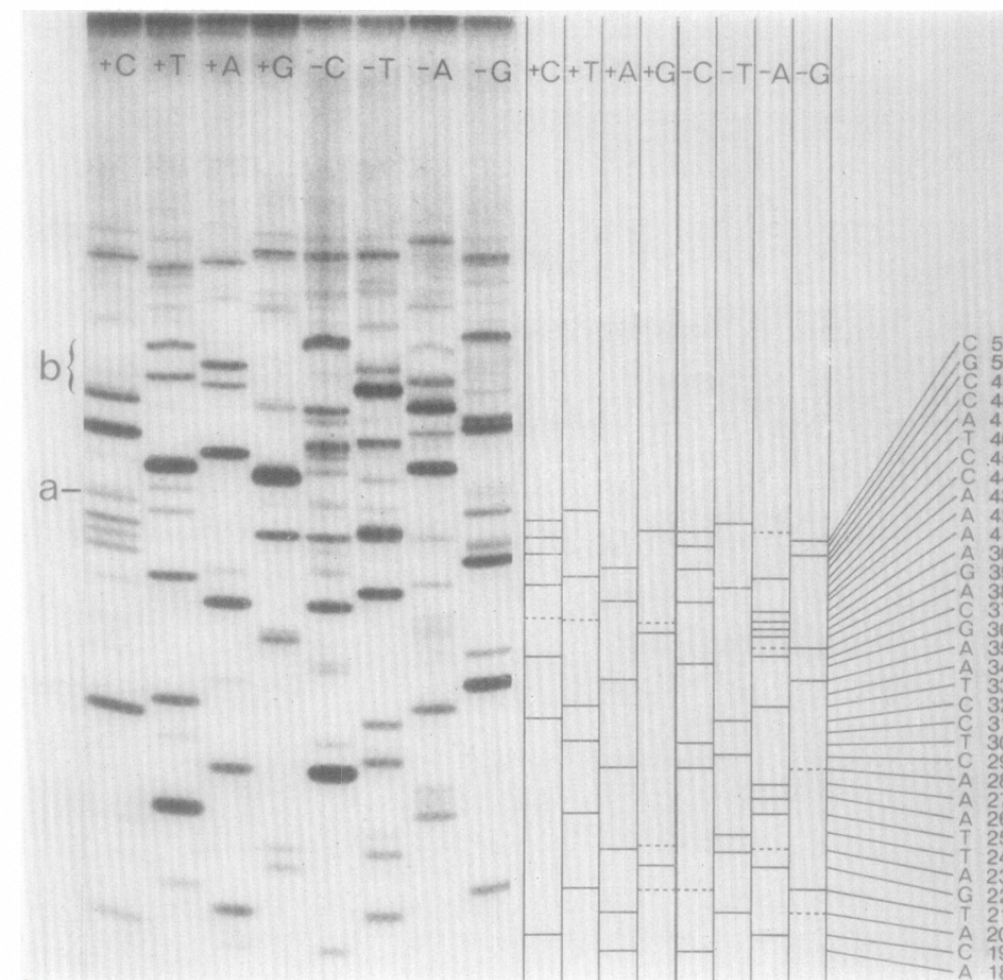
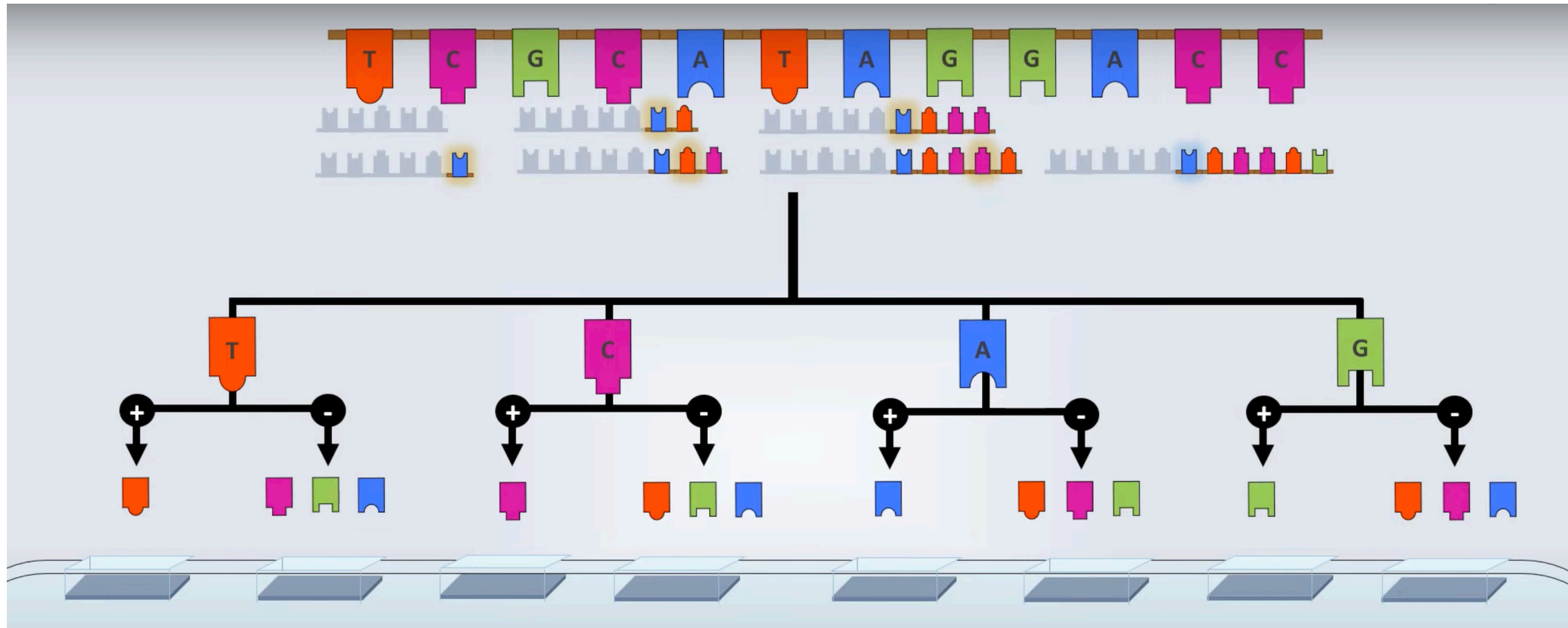


PLATE I. Radioautograph of experiment 1 (see Materials and Methods) in which the decanucleotide was used as a primer. The diagram shows the interpretation and the sequence deduced. The dashed lines represent artifact bands (see text). The bromphenol blue marker was 1 cm below the section of the gel shown in the Plate and the xylene cyanol FF marker opposite the product 50 nucleotides long.

Sanger's Plus/Minus sequencing



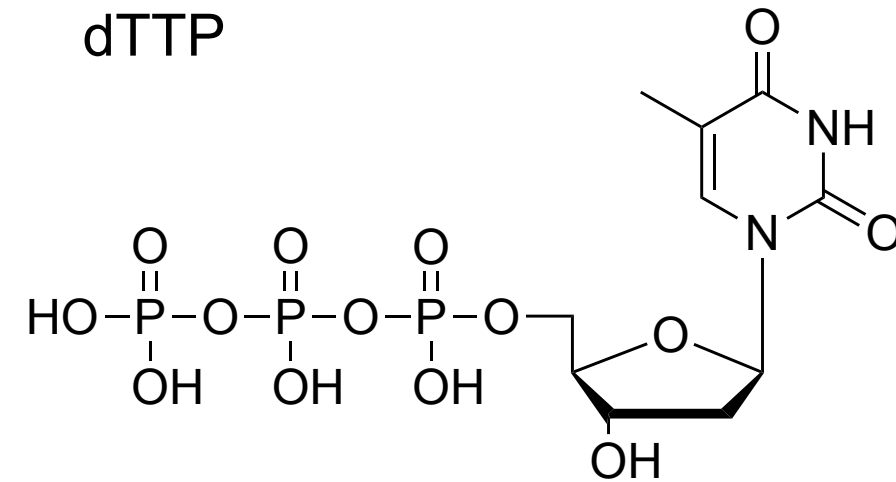
Dideoxynucleotides (ddNTP) are incorporated

Polymerases can incorporate ddNTPs.

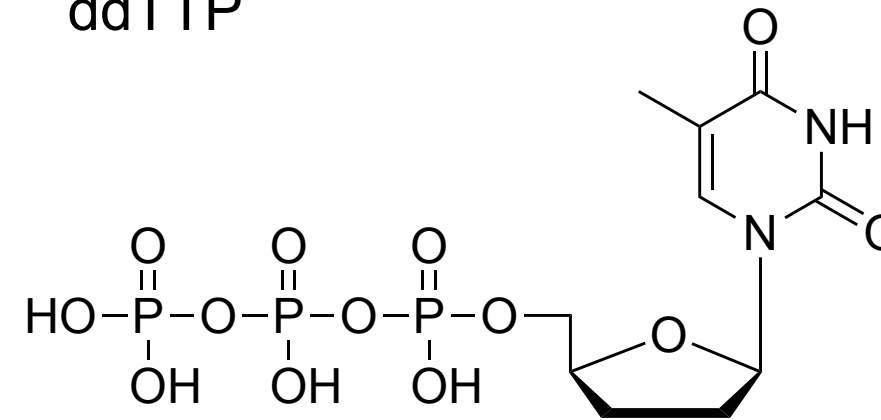
Leading to a *single base extension* followed by termination

No 3'hydroxyl group (3'OH).

dTTP



ddTTP





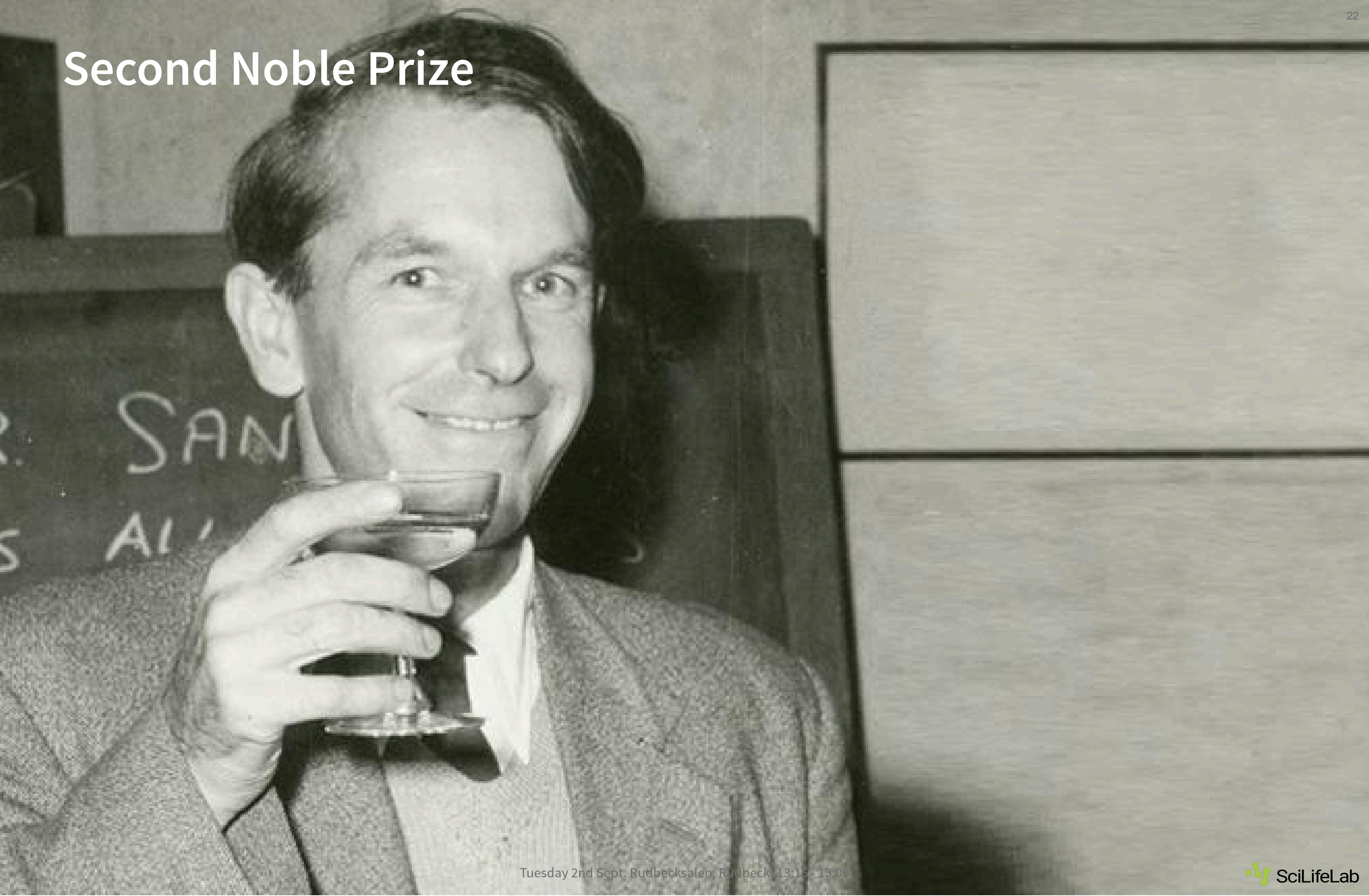
Chain-termination method 🏆



“Sanger sequencing”



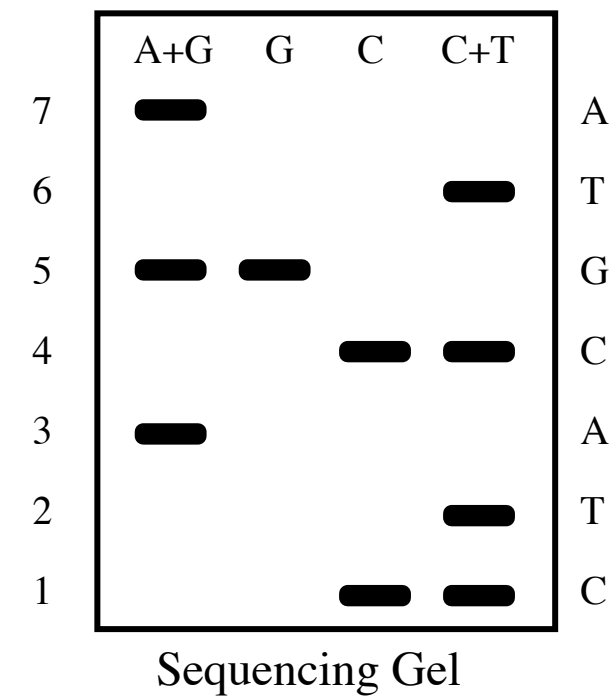
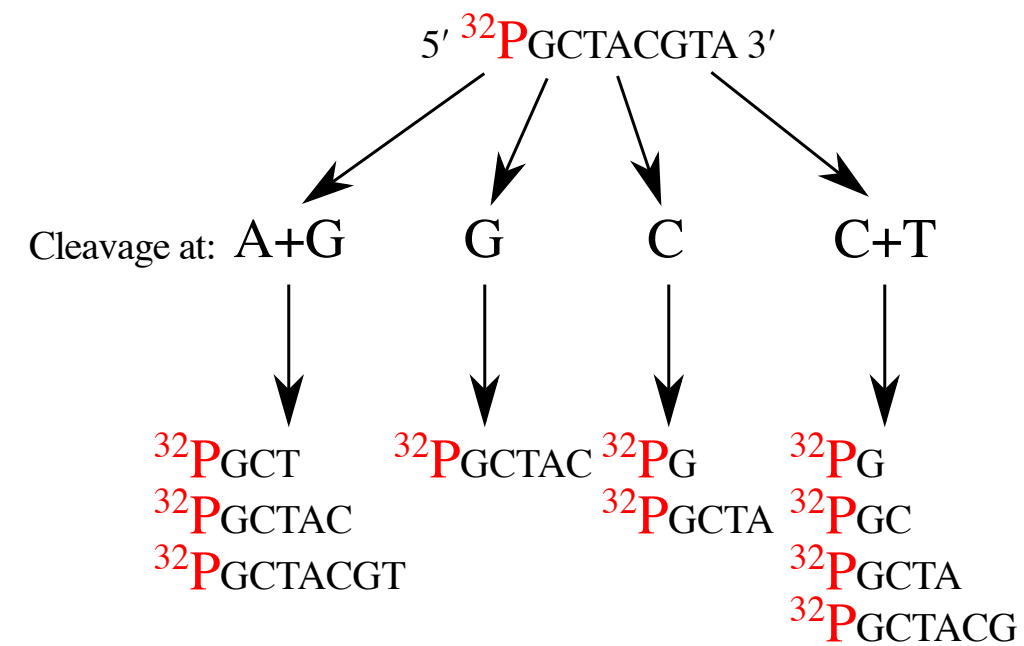
Second Noble Prize



Tuesday 2nd Sept. Rudbecksalen, Rudbeck 13:15 - 15:00

Maxam & Gilbert sequencing

- Radioactive 5' labeling
- Chemical cleavage
- PAGE



The birth of biotech

California. Genentech and recombinant technologies were already creating value. Flow cytometer and Beckman Coulter commercial machines.

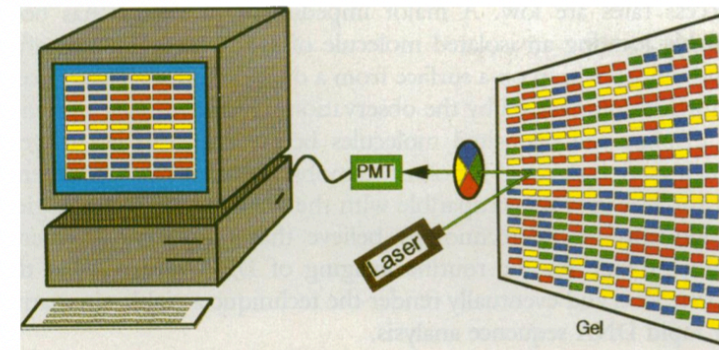
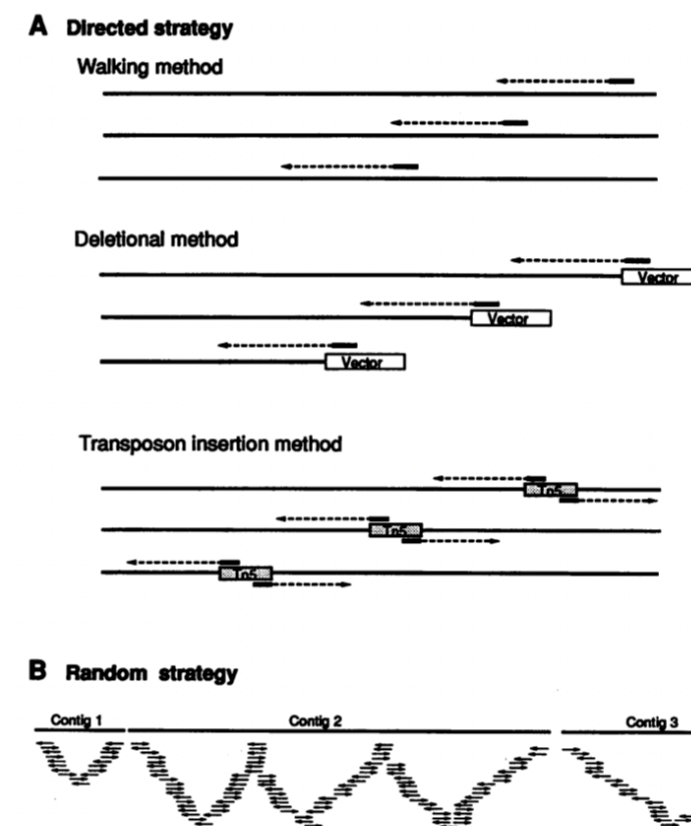


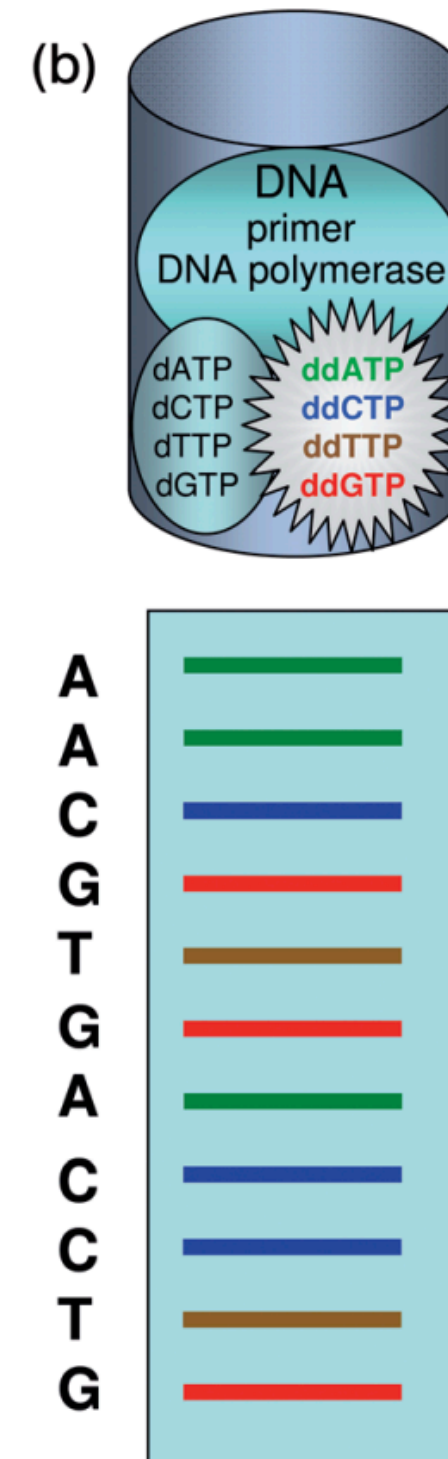
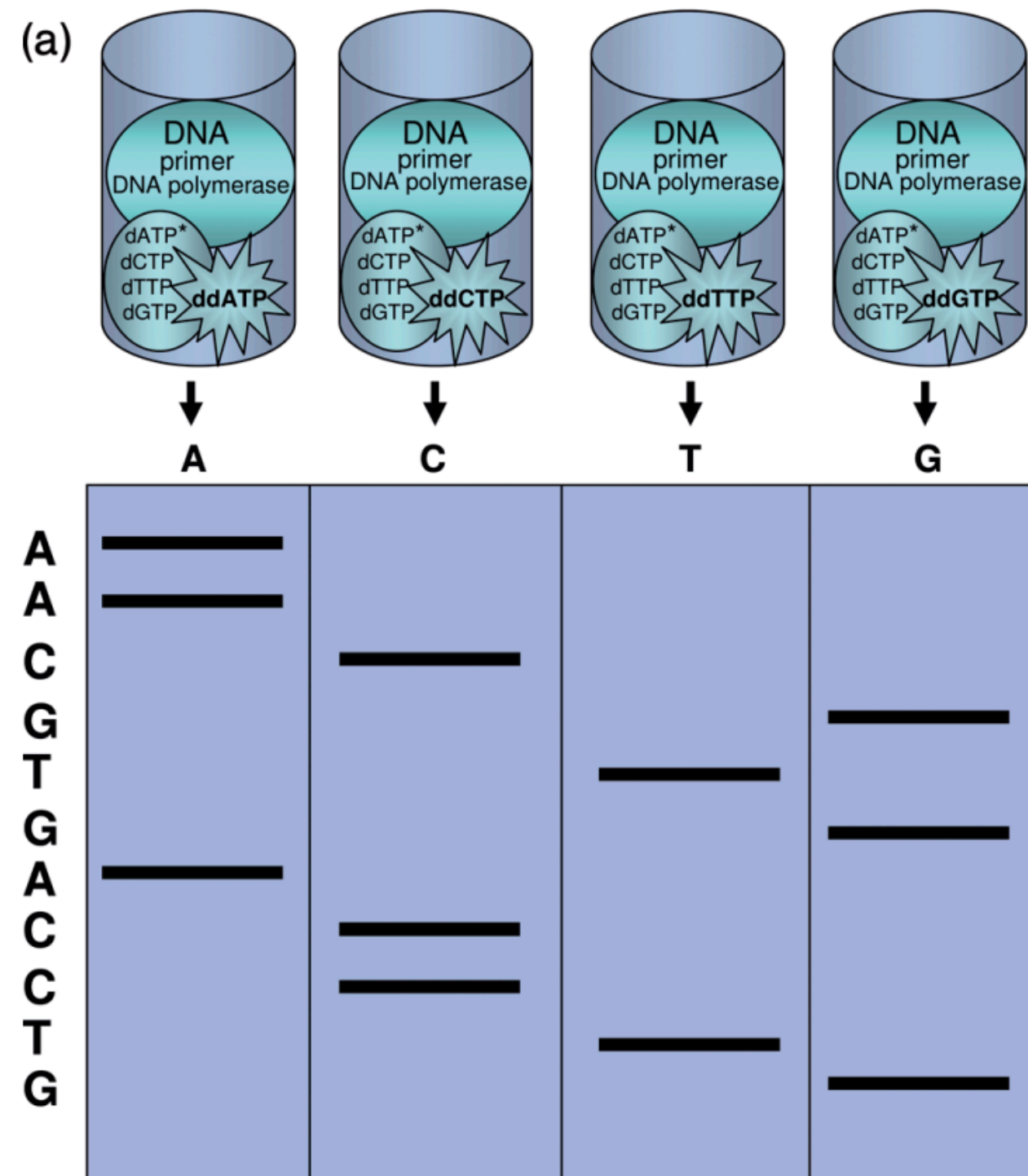
Fig. 6. A schematic illustration of the principle of operation of the ABI 373A automated sequencer (see text); PMT, photomultiplier tube. [Adapted from (42) with permission © Eaton]



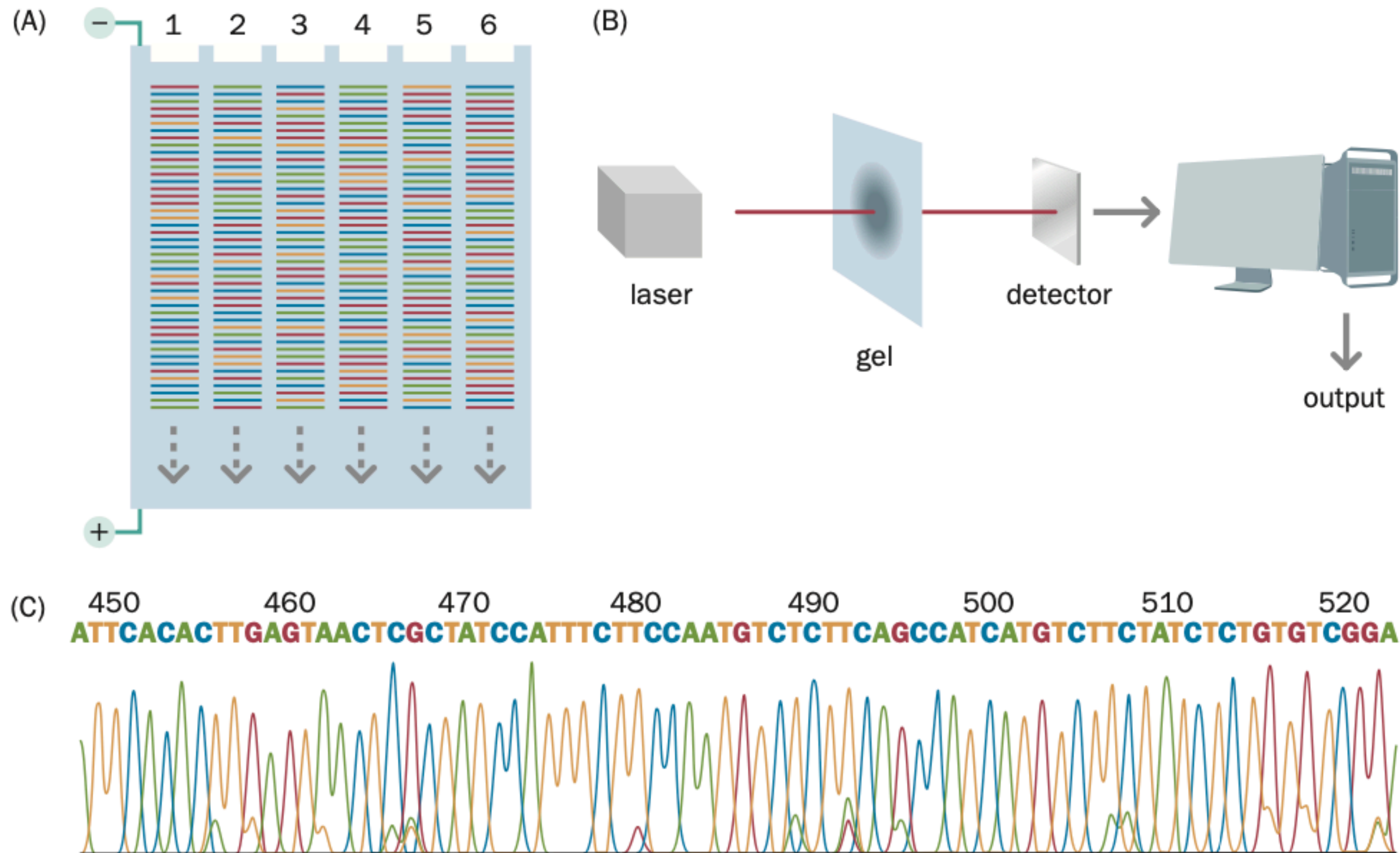
Cal. Tech. Leroy Hood and Mike Hunkapiller Applied Bioscience Inc. (ABI)

Sanger sequencing

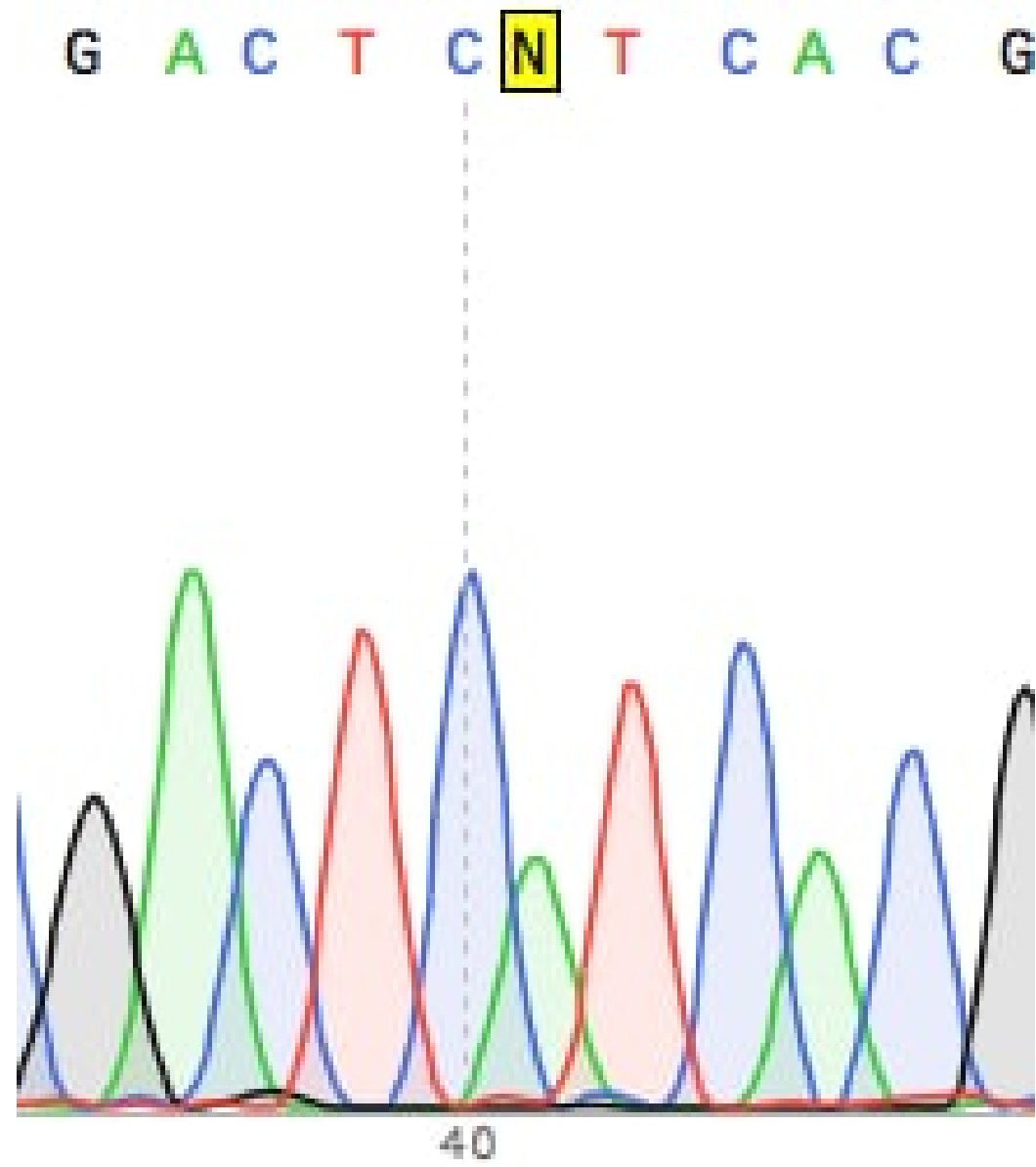
Fluorescently labeled nucleotides.



Capillary gel electrophoresis



Sanger sequencing



Base calling: assigning specific nucleotide bases to raw signals generated during sequencing.

Human Genome Project

$$\begin{aligned}
 & \sum_{j \neq i} \lambda_{ij} X_j - \frac{\lambda_{ii} X_i}{n} - \sum_{j \neq i} \lambda_{ji} X_j - \lambda_{ii} X_i + \sum_{j \neq i} 2(\lambda_{ij} + \lambda_{ji}) X_j + \lambda_{ii} 2X_i - \sum_{j \neq i} \lambda_{ji} X_j = \\
 & \left(\frac{\lambda_{ii} X_i}{n} + \sum_{j \neq i} \lambda_{ji} 2 + \sum_{j \neq i} \lambda_{ij} 2 + \lambda_{ii} 2 \right) X_i \\
 & \sum_{j \neq i} \lambda_{ij} X_j + \sum_{j \neq i} \lambda_{ji} X_j + \sum_{j \neq i} \lambda_{ii} X_i + \sum_{j \neq i} \lambda_{ji} X_j - \frac{\lambda_{ii} X_i}{n} - \lambda_{ii} X_i + (\lambda_{ii} - \lambda_{ii} - \lambda_{ii} + \sum_{j \neq i} \lambda_{ji}) 2 + \sum_{j \neq i} \lambda_{ij} X_j = \\
 & \left(\frac{1}{n} - 1 \right) \lambda_{ii} X_i + (\sum_{j \neq i} \lambda_{ji} + \sum_{j \neq i} \lambda_{ij} - \lambda_{ii}) 2 + \sum_{j \neq i} \lambda_{ij} X_j = \\
 & \lambda_{ii} X_i + \sum_{j \neq i} \lambda_{ij} X_j + \sum_{j \neq i} \lambda_{ji} X_j + \sum_{j \neq i} \lambda_{ii} X_i - \frac{\lambda_{ii} X_i}{n} - \lambda_{ii} X_i = \left(\frac{1}{n} - 1 \right) \lambda_{ii} X_i + (\sum_{j \neq i} \lambda_{ji} - \lambda_{ii} + \lambda_{ii}) 2 + \sum_{j \neq i} \lambda_{ij} X_j =
 \end{aligned}$$

Next-generation sequencing

All about how to massively multiplex sequencing.

- Pyrosequencing
- Sequencing by ligation
- Reversible dye-terminators

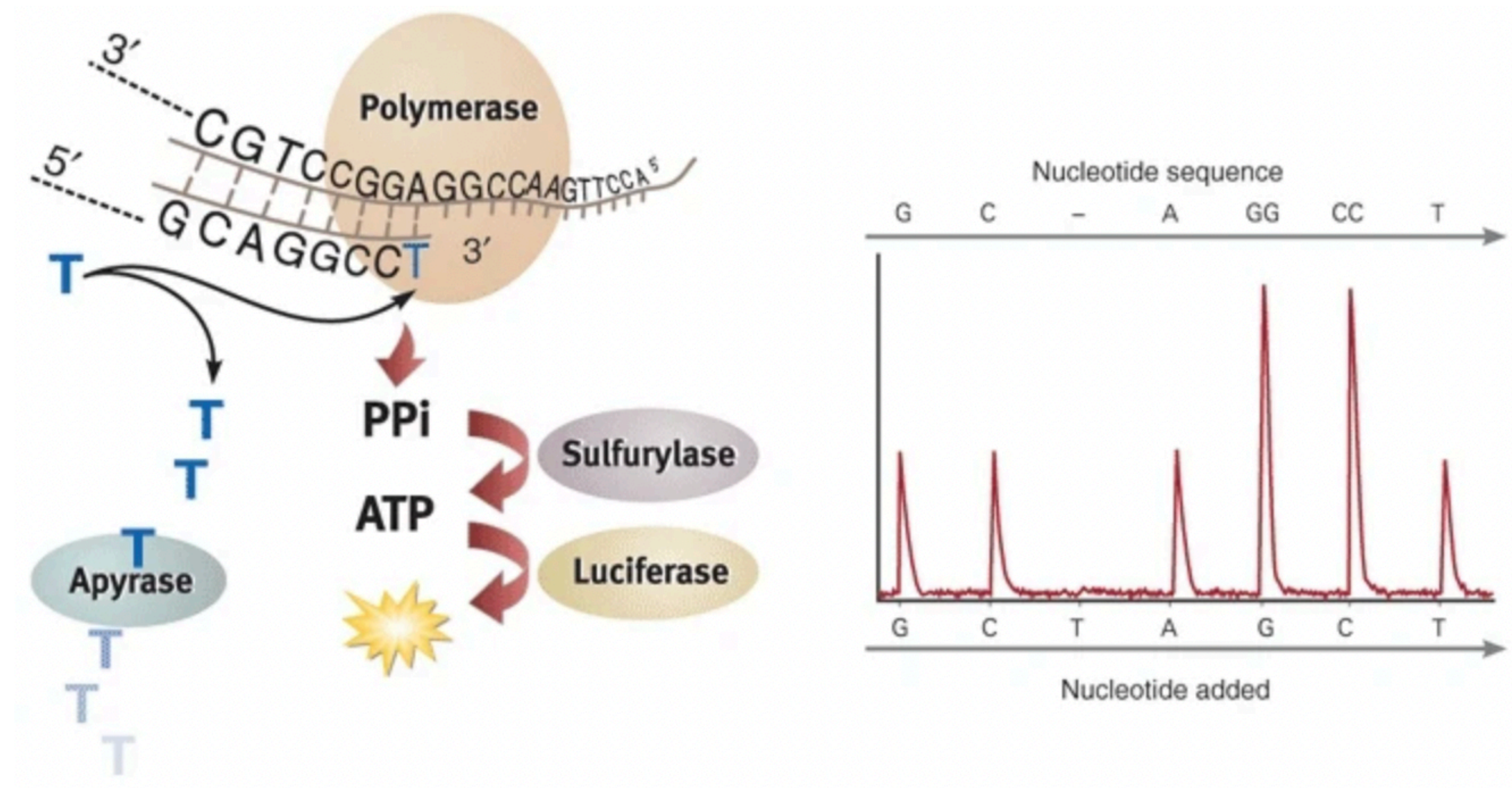
Solid-phase substrate

- Bridge amplification
- Emulsion PCR
- DNA nanoballs

Primer extension

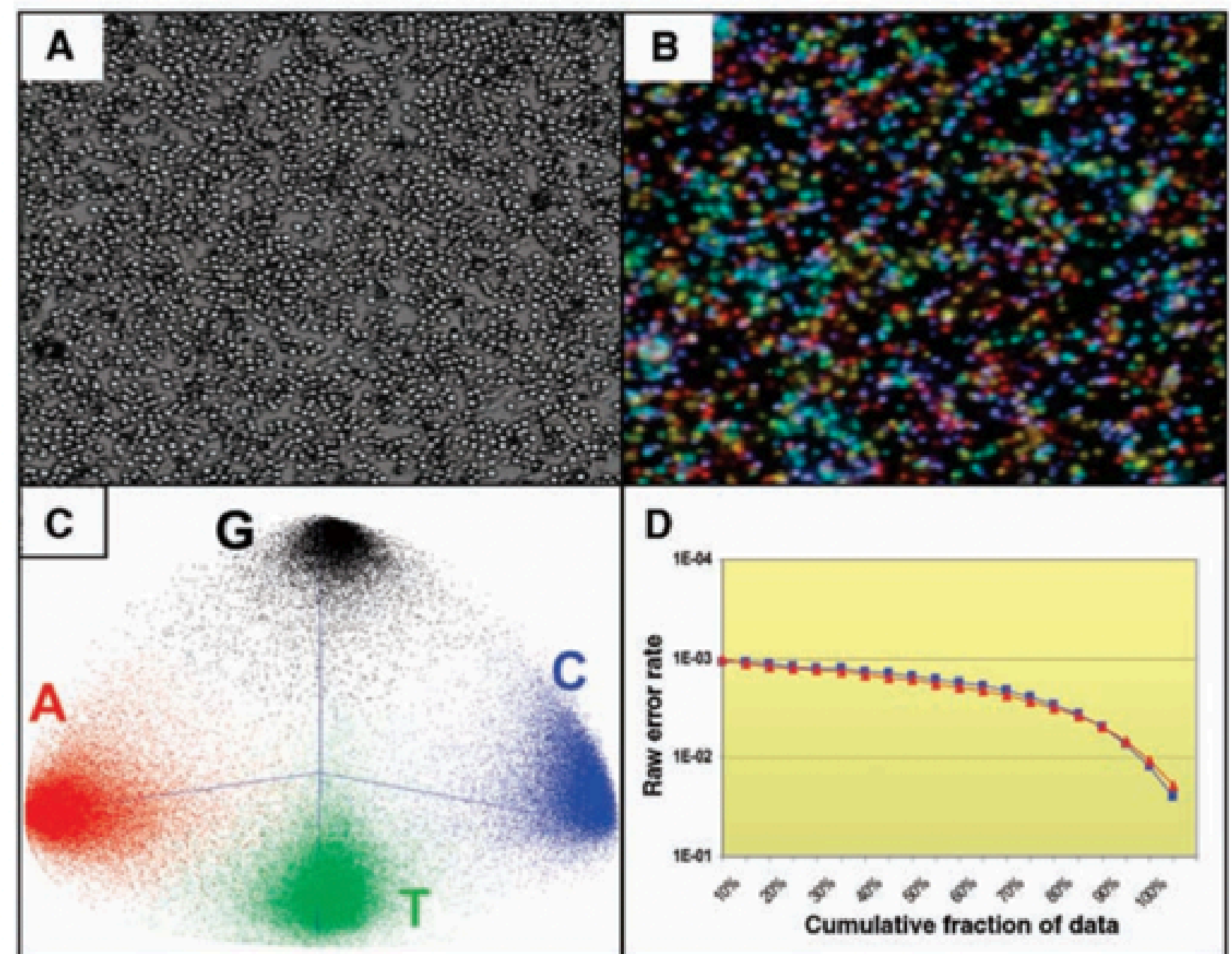
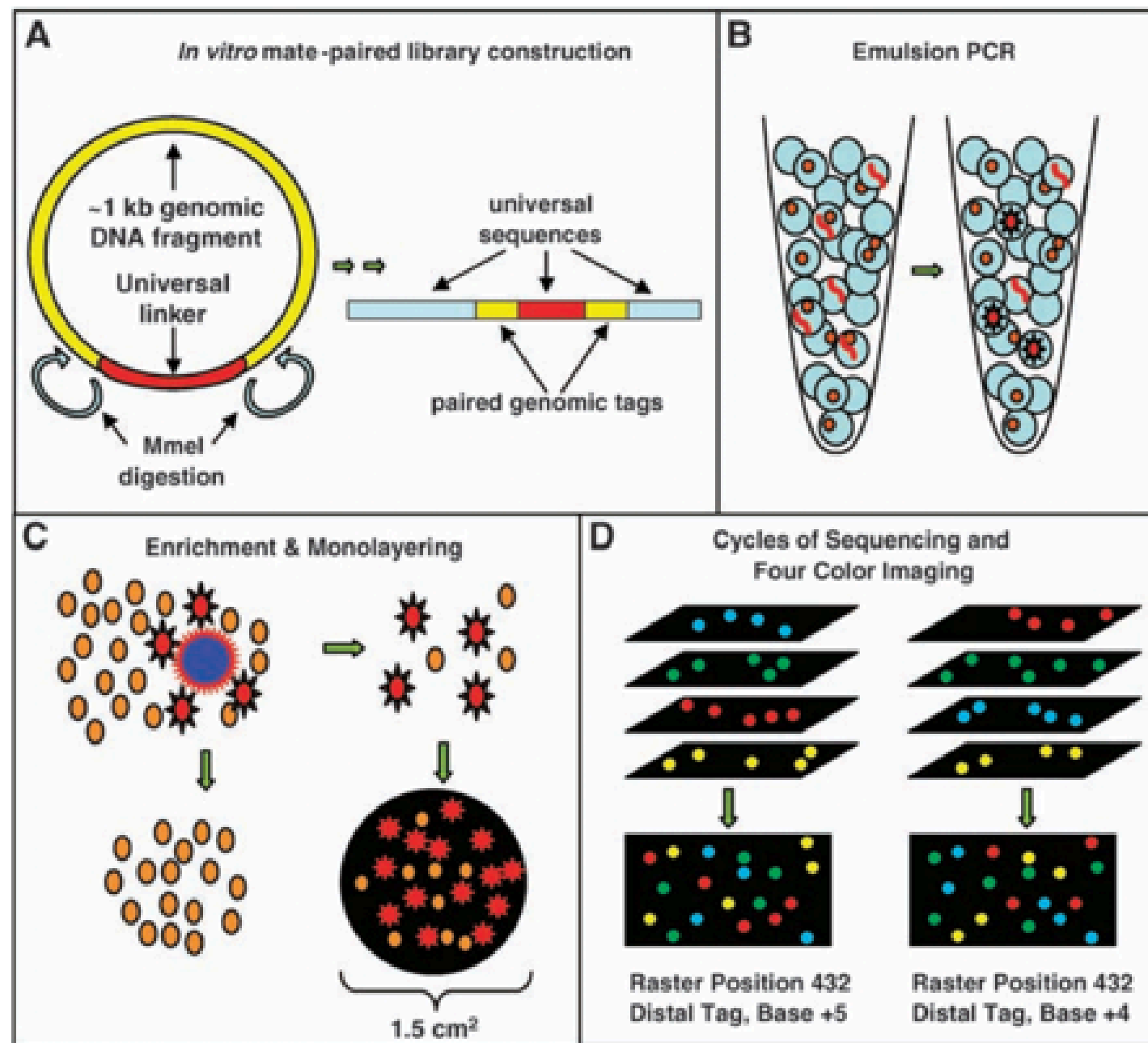
- Sequencing by synthesis (SBS)
- Sequencing by ligation (SBL)

Pyrosequencing



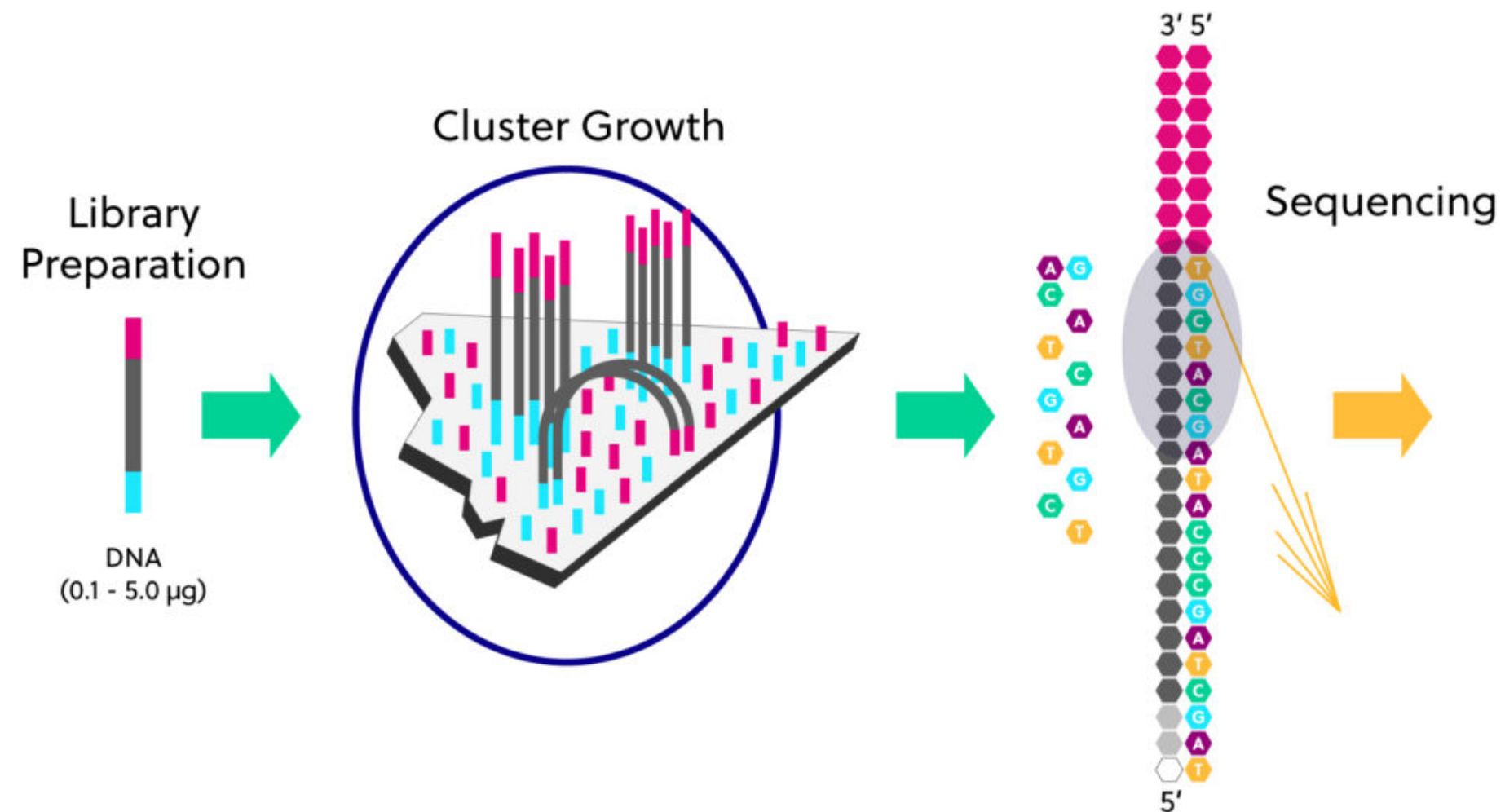
- **Principle:** DNA polymerase incorporates nucleotides, releasing pyrophosphate (PPI).
- **Signal:** PPI is converted to light via sulfurylase and luciferase, where the light intensity indicates the number of incorporated bases.
- **Analysis:** Light signals are used to determine the DNA sequence in real time.
- **Question:** How can individual bases (A, T, C, G) be discriminated?

Polony sequencing

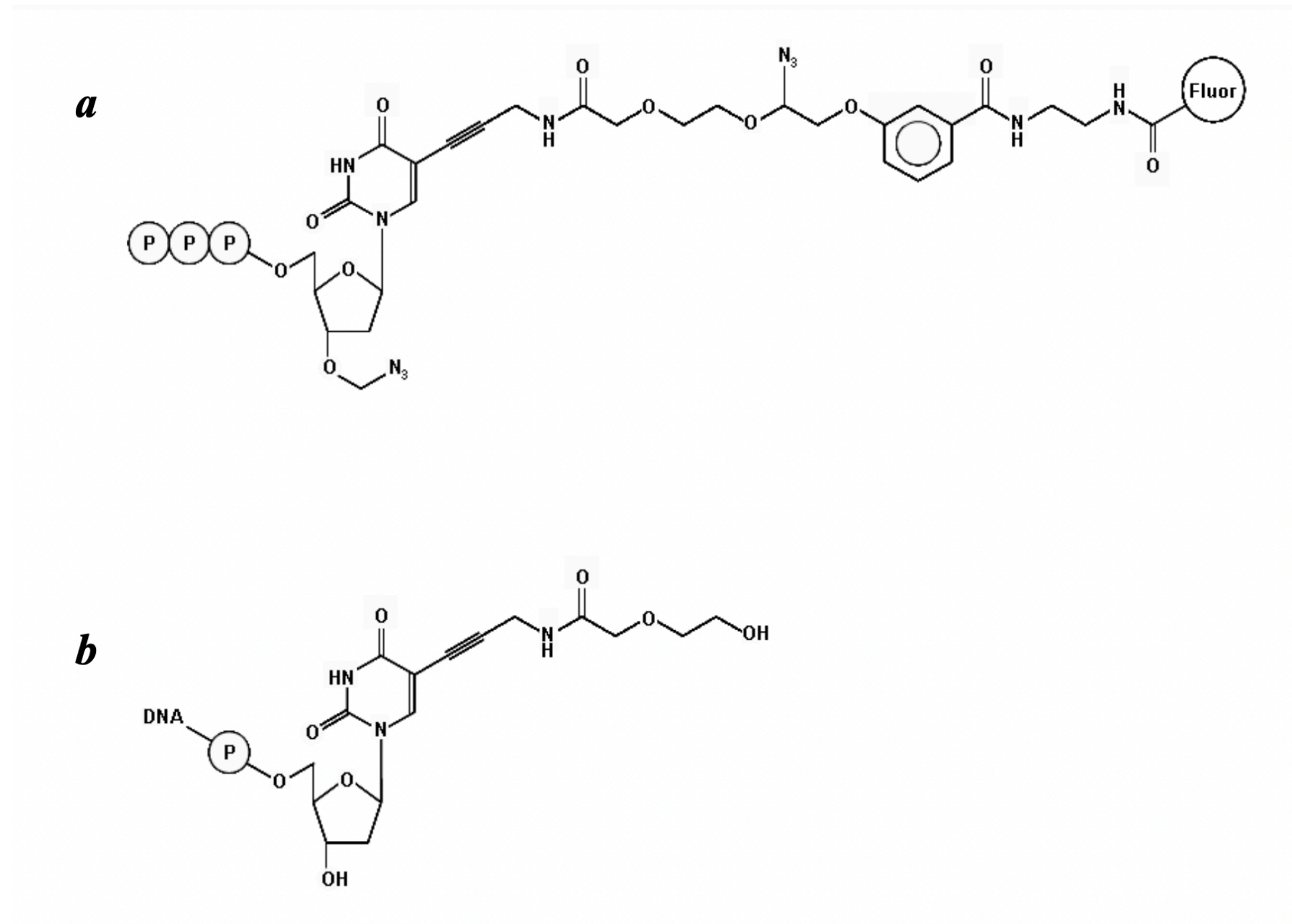


Solexa/Illumina sequencing

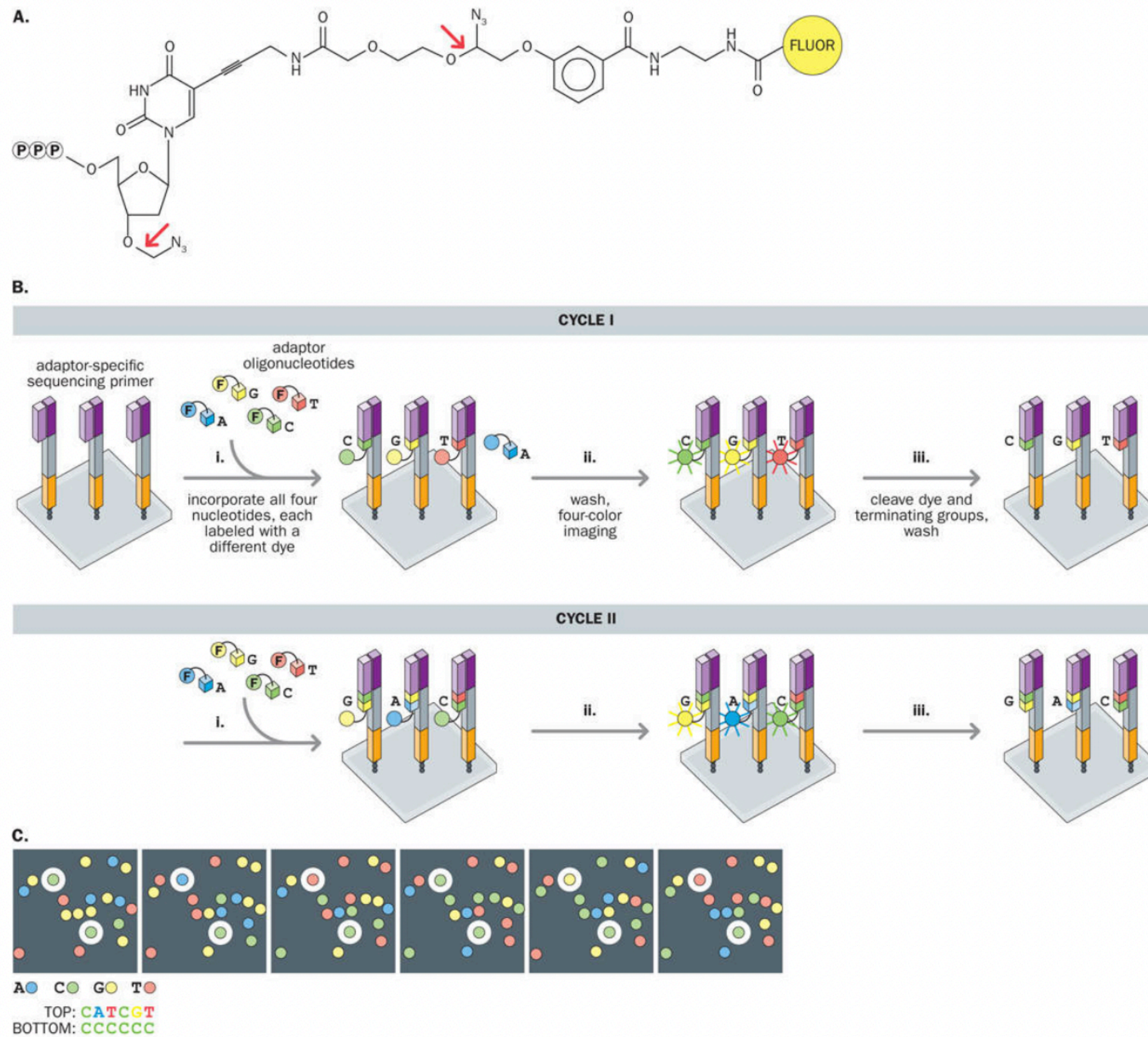
- **Solid support:** PCR bridge amplification
- **Primer extension:** Reversible dye-terminators (patent 444 expired last year!)



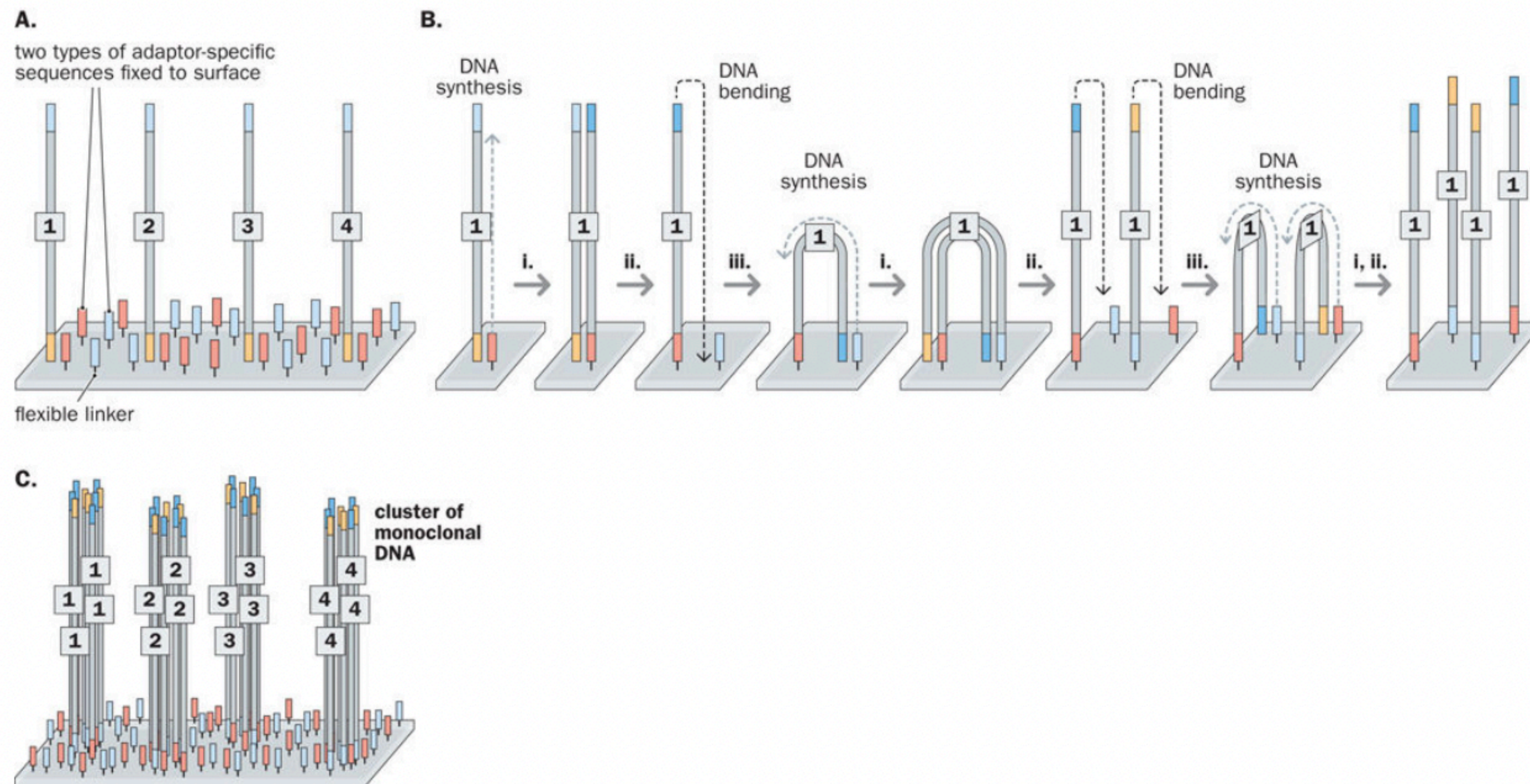
Solexa/Illumina sequencing



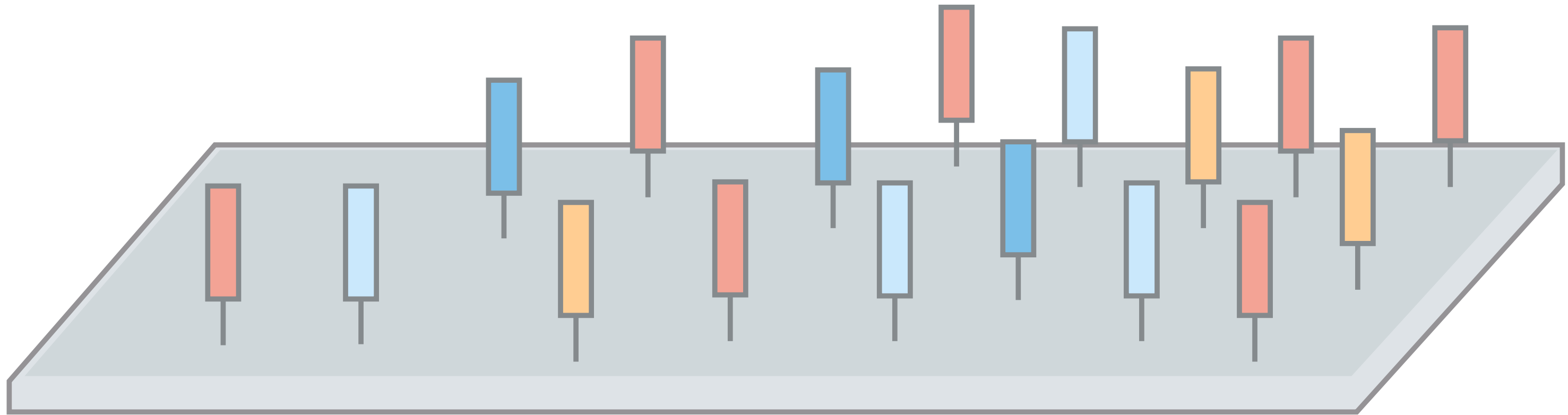
Reversible dye-terminators



Bridge amplification



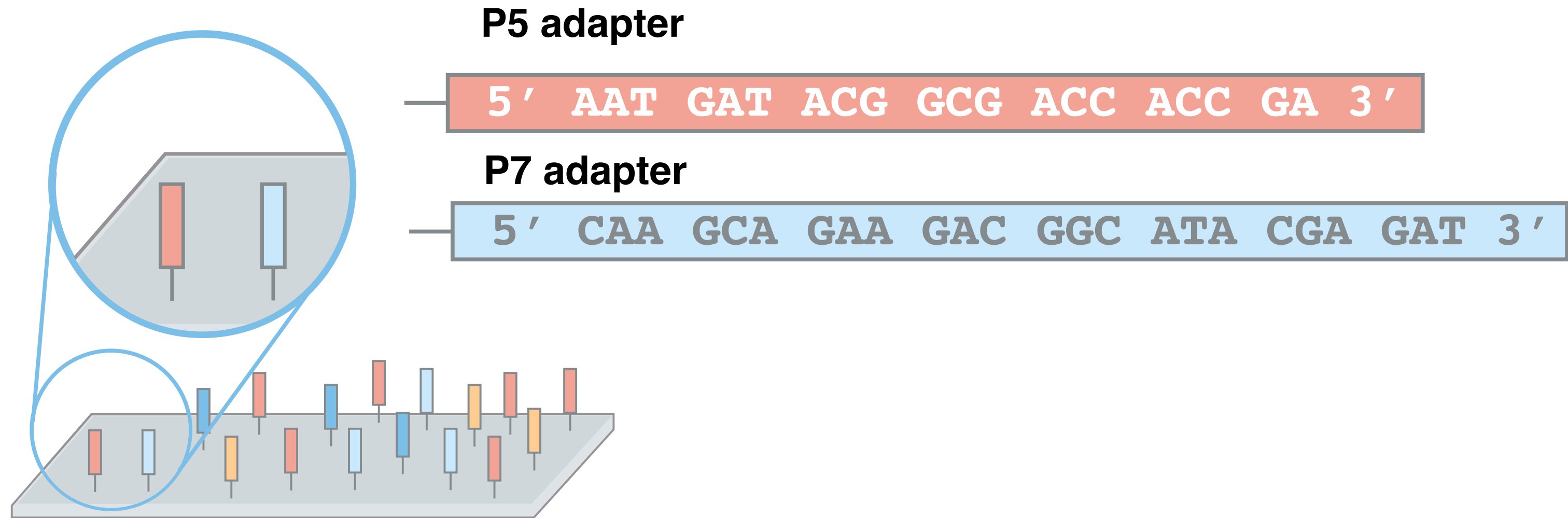
Bridge amplification - Flow cell



Flow cell

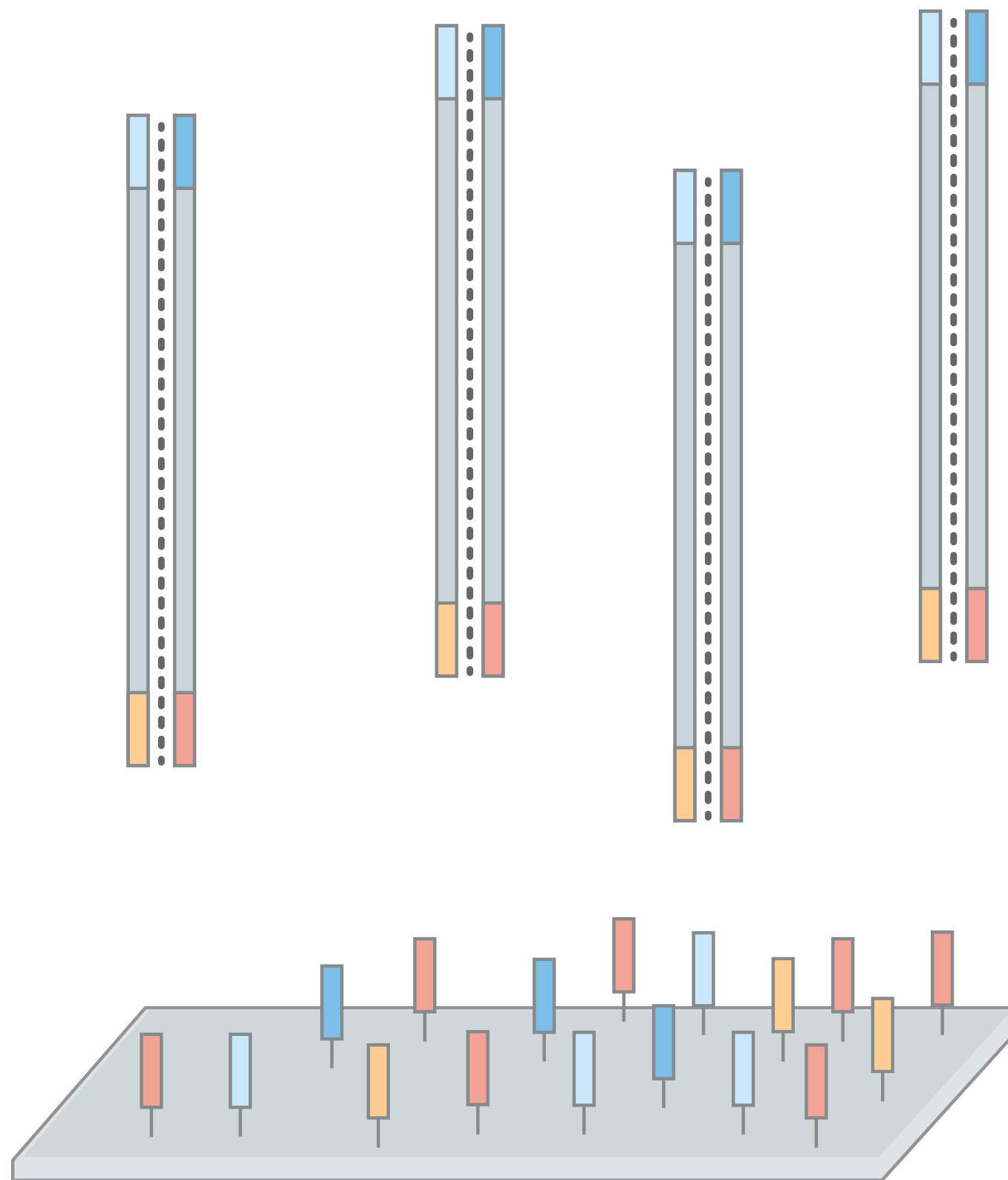
- **Flow cell:** A microscope slide where oligonucleotides have been *functionalized* with a linker so that they are covalently attached to the glass.

Bridge amplification - Illumina adapters



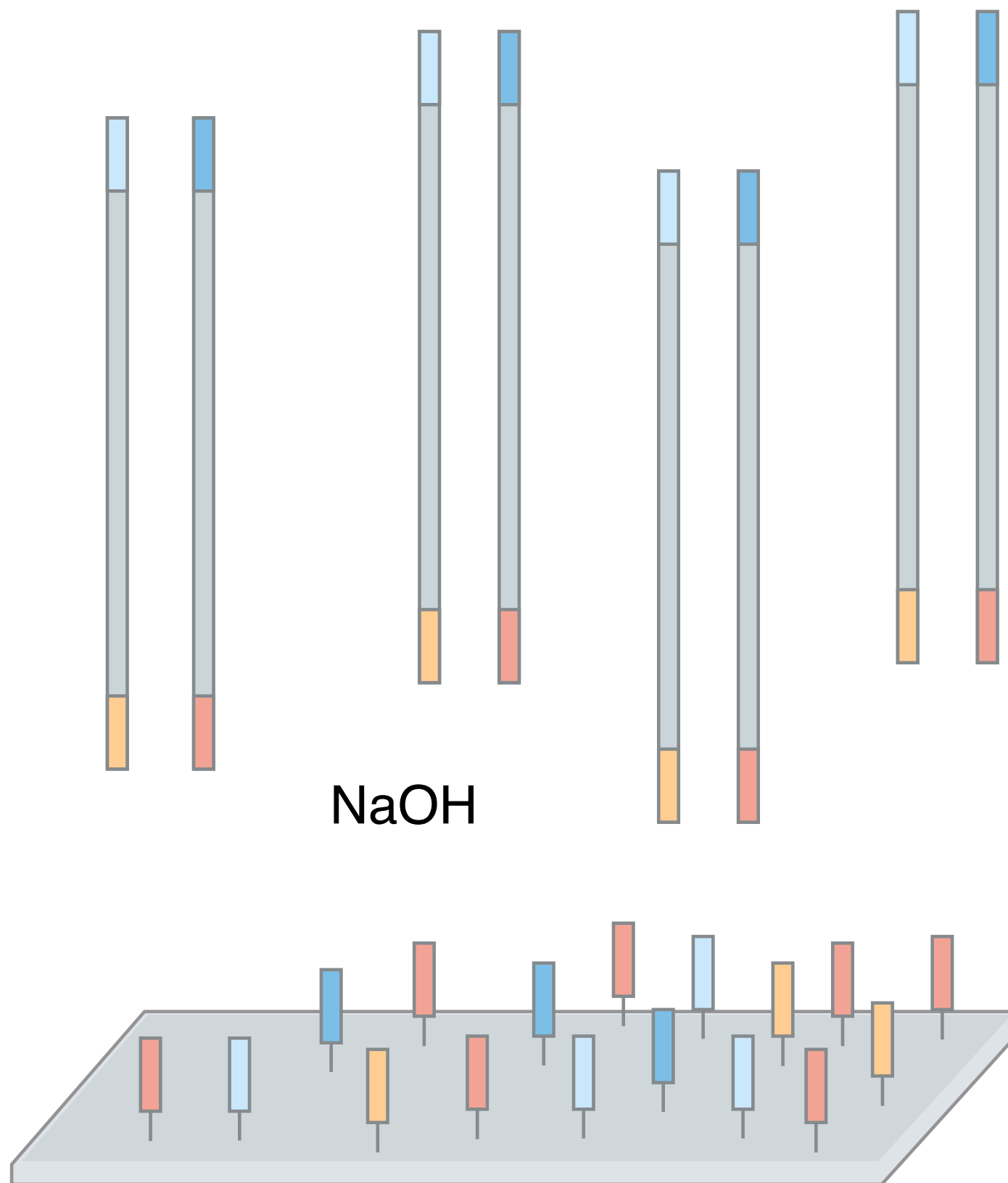
- **P5 adapter:** common that the first read (R1) starts from this end.
- **P7 adapter:** common that the second read (R2) as well as the *multiplex index* (what sample) reads from this end (i7 = index P7).

Bridge amplification - sekvenseringsbibliotek



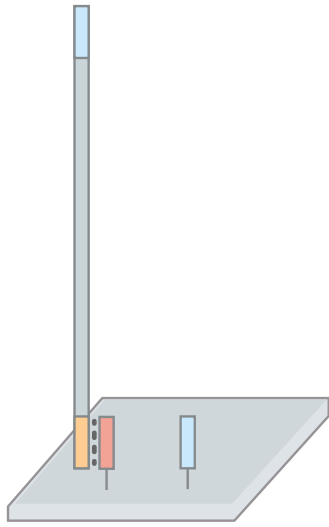
- PCR used to *flank* our DNA library with P5 and P7 adapter sequence.
- The library cannot be directly placed on the flow cell because it is double-stranded DNA (dsDNA) and therefore cannot bind to the flow cell oligos.

Bridge amplification - denaturing and laddering



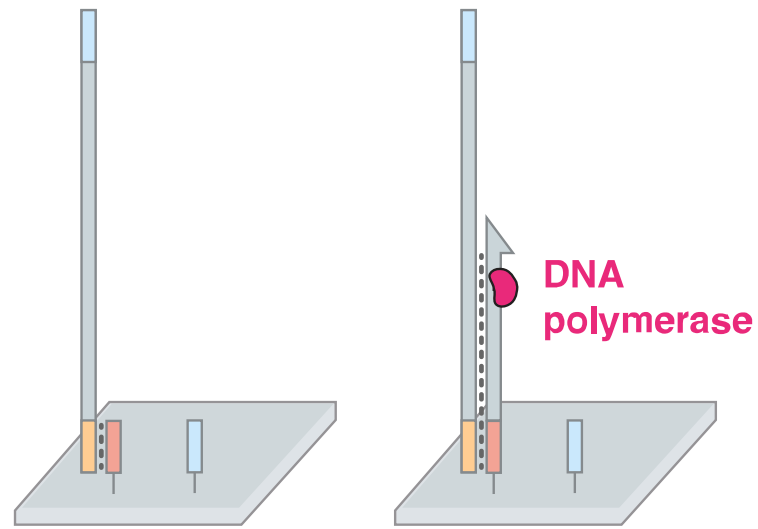
- The sequencing library is diluted to very low concentrations together with NaOH (sodium hydroxide).
- NaOH is strongly basic, which denatures dsDNA into ssDNA.

Bridge amplification - PCR



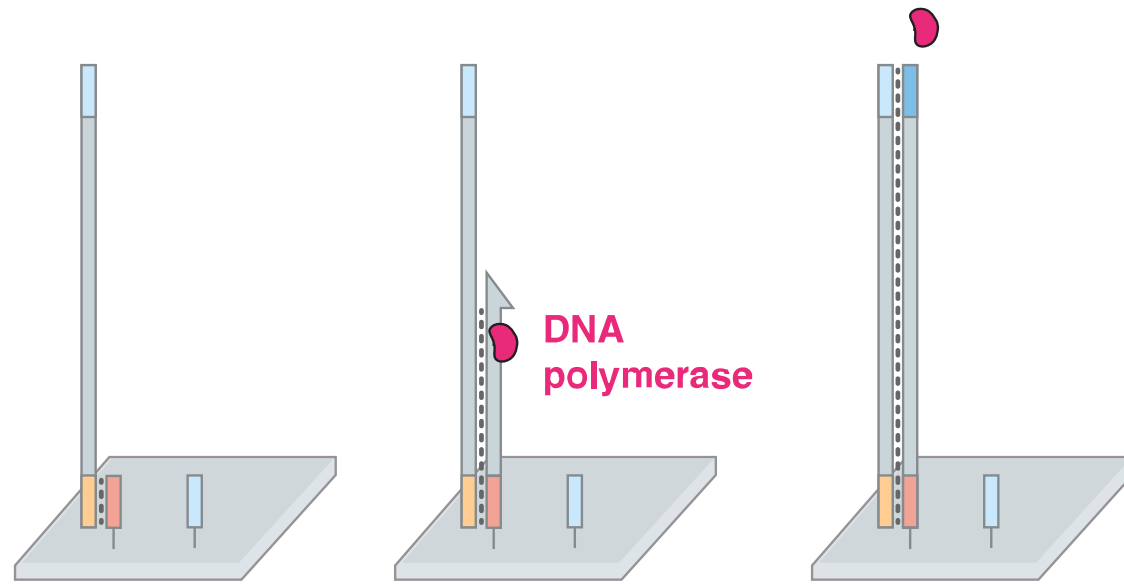
- ssDNA annealed to flowcell

Bridge amplification - PCR



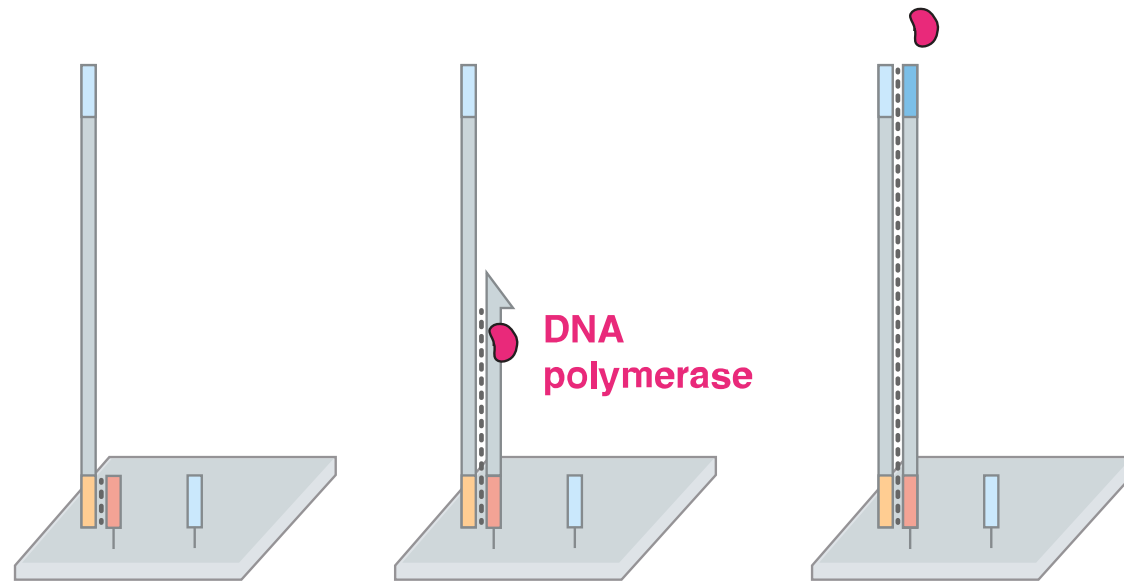
- Polymerase extension from P5/P7 adapter.

Bridge amplification - PCR



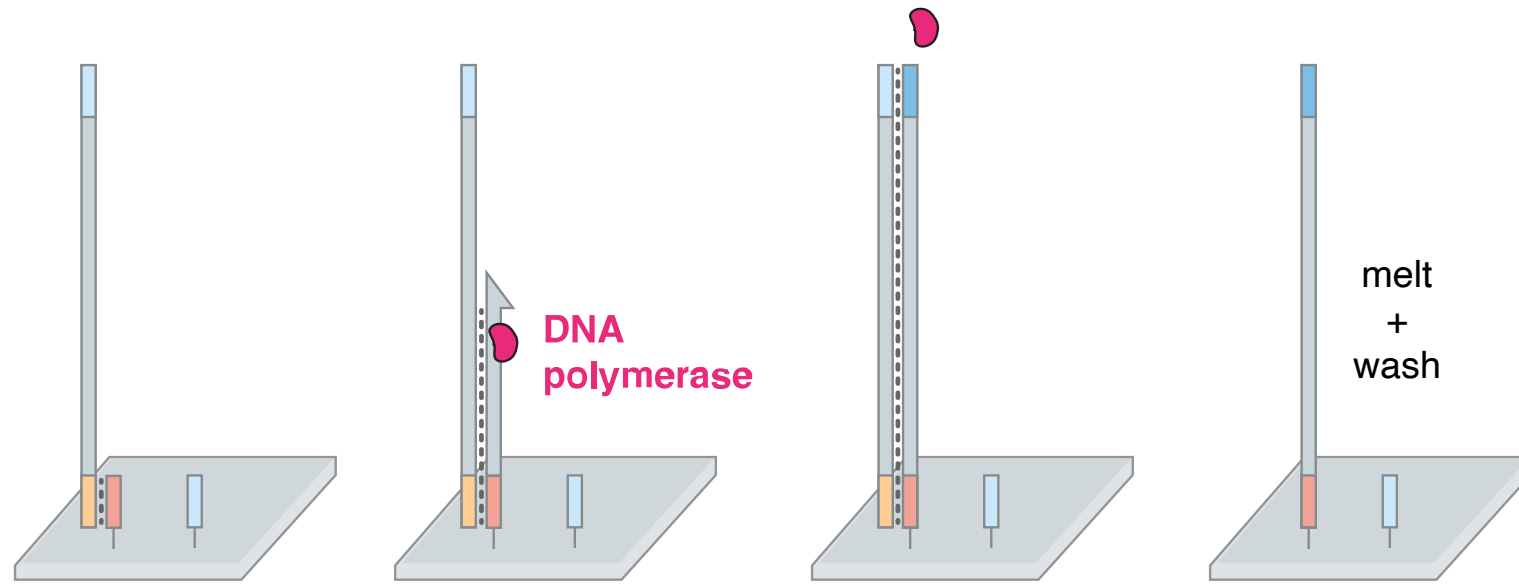
- dsDNA

Bridge amplification - PCR



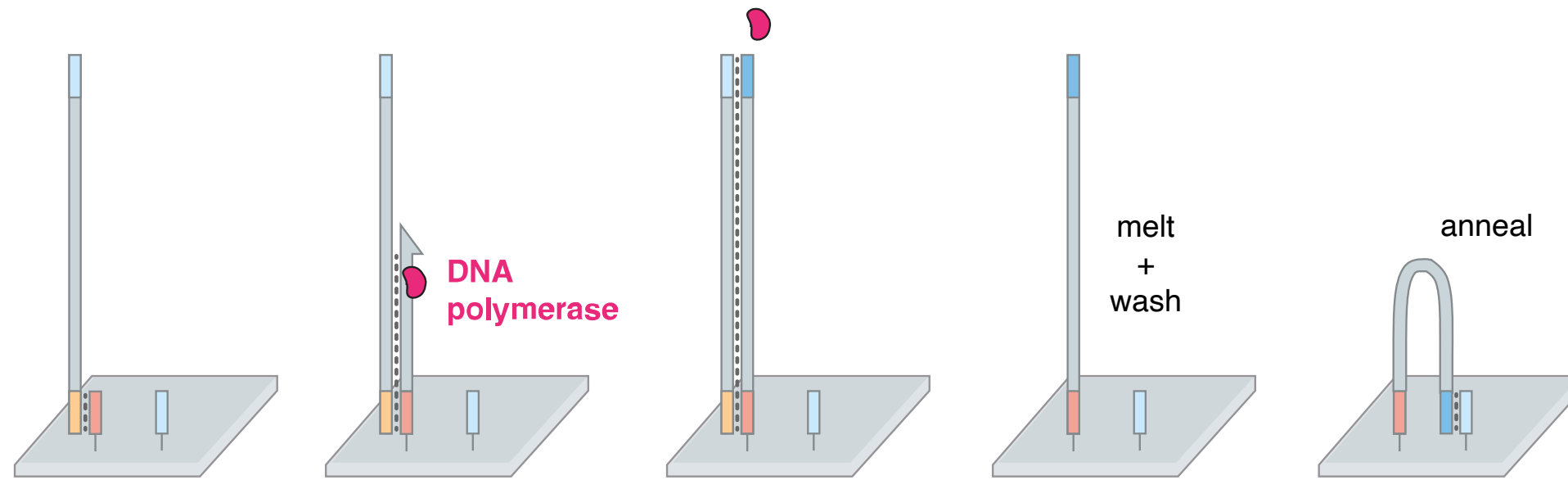
- dsDNA

Bridge amplification - PCR



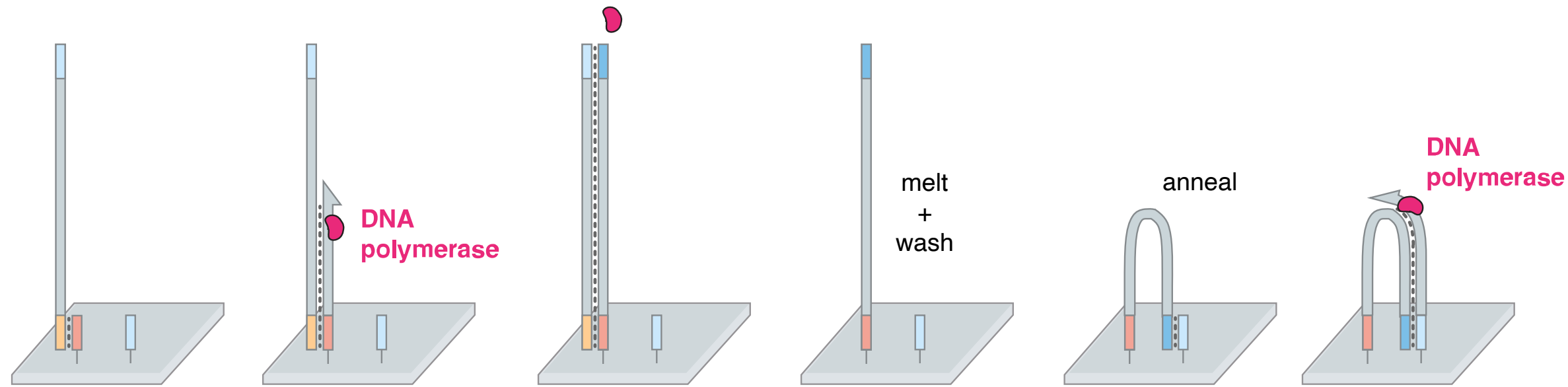
- Increase temperature och wash away non-covalent ssDNA.

Bridge amplification - PCR



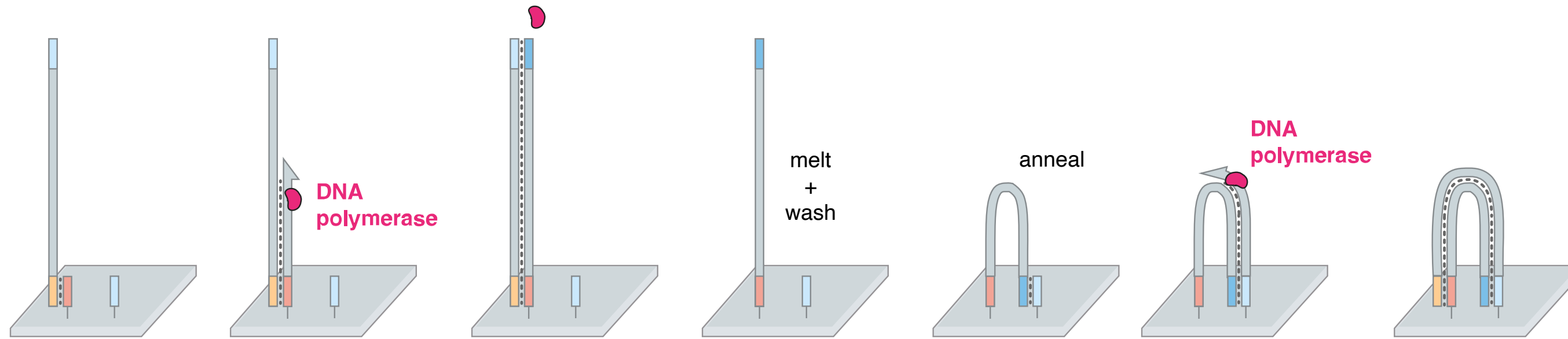
- Decrease temperature
- Hybridize to complementary adapter at the flow cell.

Bridge amplification - PCR



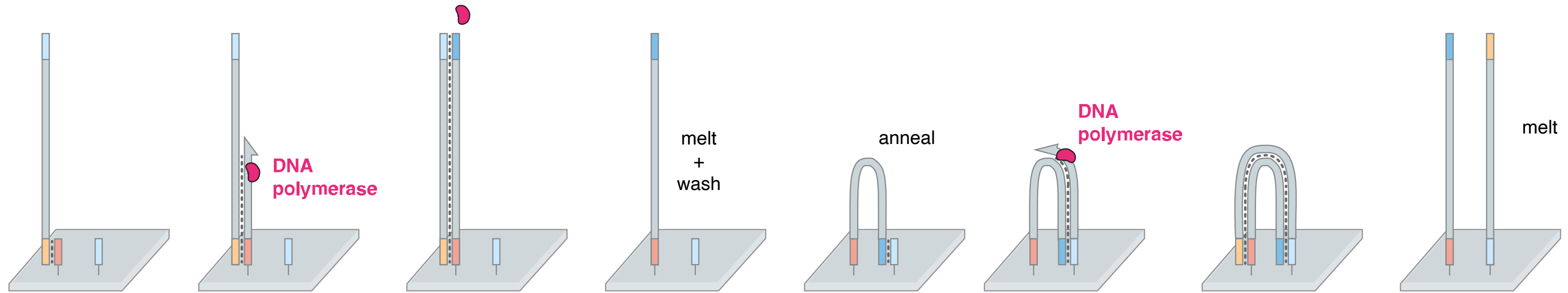
- Decrease temperature
- Hybridize to complementary adapter at the flow cell.
- Polymerase copies the covalent ssDNA.

Bridge amplification - PCR



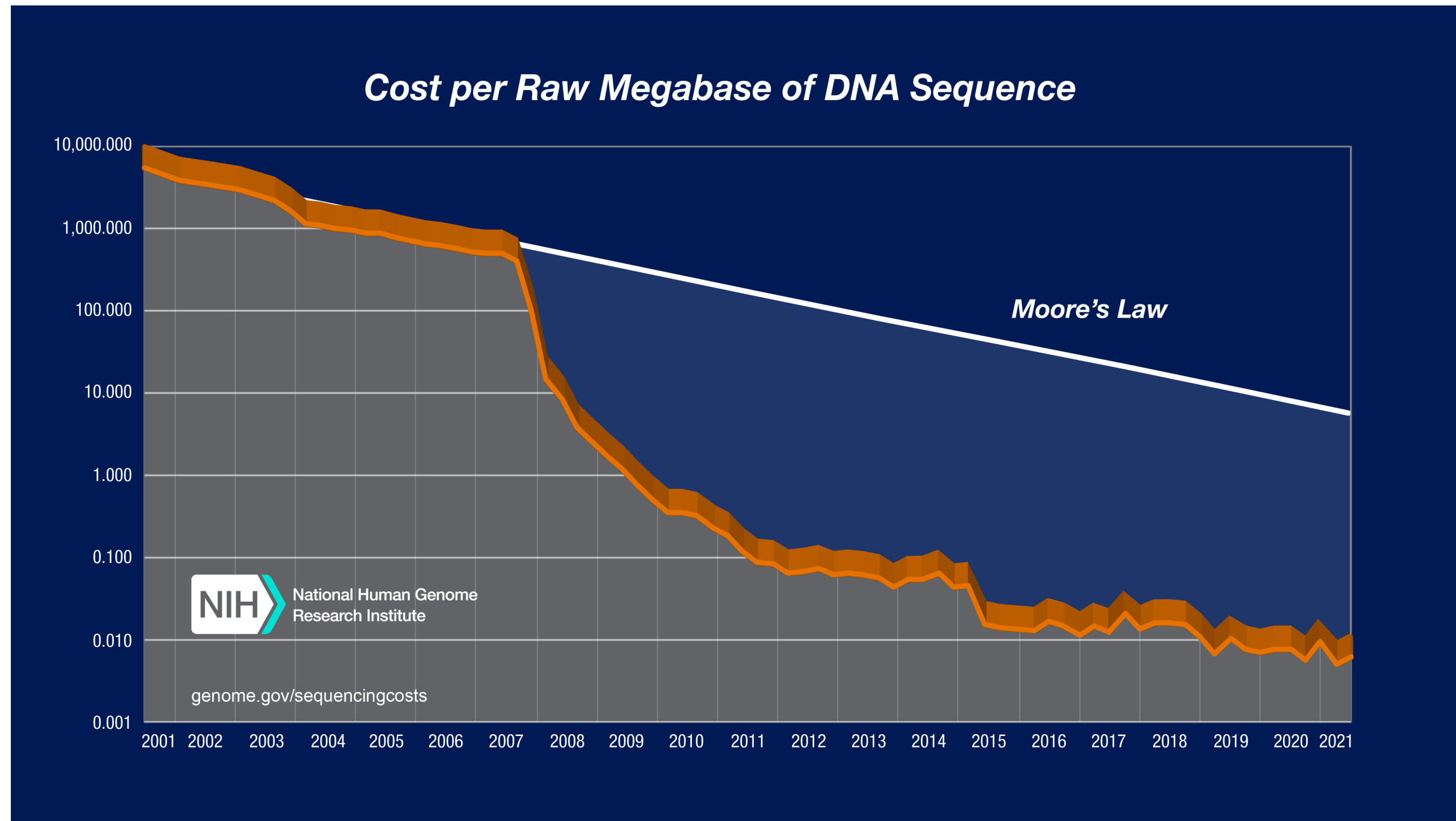
- Decrease temperature
- Hybridize to complementary adapter at the flow cell.
- Polymerase copies the covalent ssDNA.
- dsDNA

Bridge amplification - PCR

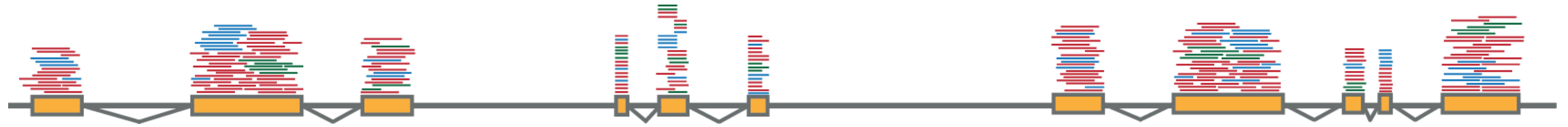


- Decrease temperature
- Hybridize to complementary adapter at the flow cell.
- Polymerase copies the covalent ssDNA.
- dsDNA
- Increase temperature and start the PCR cycle again.

Democratization of sequencing

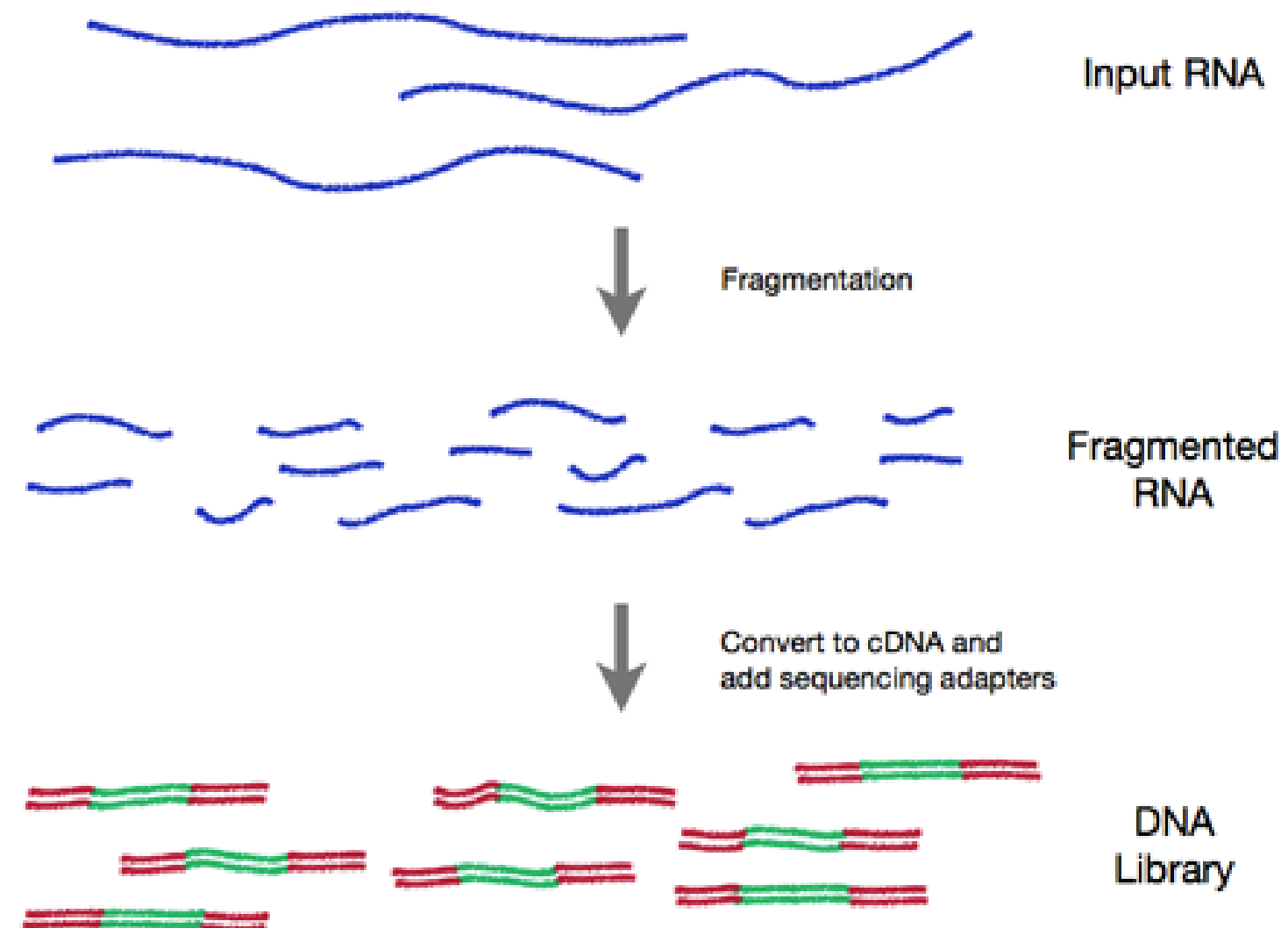


RNA sequencing

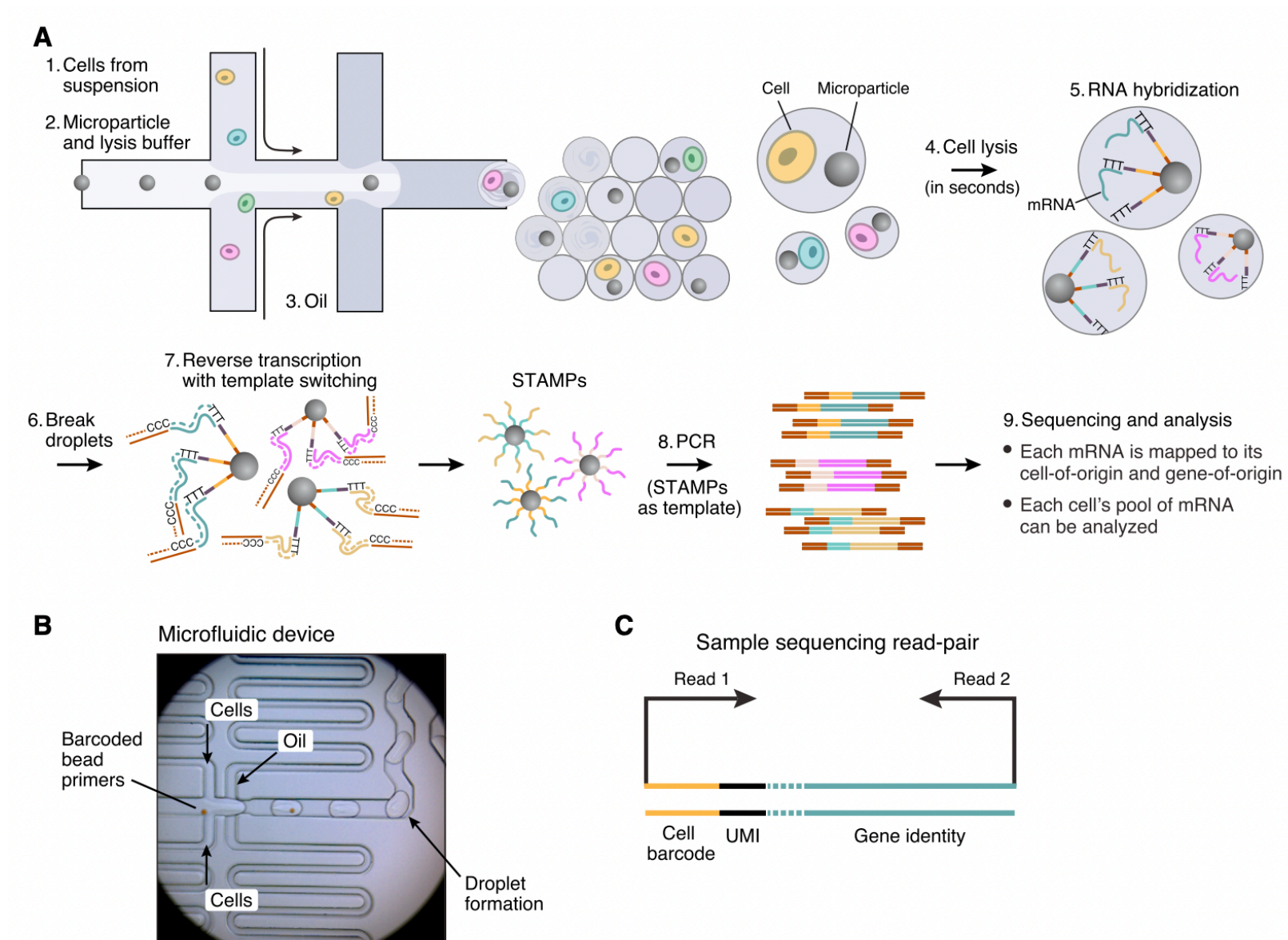


- Is actual DNA sequencing.
- But the RNA is converted to cDNA by reverse transcriptase.

RNA sequencing

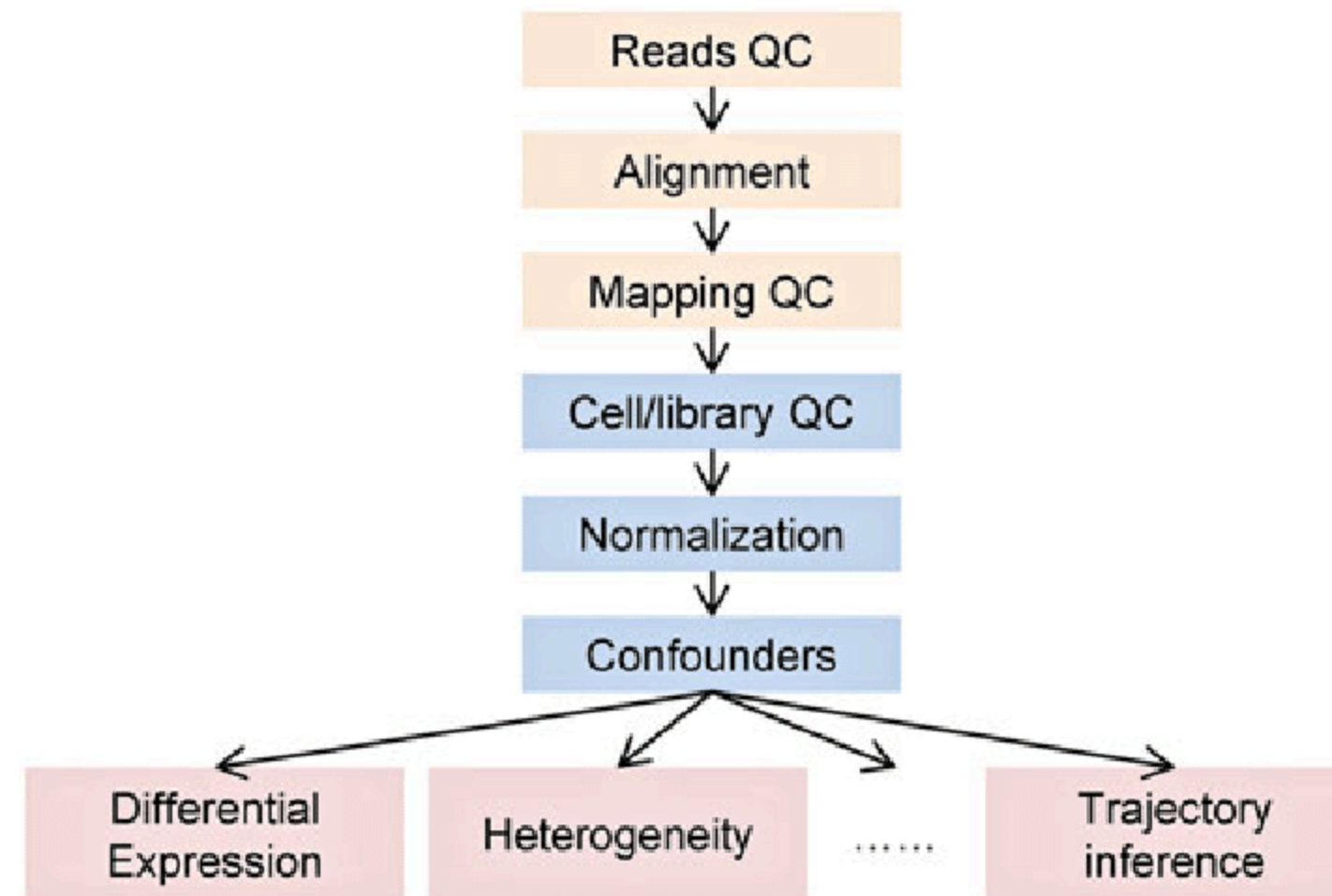


Single-cell RNA sequencing

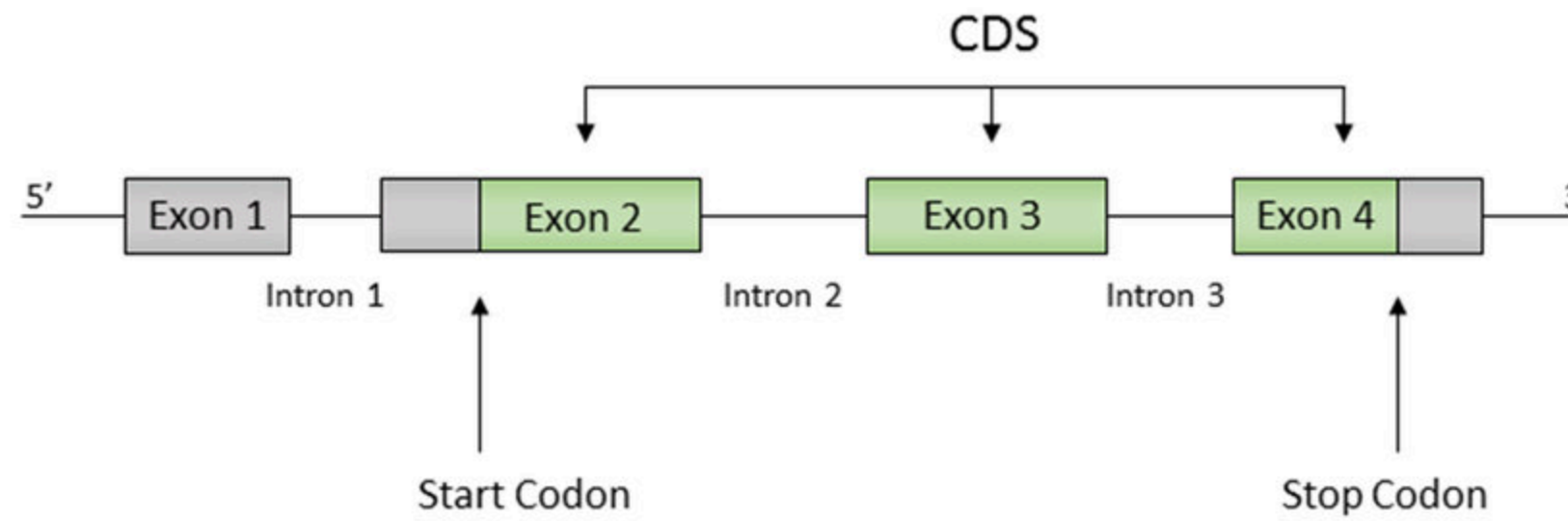


Transcriptomic analysis

Primary, secondary, tertiary



Gene body and why to sequence



- Alternative splicing
- Intron retention, pre-mRNA
- Alternative polyadenylation (3'UTR exon)

Strand-specificity

- Allele-specific expression



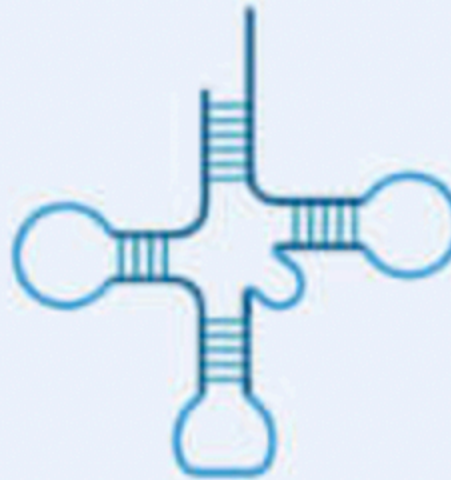
Different types of RNAs and their functions

mRNA



Encodes proteins

tRNA



Acts as adaptor between
mRNA and amino acids

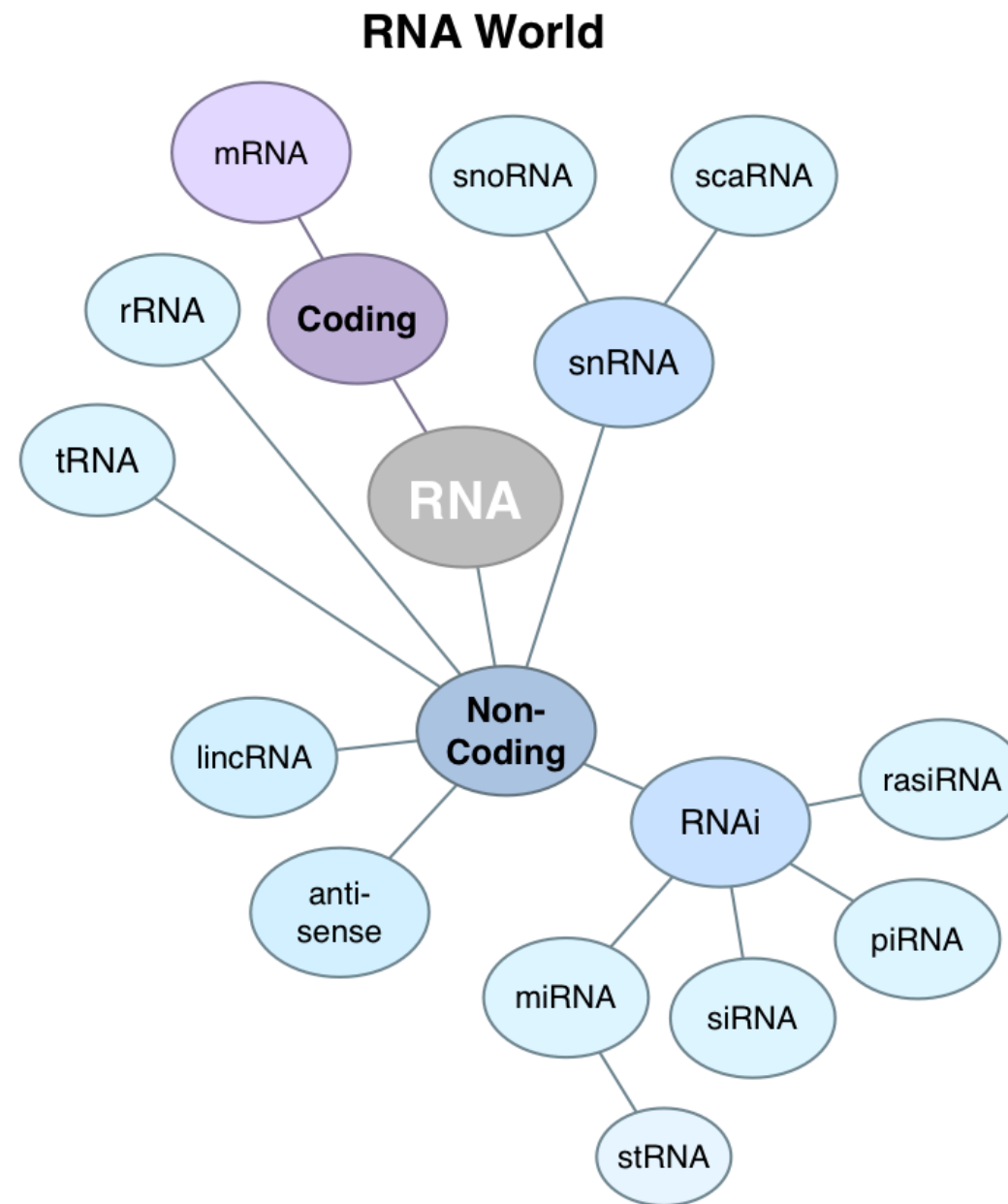
rRNA



Forms the ribosome



Different types of RNAs and their functions





Different types of RNAs and their functions

Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10 ⁶ (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10 ⁷	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10 ⁵	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10 ³	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10 ³	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10 ⁵	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10 ⁵	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10 ⁵	0.02	0.001 to 0.003	About 10 ⁵ molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10 ⁴	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10 ³	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10 ³	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Ramsköld et al. (2009), Menet et al. (2012)

**The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1976).*

***Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.*