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

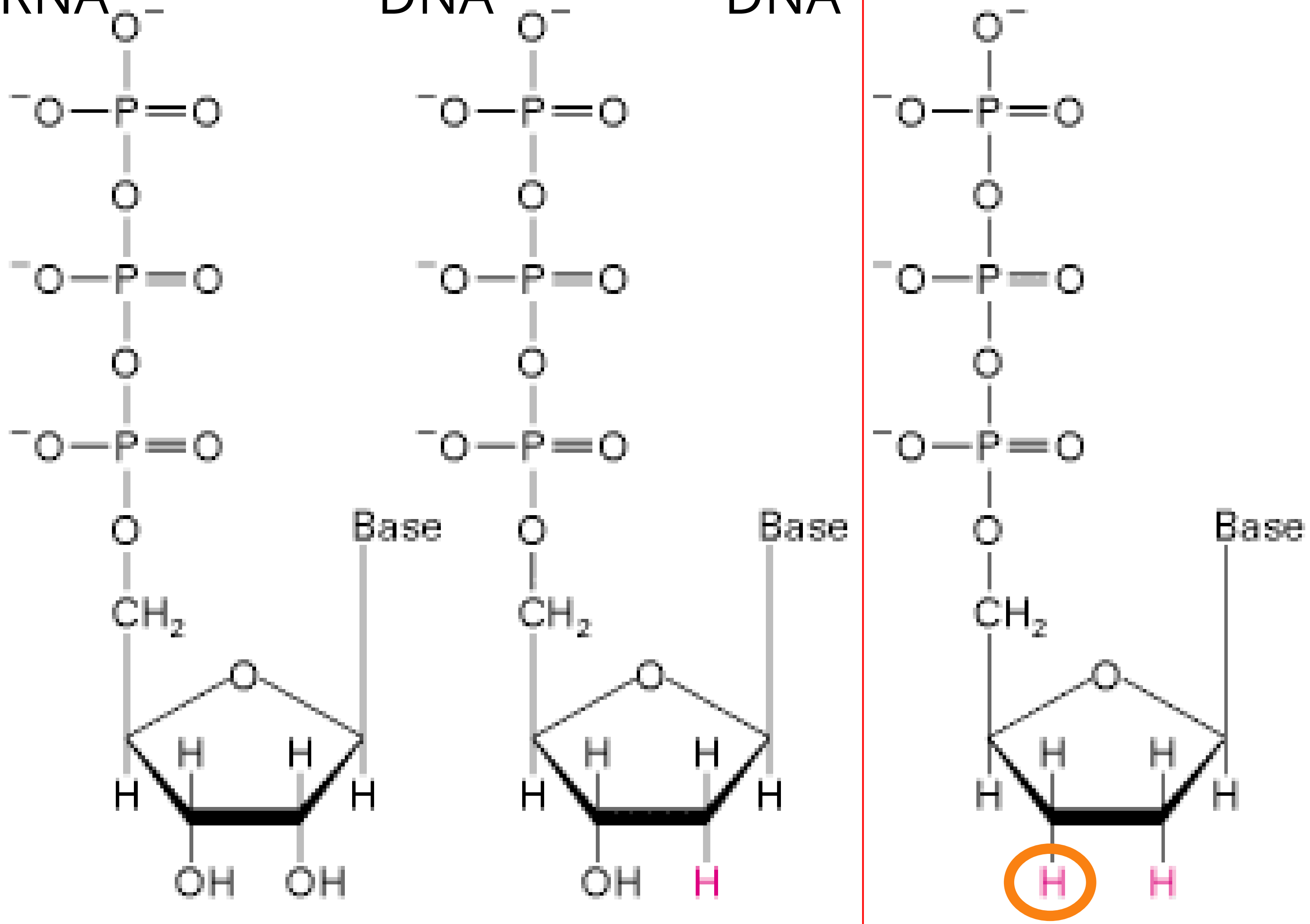
B. Next generation sequencing: many reactions at the same time

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 DNA DNA



Ribonucleoside triphosphate

Deoxyribonucleoside triphosphate

Dideoxyribonucleoside triphosphate

rNTP

dNTP

ddNTP: no 3′ -OH

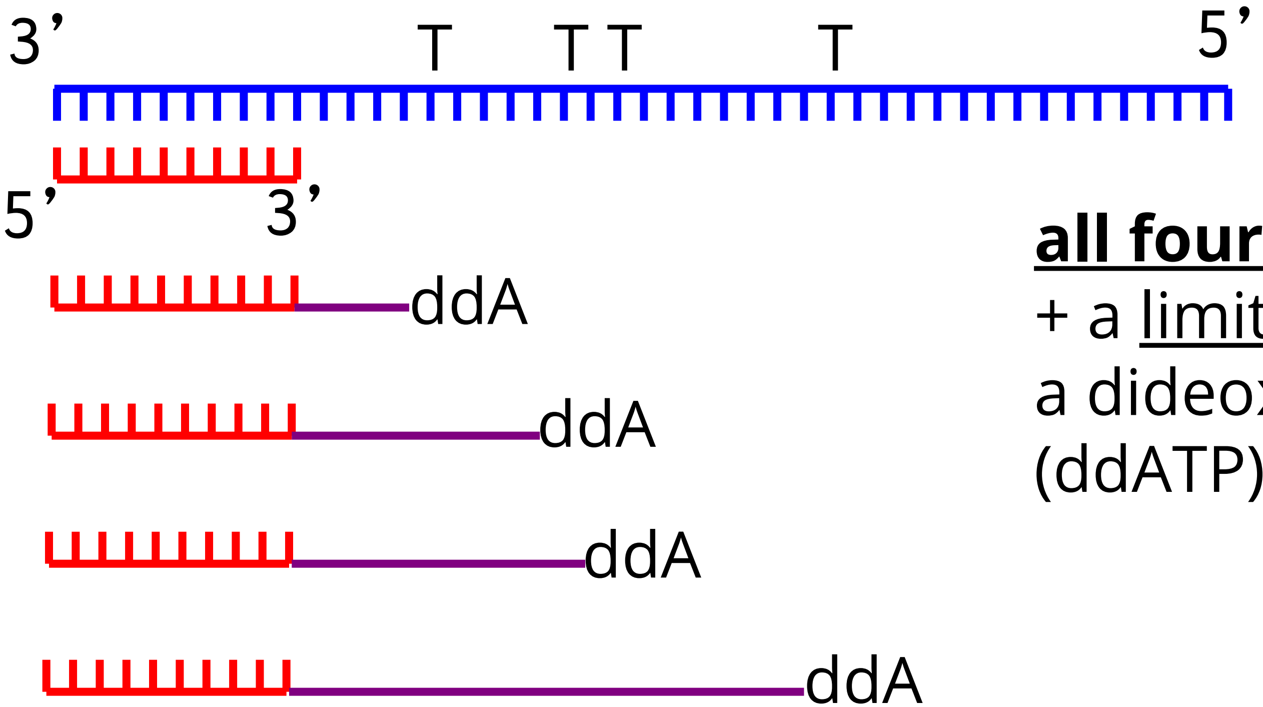
DNA polymerase for sequencing

- 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

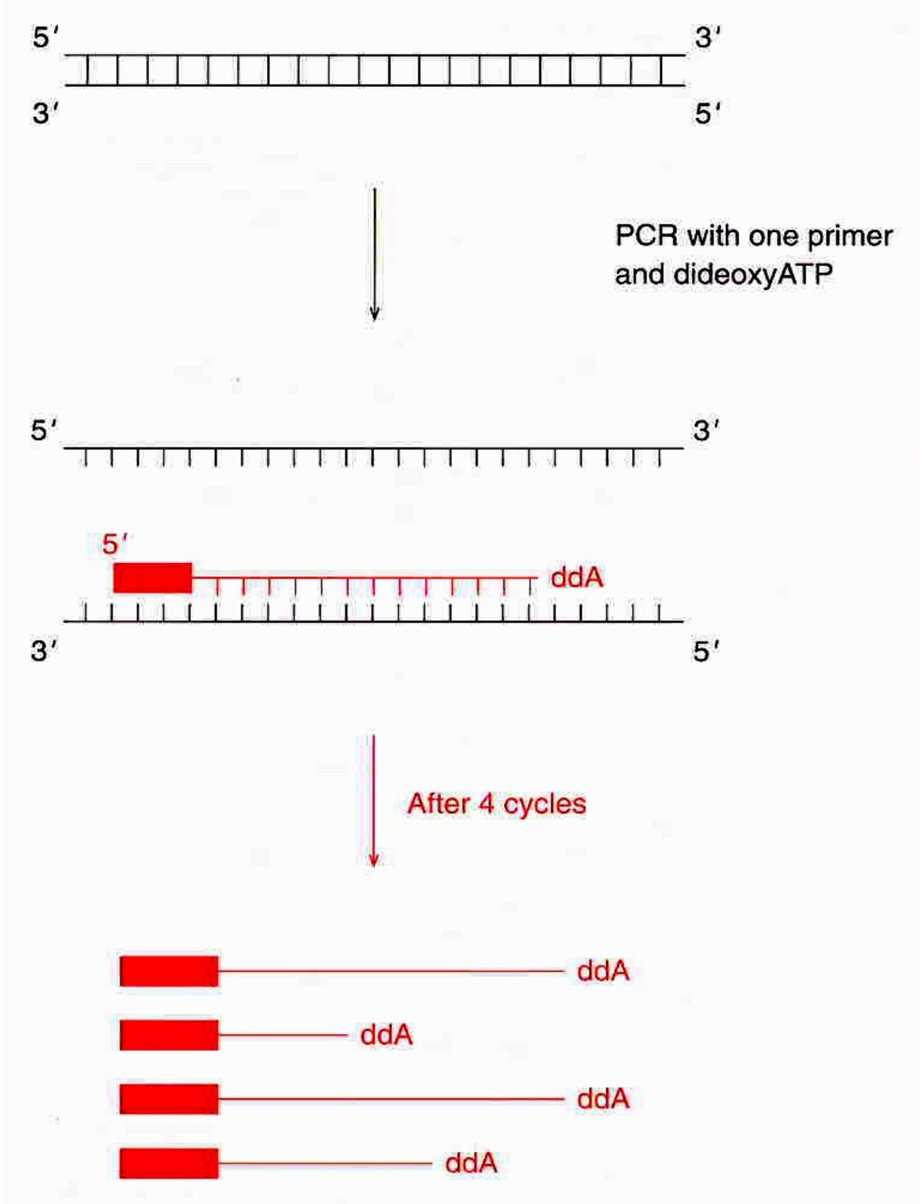


all four dNTPs

+ 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

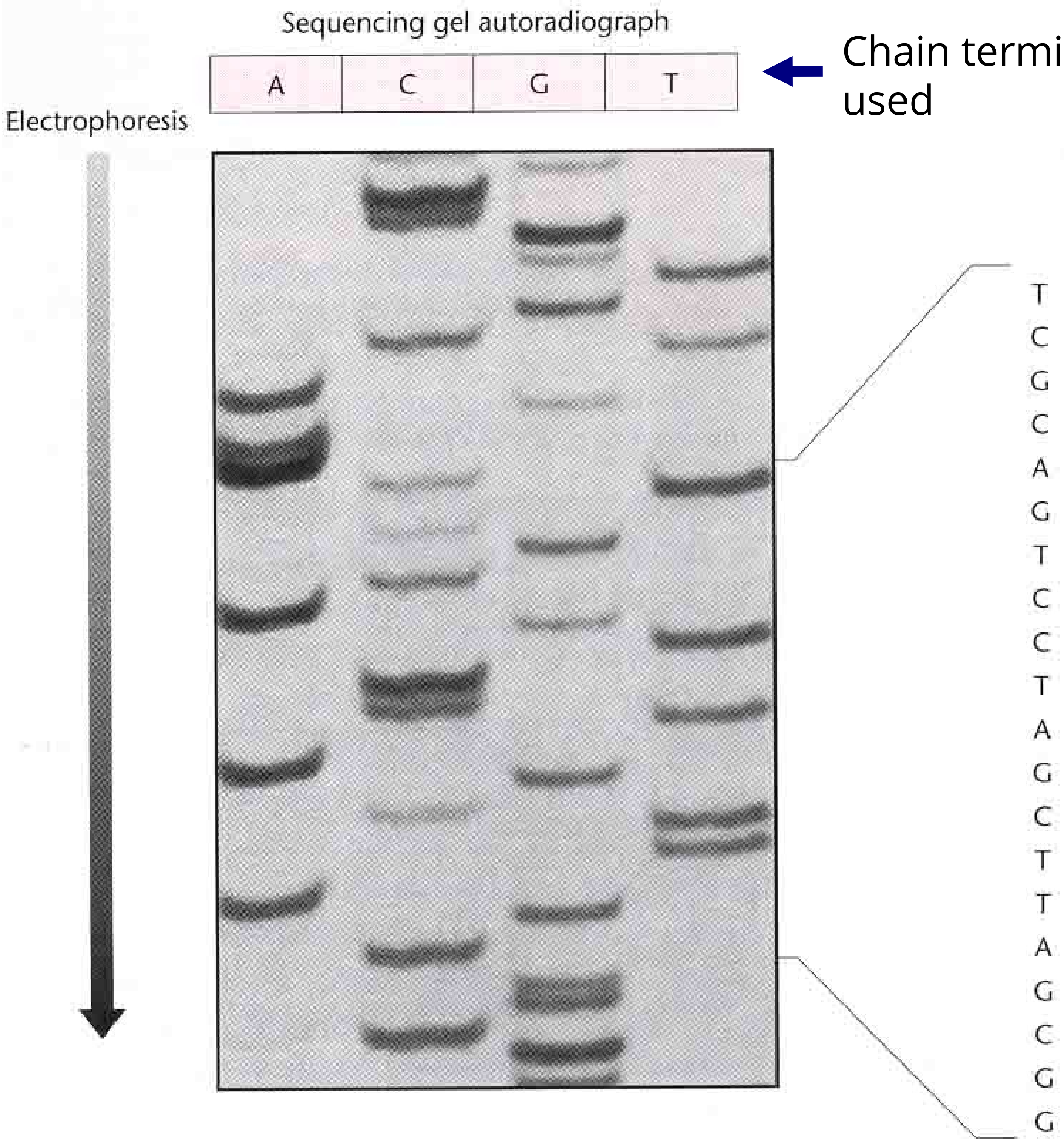
- Radioactivity
  - Radiolabeled primers (kinase with ^{32}P)
  - Radiolabelled dNTPs (γ ^{35}S or ^{32}P)
- Fluorescence
  - ddNTPs chemically synthesized to contain a different fluorophore
  - Each fluorophore is a different color

Analysis of sequencing products:

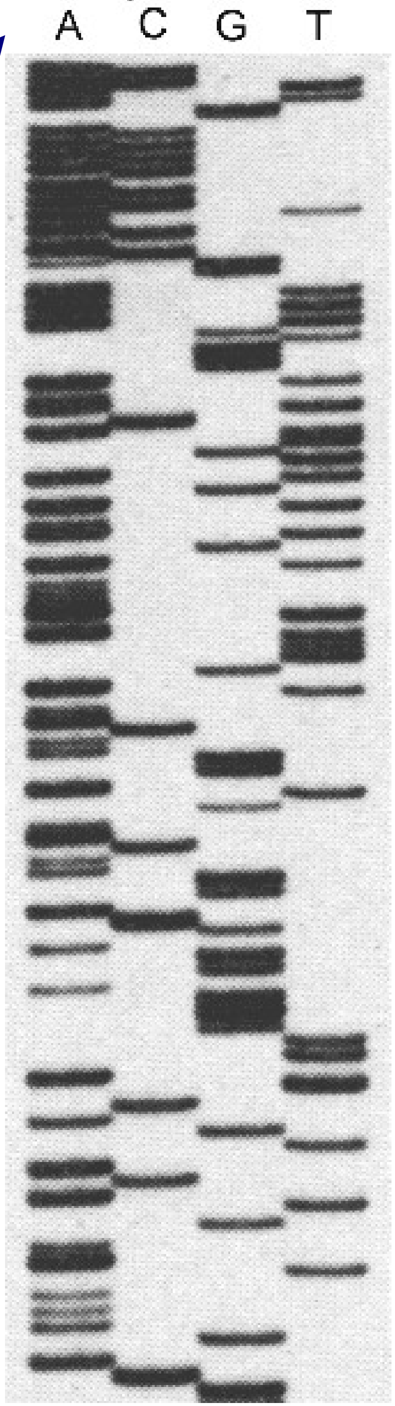
Polyacrylamide gel electrophoresis -- resolves of fragments differing by a single dNTP

- ‘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



Chain terminator used



Animation of cycle sequencing:

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

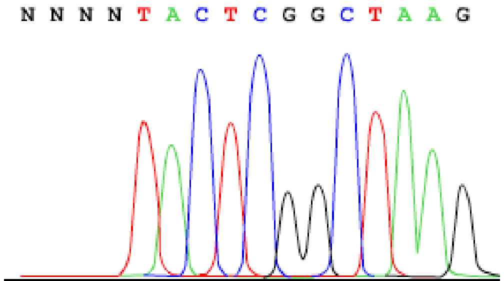
Sequencing in a typical lab

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

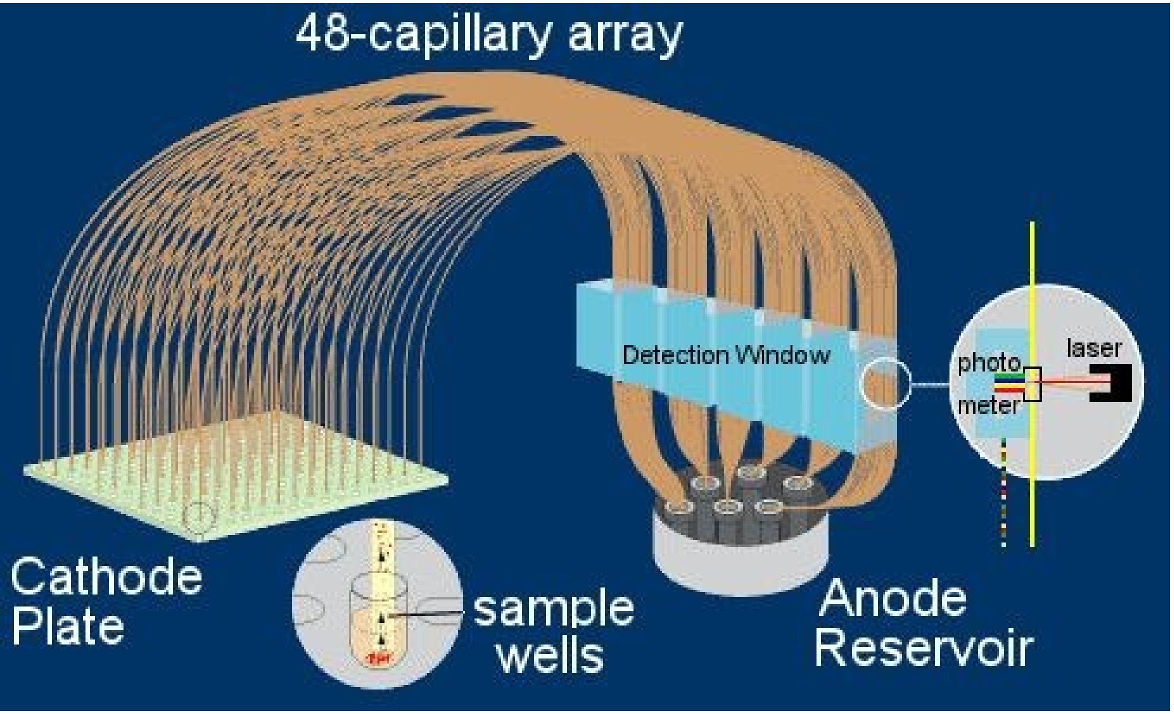
-- costly equipment and materials
-- time consuming protocols

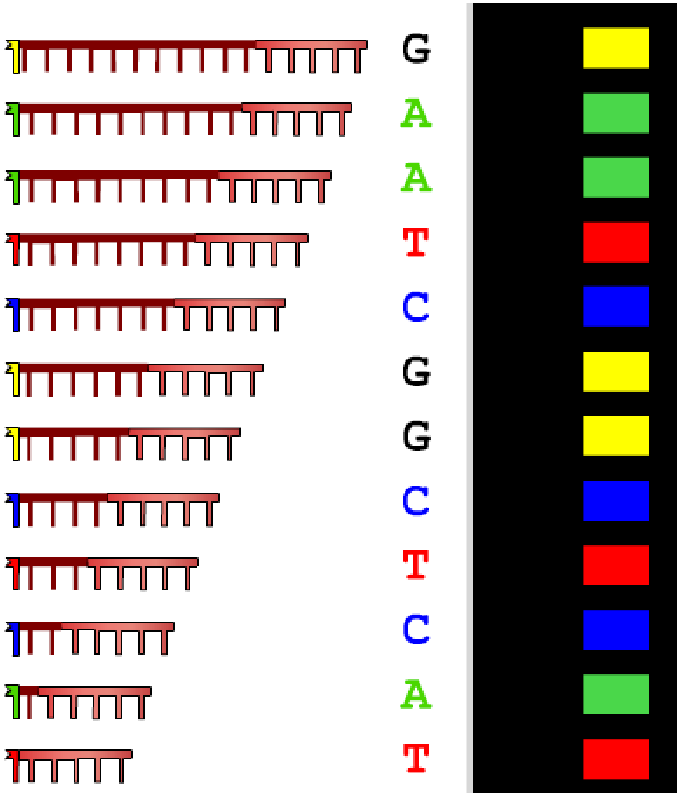
Most labs send out for sequencing:

- 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

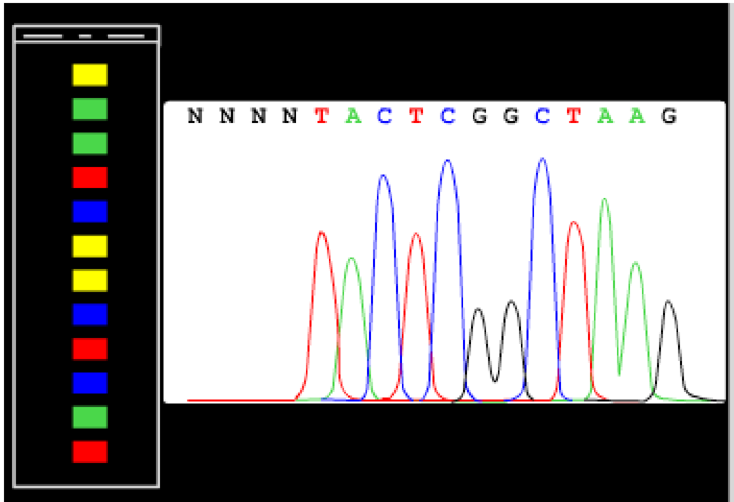


An automated sequencer





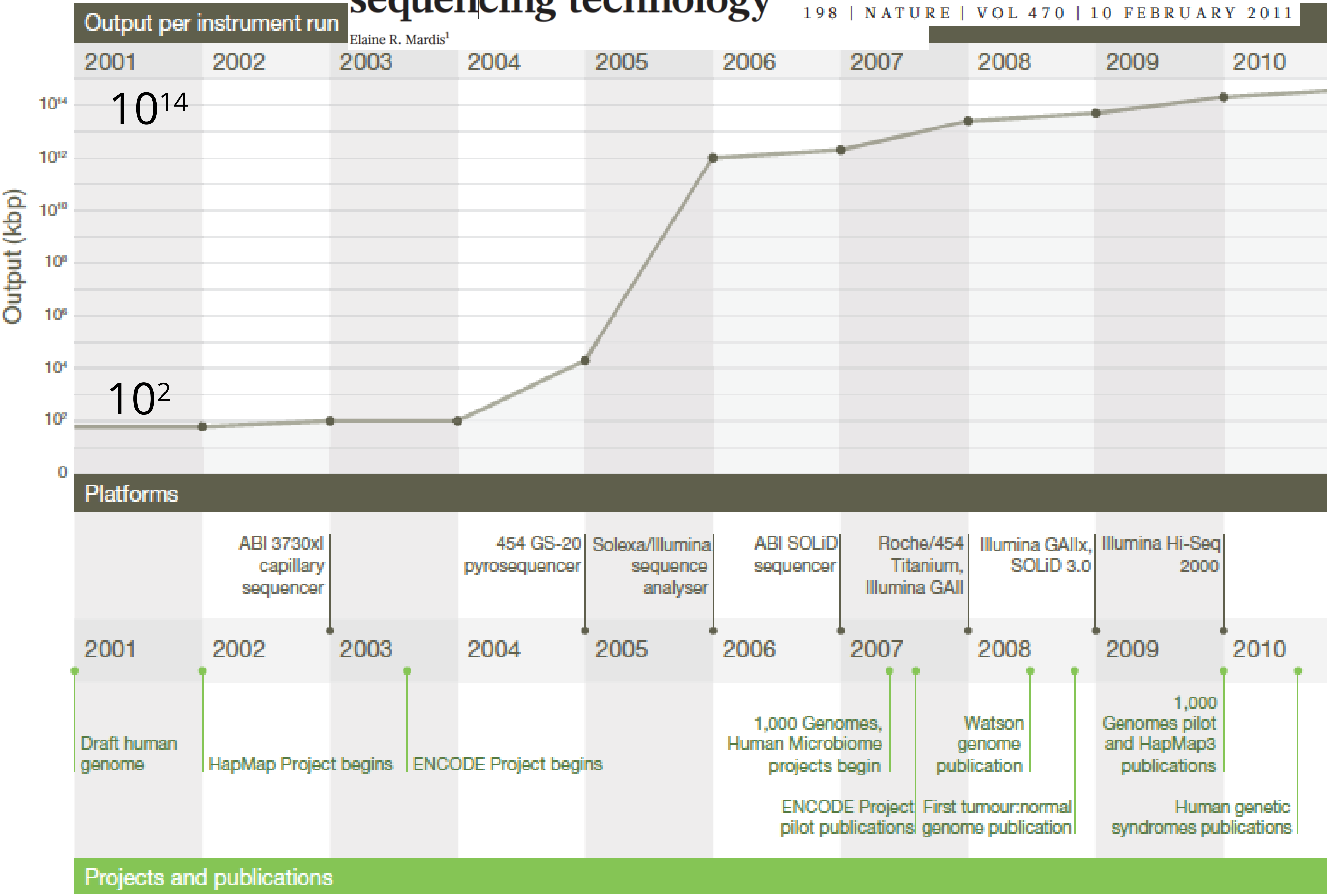
The data: electropherogram



A decade's perspective on DNA sequencing technology

198 | NATURE | VOL 470 | 10 FEBRUARY 2011

Elaine R. Mardis[1]



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^+ 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)

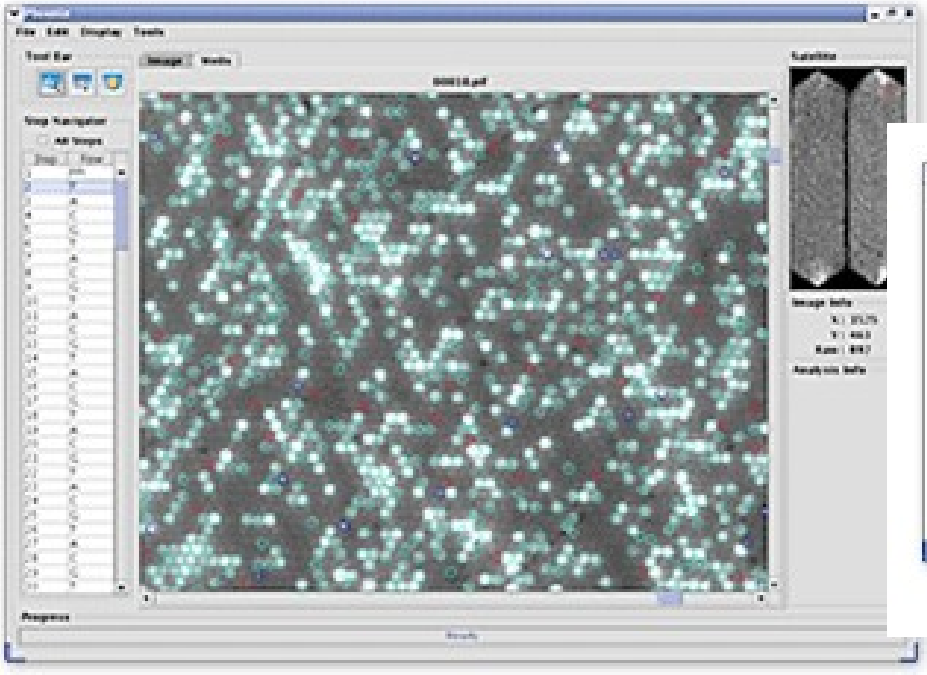
Separate each bead on a plate with up to 1.6 million wells

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

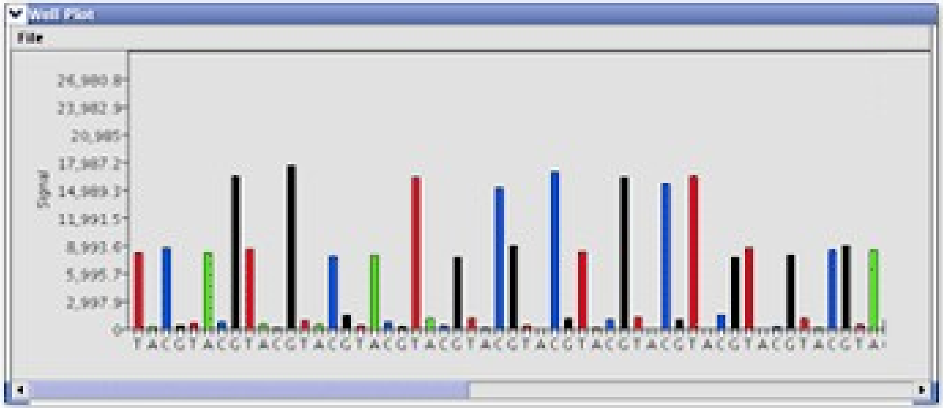
Sequence by DNA polymerase-dependent chain extension, one base at a time in the presence of a reporter (luciferase)

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



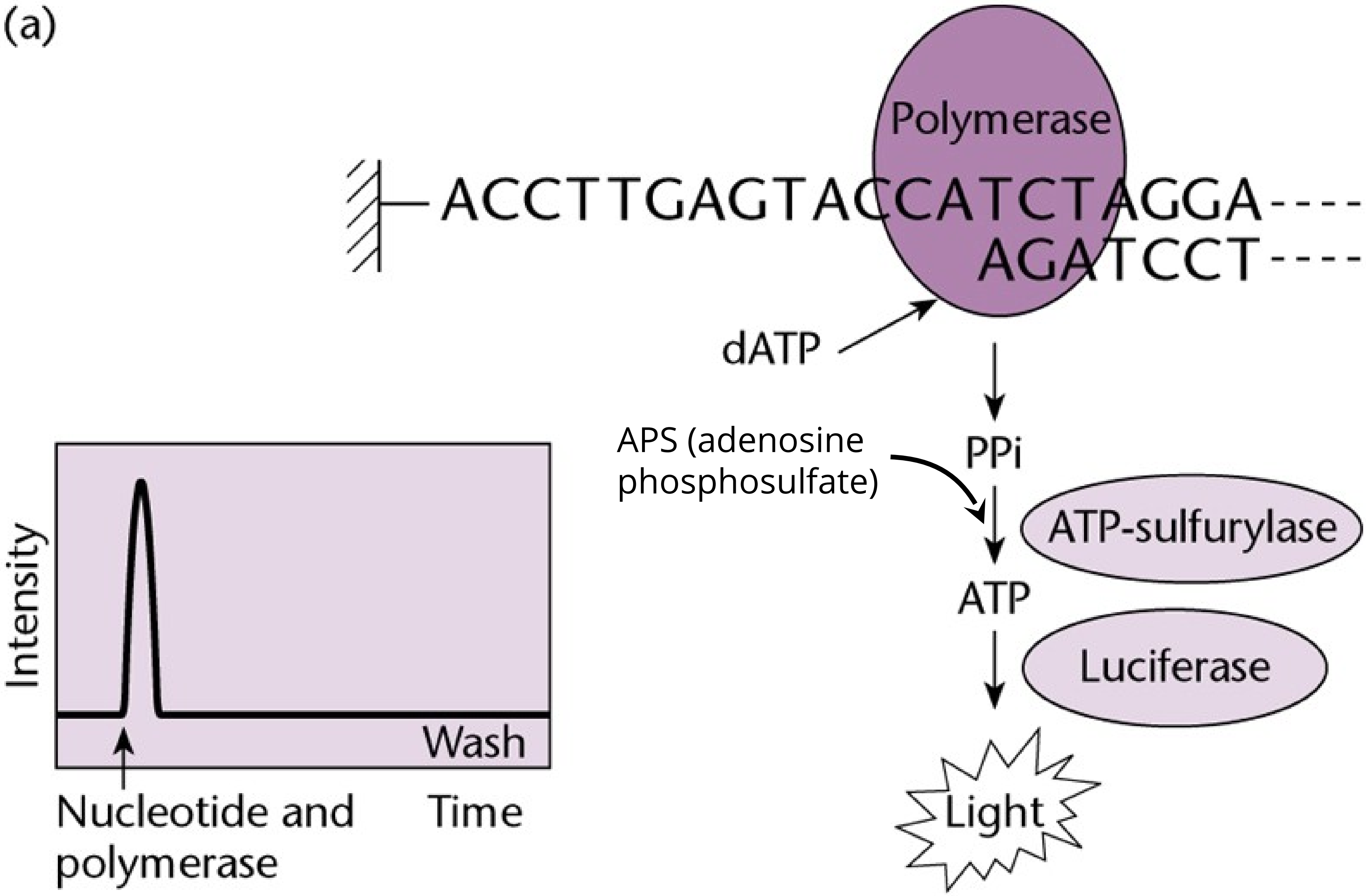
DATA ANALYSIS: OUTPUT PACKAGE



FLOWGRAM (SIGNAL OUTPUT FROM A SINGLE WELL)

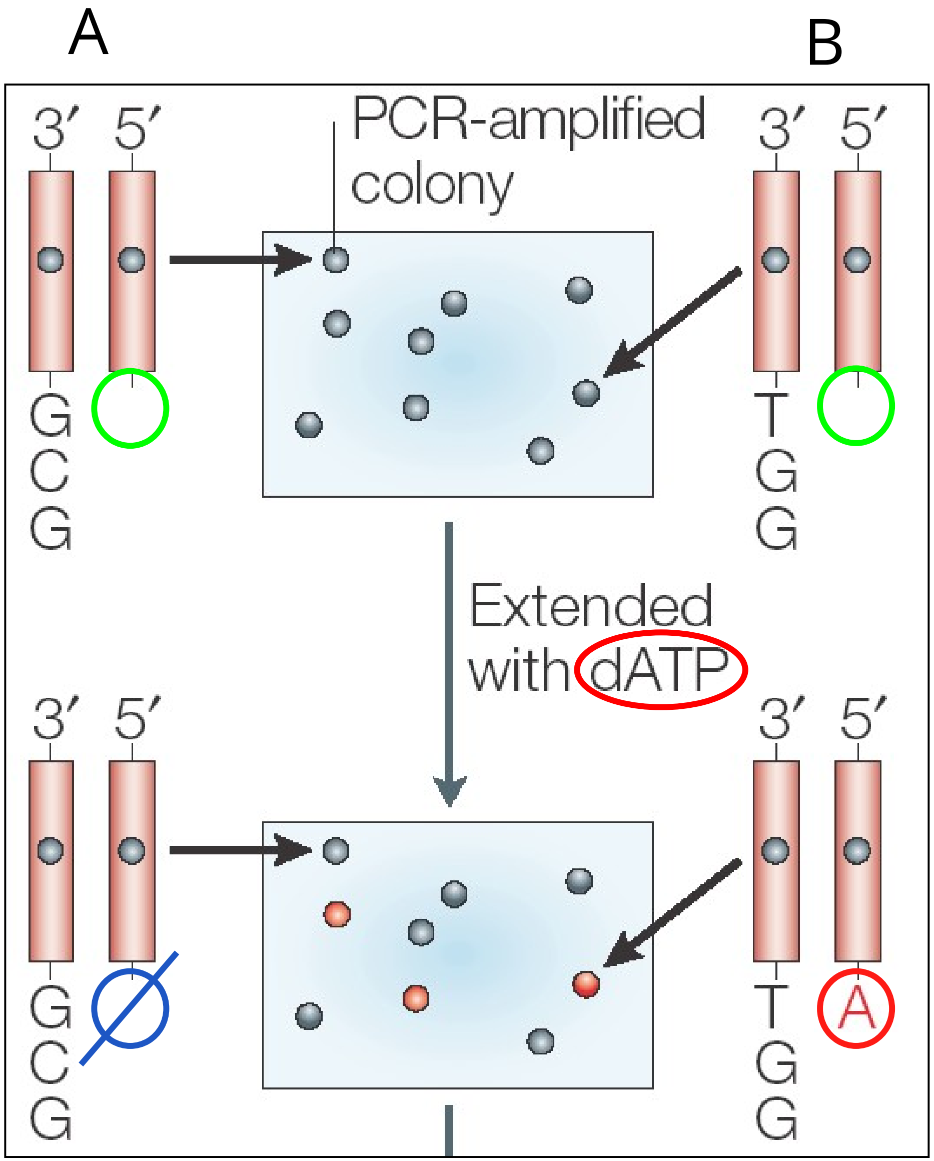
Read length: about 200 bp

How the extension reaction is detected (version A)

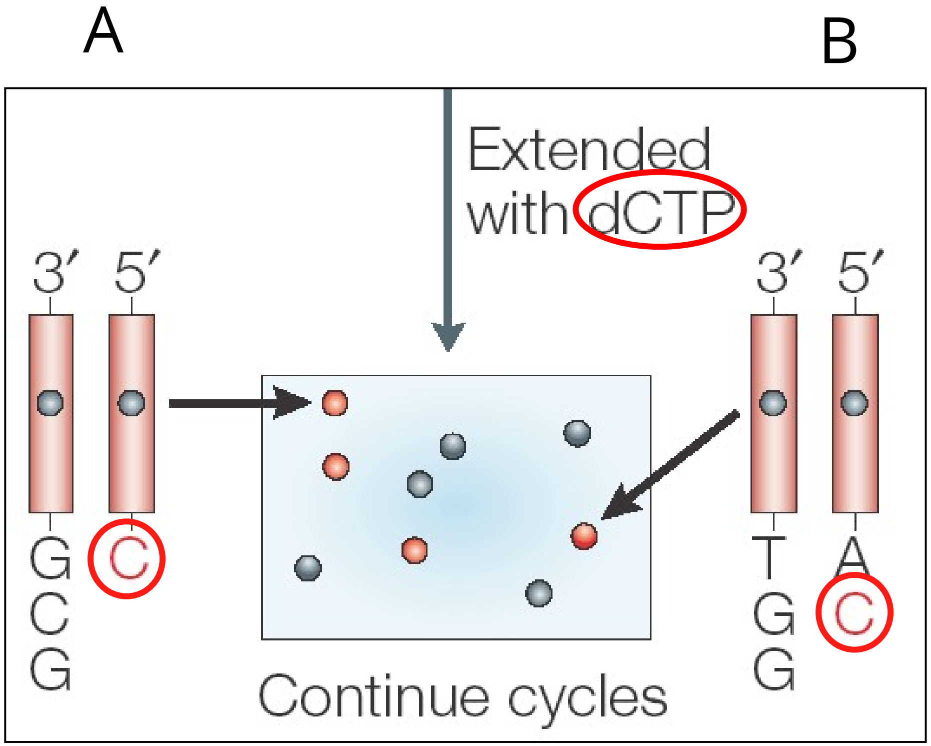


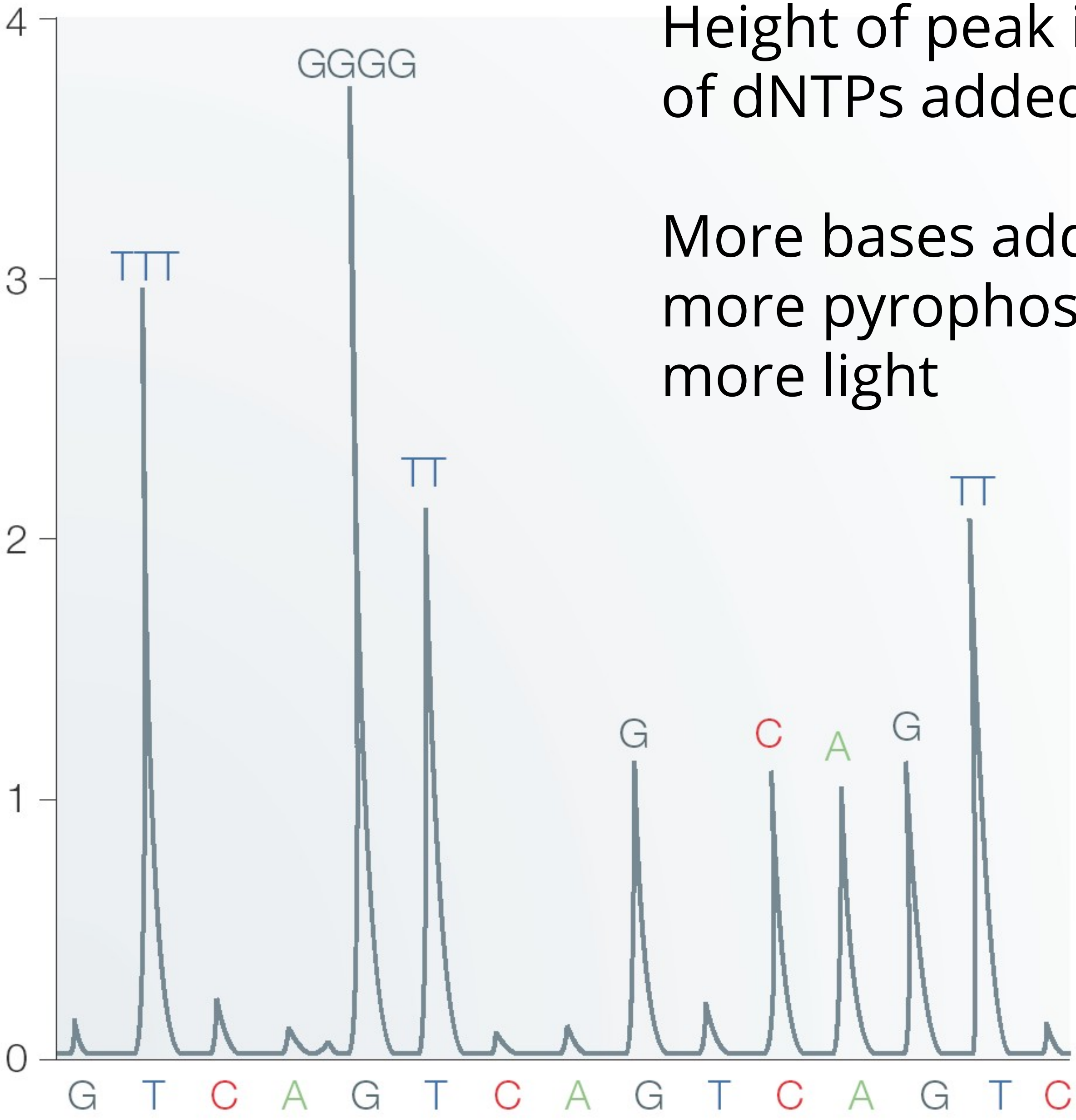
APS (adenosine phosphosulfate)

Extension with individual dNTPs gives a readout



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



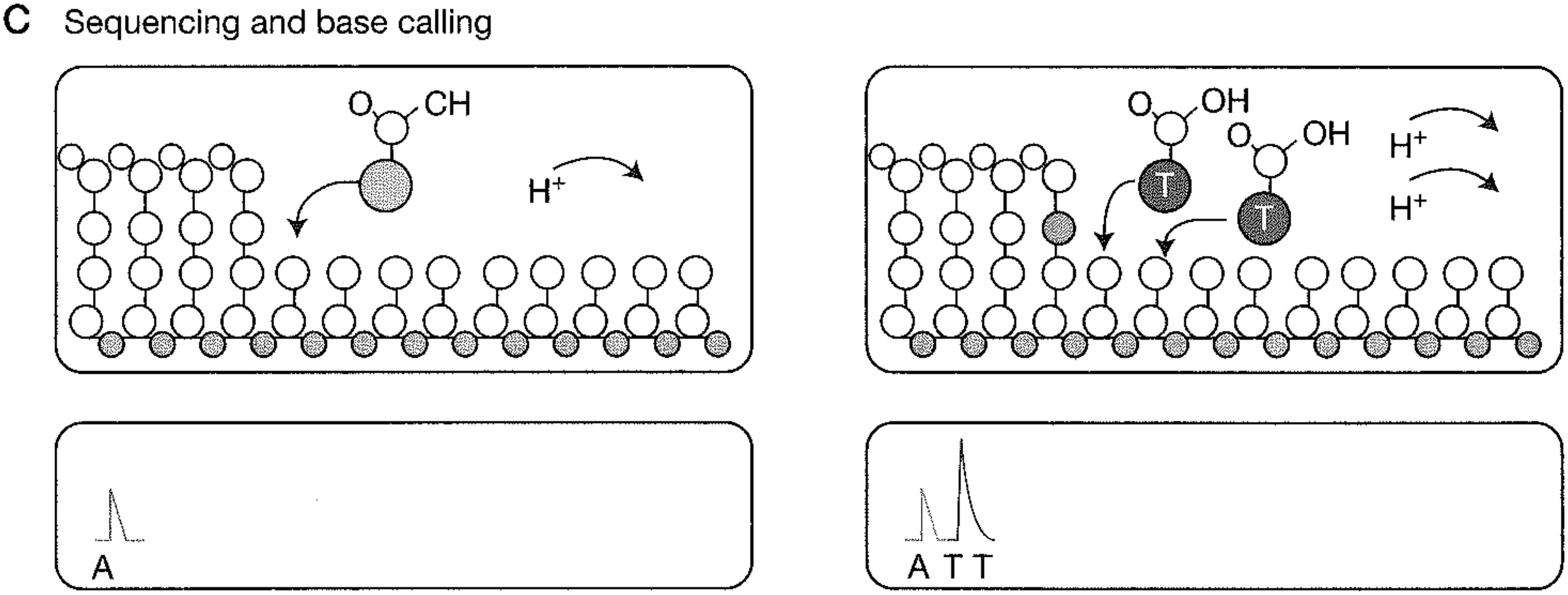


Height of peak indicates the number of dNTPs added

More bases added = more pyrophosphate = more light

This sequence: TTTGGGGTTGCAGTT

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



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^+ 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

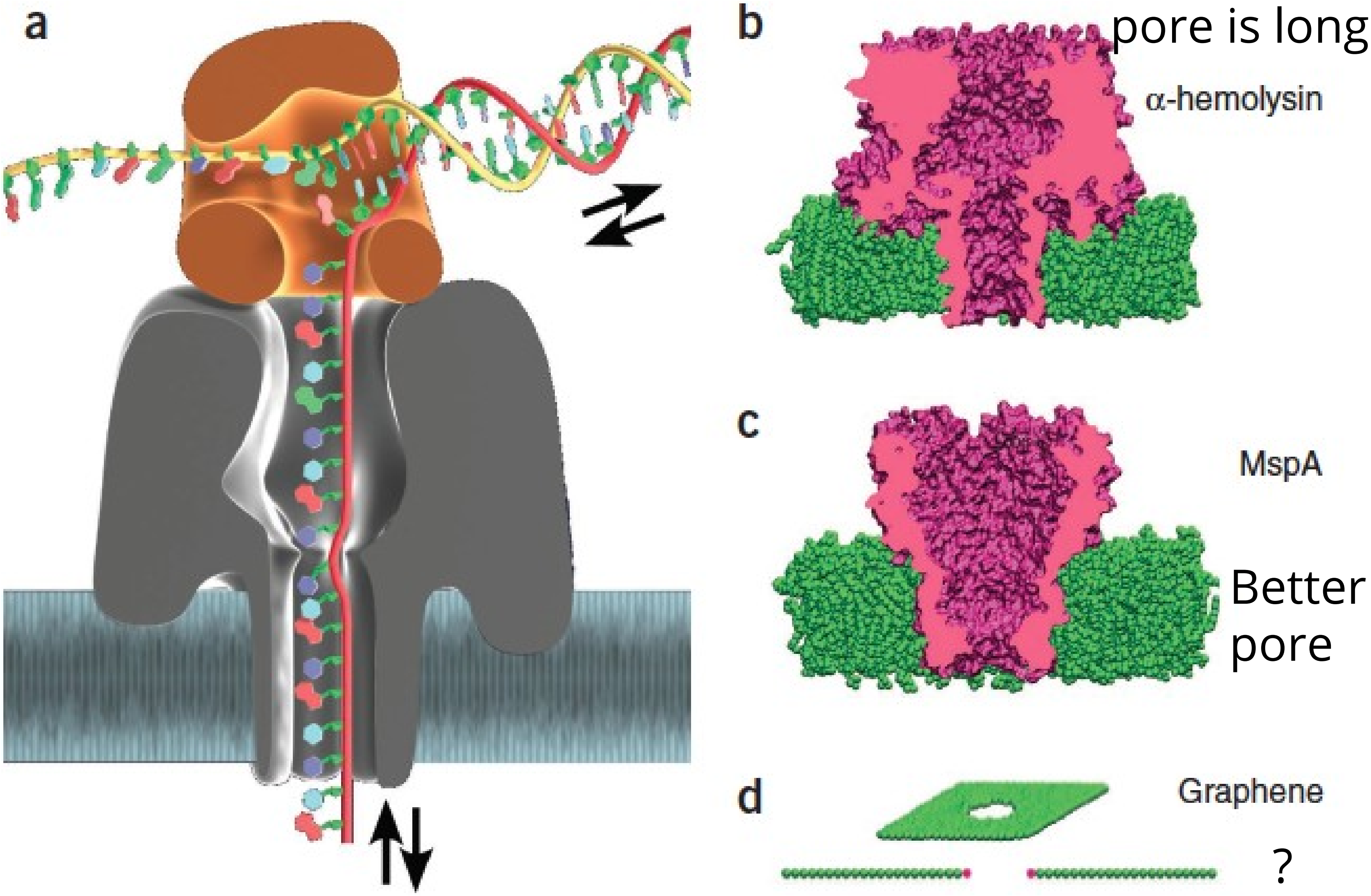
4) Pore sequencing: DNA through pore, record each base

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



Restrictive region of pore is long

Better pore

?

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

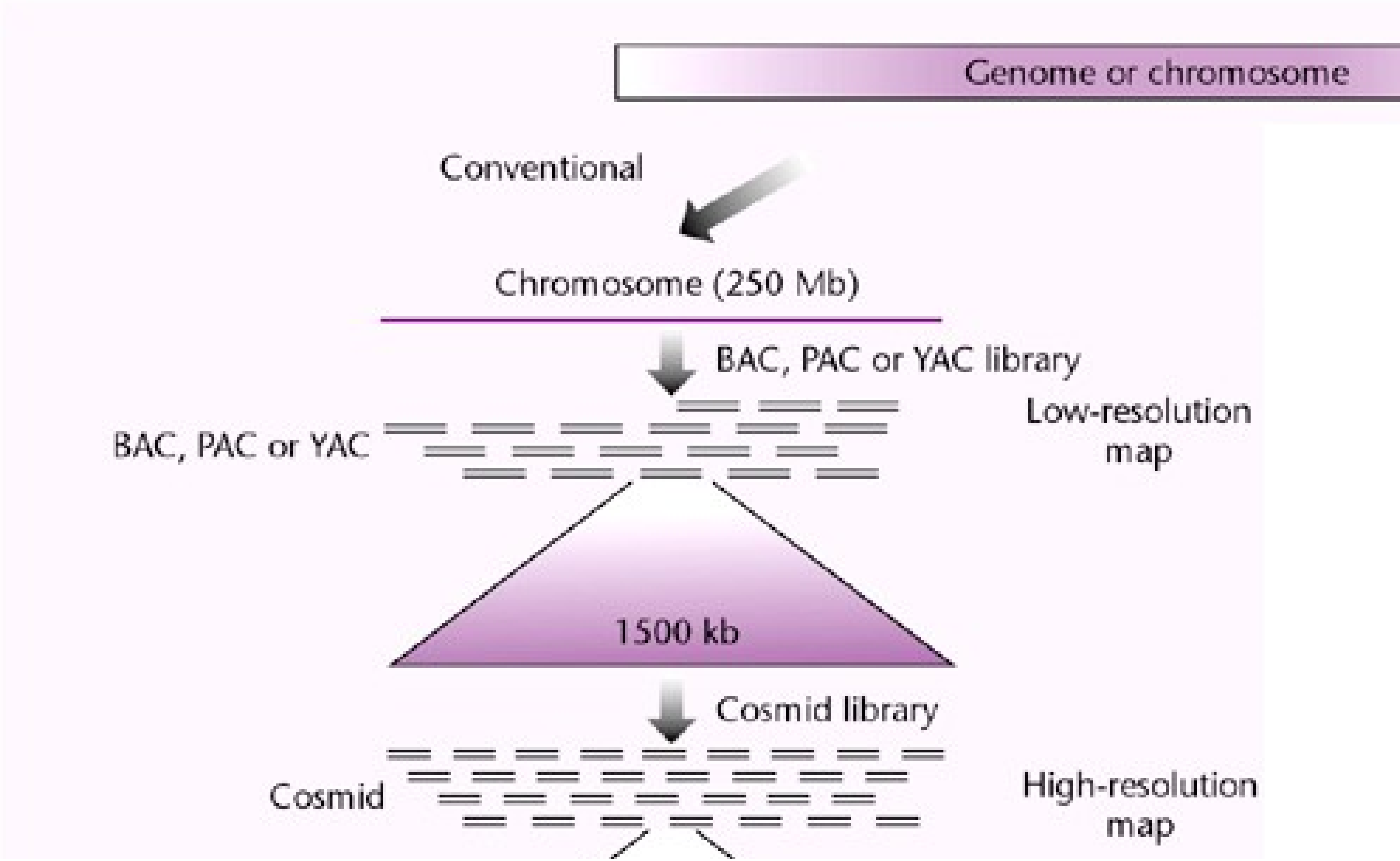
- Very high speed sequencing, so it could be useful for speedy diagnosis in clinic: e.g. ID infectious agent to help give best treatment

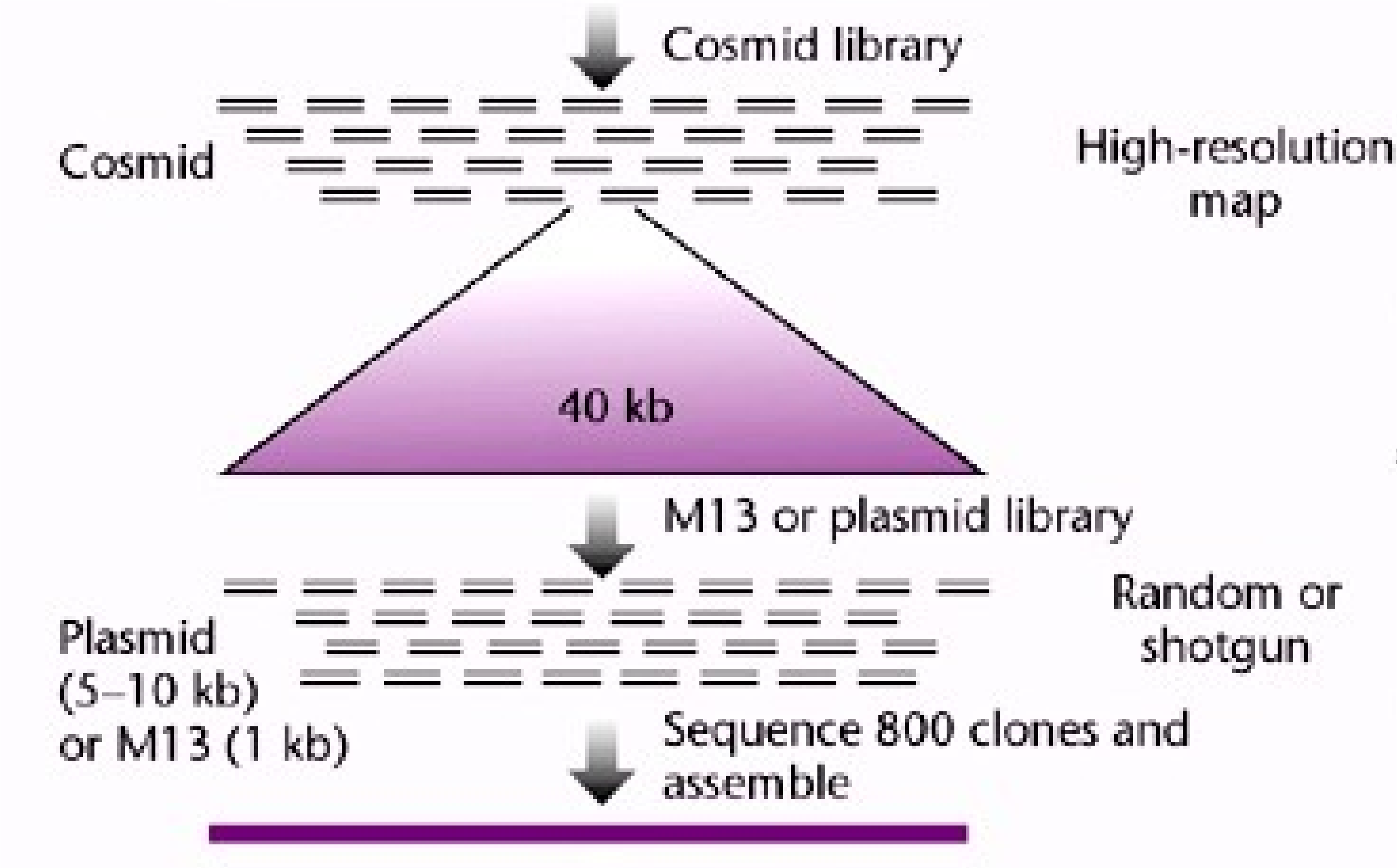
Movies

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

Whole genome sequences:

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



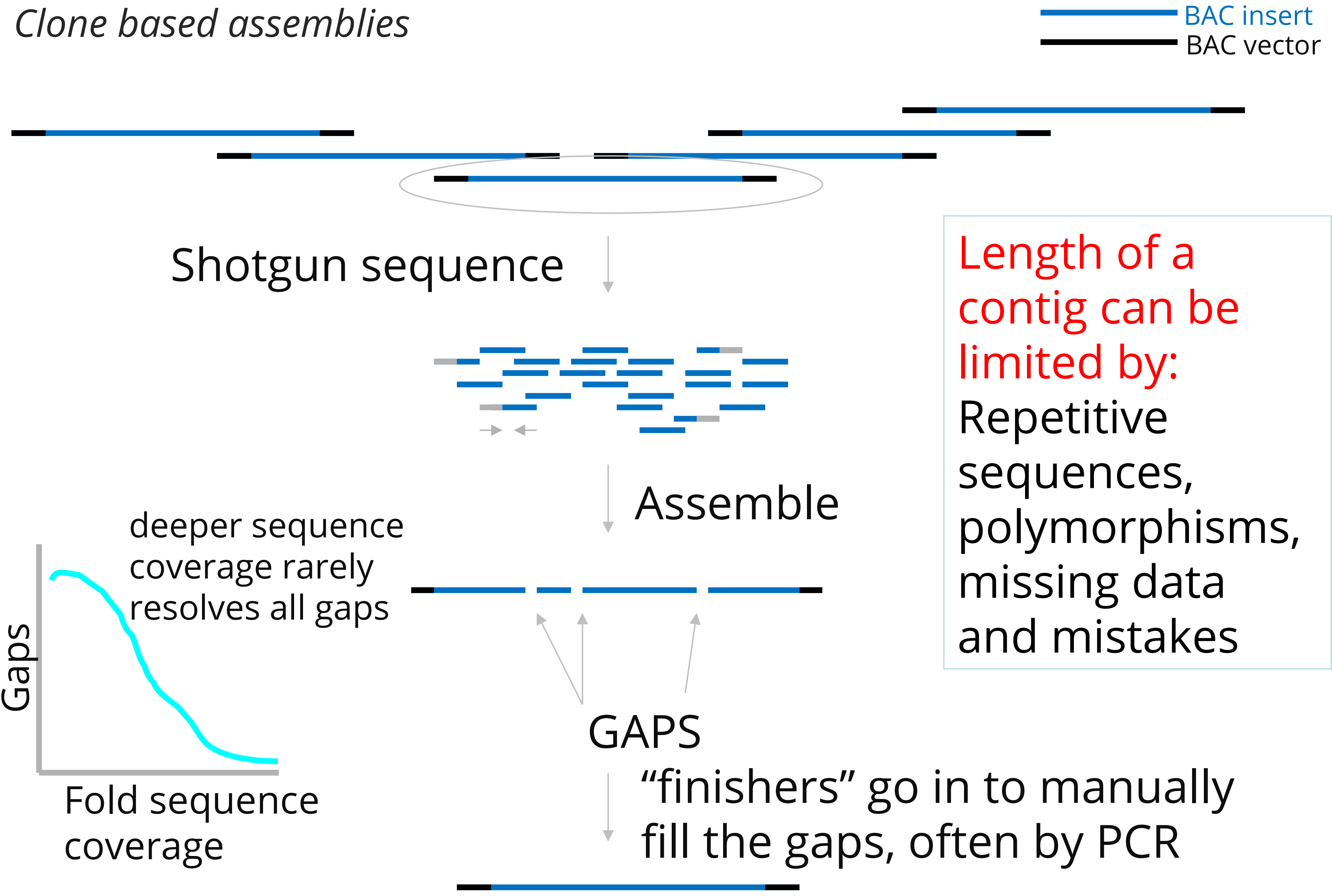


Sequencing is easy, mapping/assembly is more difficult

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



“finishers” go in to manually fill the gaps, often by PCR

Length of a contig can be limited by:

Repetitive sequences, polymorphisms, missing data and mistakes

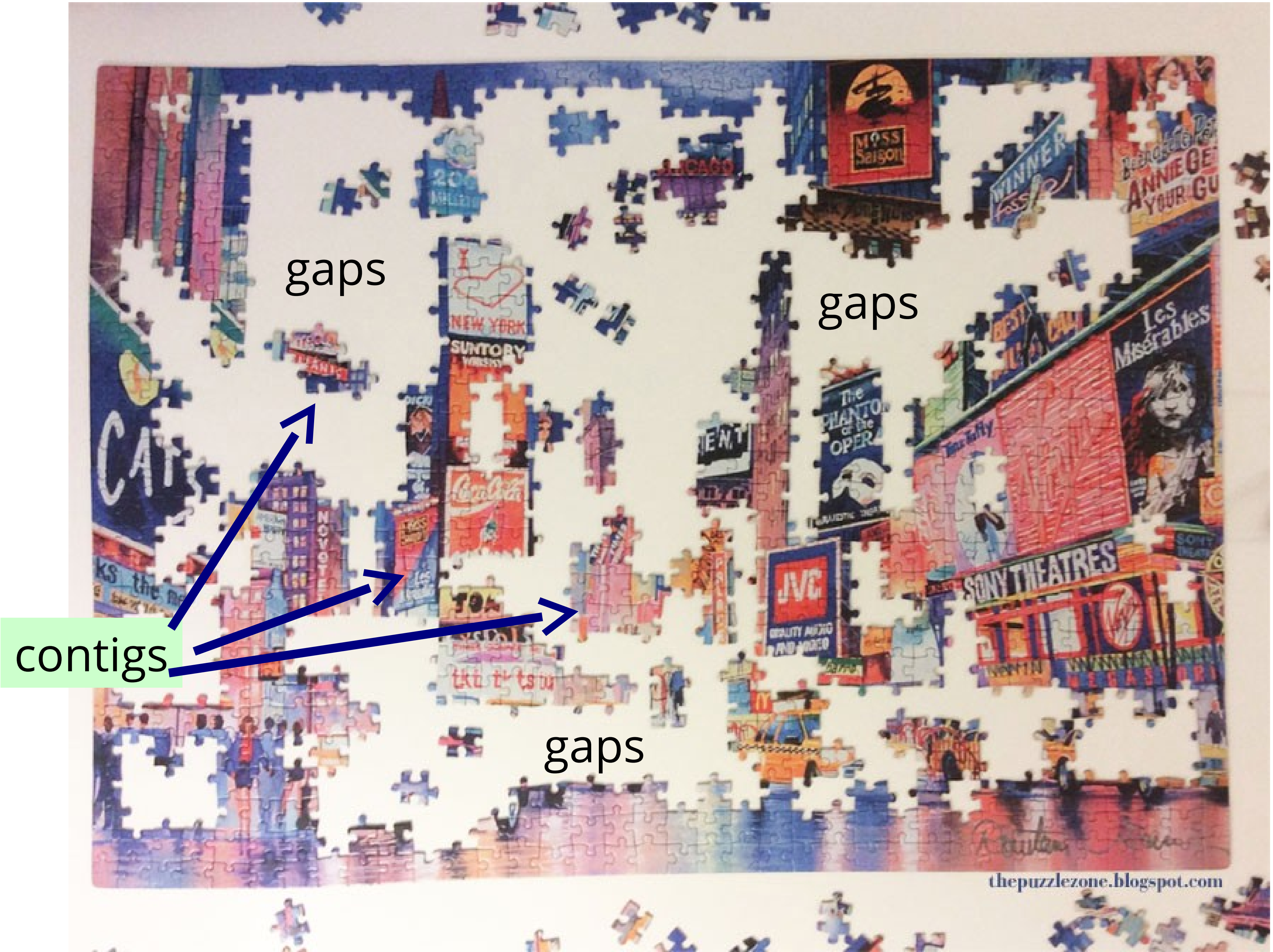
Vocabulary

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



gaps

gaps

gaps

contigs

thepuzzlezone.blogspot.com

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*’

(M. Baker, Nature Methods | VOL.9 NO.4 | APRIL 2012 | 333)

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)

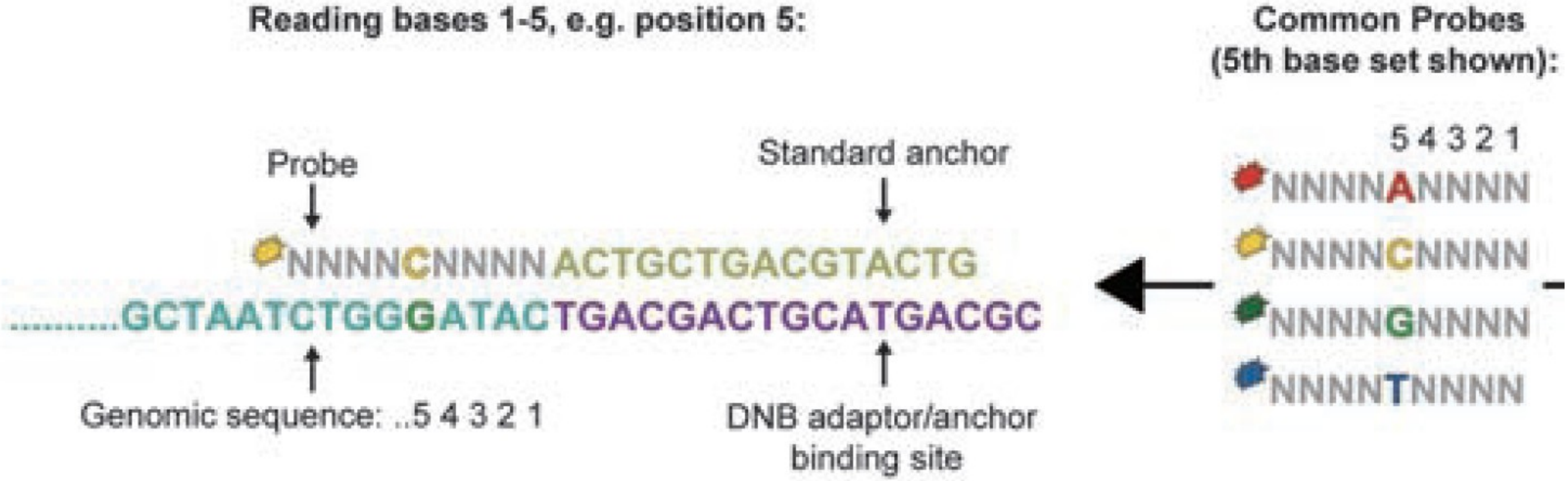
- (see perspective by Worley and Gibbs, 2010)

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:
Probe
Standard anchor
NNNNCNNNNN ACTGCTGACGTACTG
..........GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..5 4 3 2 1
DNB adaptor/anchor binding site
Common Probes (5th base set shown):
5 4 3 2 1
NNNNANNNN
NNNNCNNNN
NNNNGNNNN
NNNNTNNNN

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