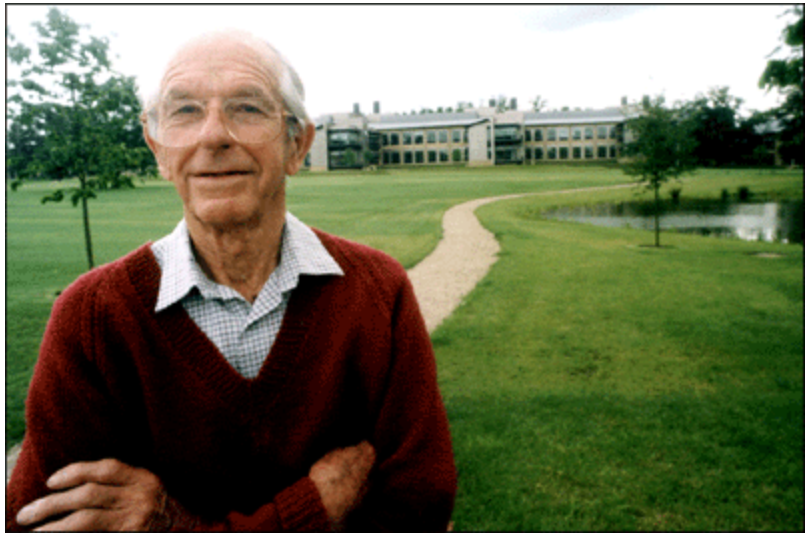
DNA sequencing

Matt Hudson

DNA Sequencing

- **Dideoxy sequencing was developed by Fred Sanger at Cambridge in the 1970s. Often called “Sanger sequencing”.**



Nobel prize number 2 for Fred Sanger in 1980, shared with Walter Gilbert from Harvard (inventor of the now little-used Maxam-Gilbert sequencing method).

Sanger's Dideoxy DNA sequencing method -How it works:

1. DNA template is denatured to single strands.

2. DNA primer (with 3' end near sequence of interest) is annealed to the template DNA and extended with DNA polymerase.

3. Four reactions are set up, each containing:

    1. DNA template – eg a plasmid
    2. Primer
    3. DNA polymerase
    4. dNTPS (dATP, dTTP, dCTP, and dGTP)

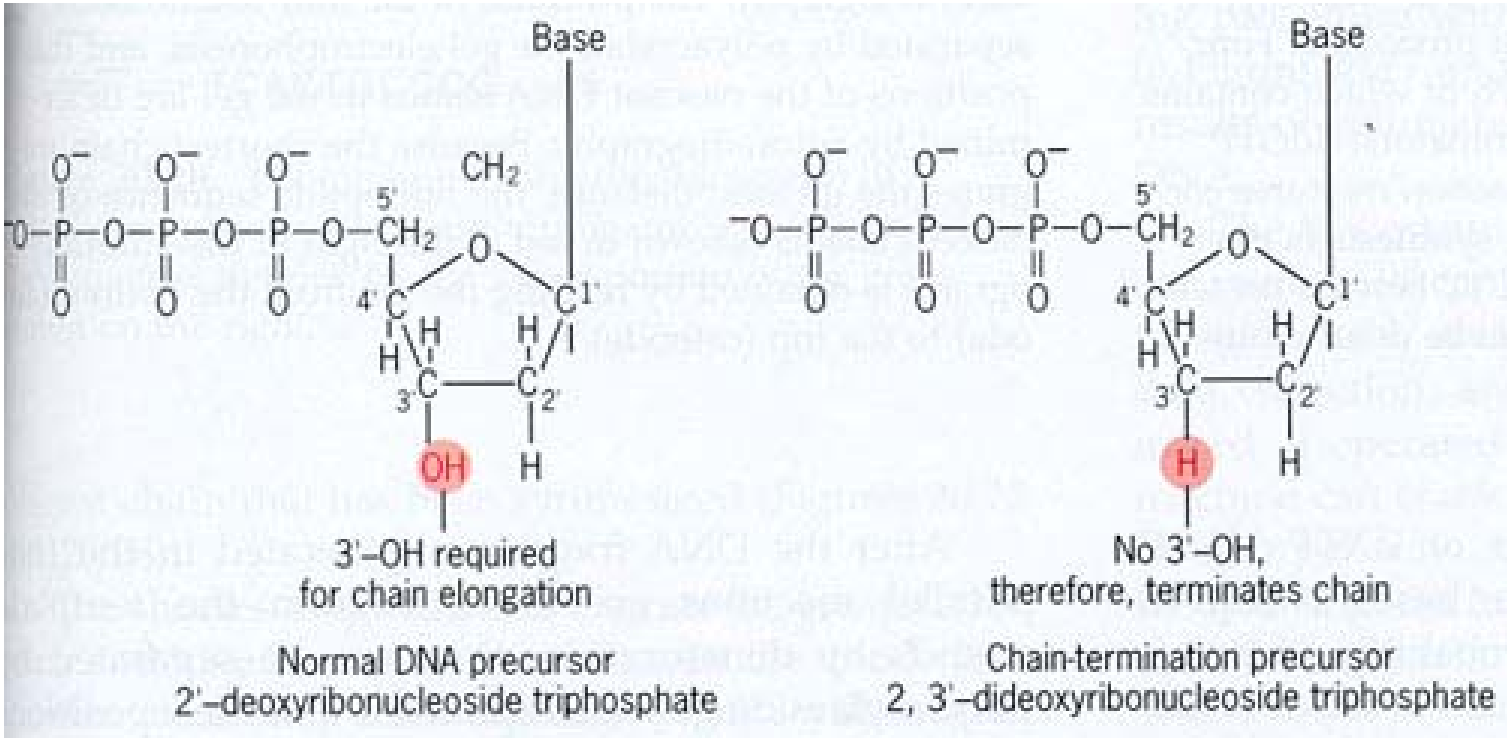
4. Next, a different radio-labeled dideoxynucleotide (ddATP, ddTTP, ddCTP, or ddGTP) is added to each of the four reaction tubes at 1/100th the concentration of normal dNTPs……


THE
TERMINATOR

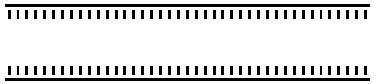
Terminators stop further elongation of a DNA deoxyribose-phosphate backbone

ddNTPs are terminators: they possess a 3'-H instead of 3'-OH, compete in the reaction with normal dNTPS, and produce no phosphodiester bond.

Whenever the radio-labeled ddNTPs are incorporated in the chain, DNA synthesis terminates.

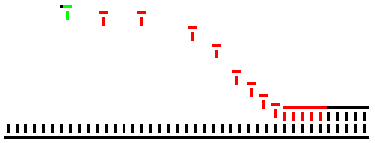


purify DNA



separate strands

prime synthesis



elongate

until

“hasta la vista”

Manual Dideoxy DNA sequencing-How it works (cont.):

5. Each of the four reaction mixtures produces a population of DNA molecules with DNA chains terminating at each “terminator”base..

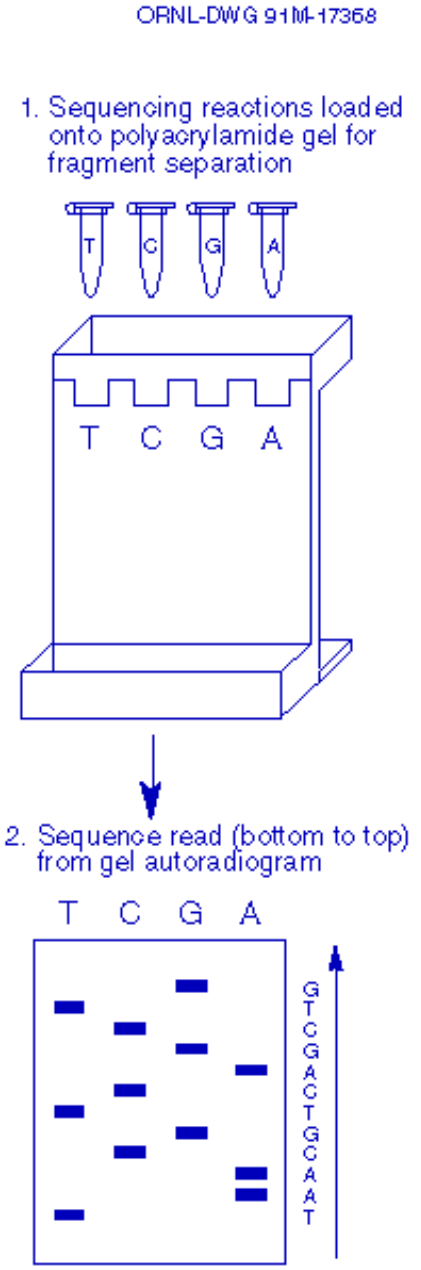
6. Extension products in each of the four reaction mixutes also end with a different radio-labeled ddNTP (depending on the base).

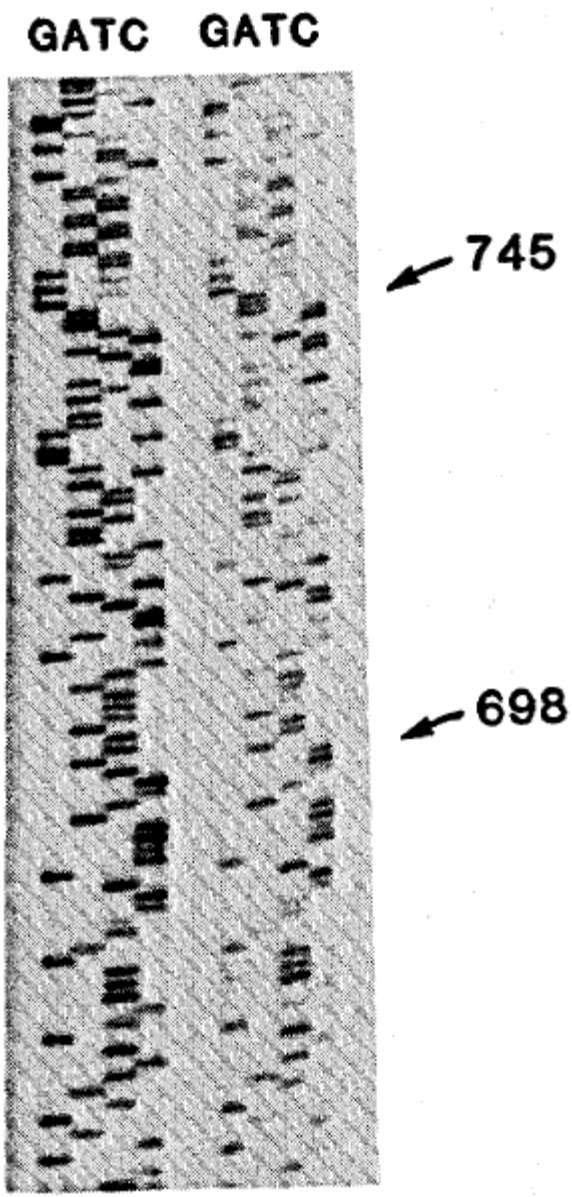
7. Next, each reaction mixture is electrophoresed in a separate lane (4 lanes) at high voltage on a polyacrylamide gel.

8. Pattern of bands in each of the four lanes is visualized on X-ray film.

9. Location of “bands” in each of the four lanes indicate the size of the fragment terminating with a respective radio-labeled ddNTP.

10. DNA sequence is deduced from the pattern of bands in the 4 lanes.



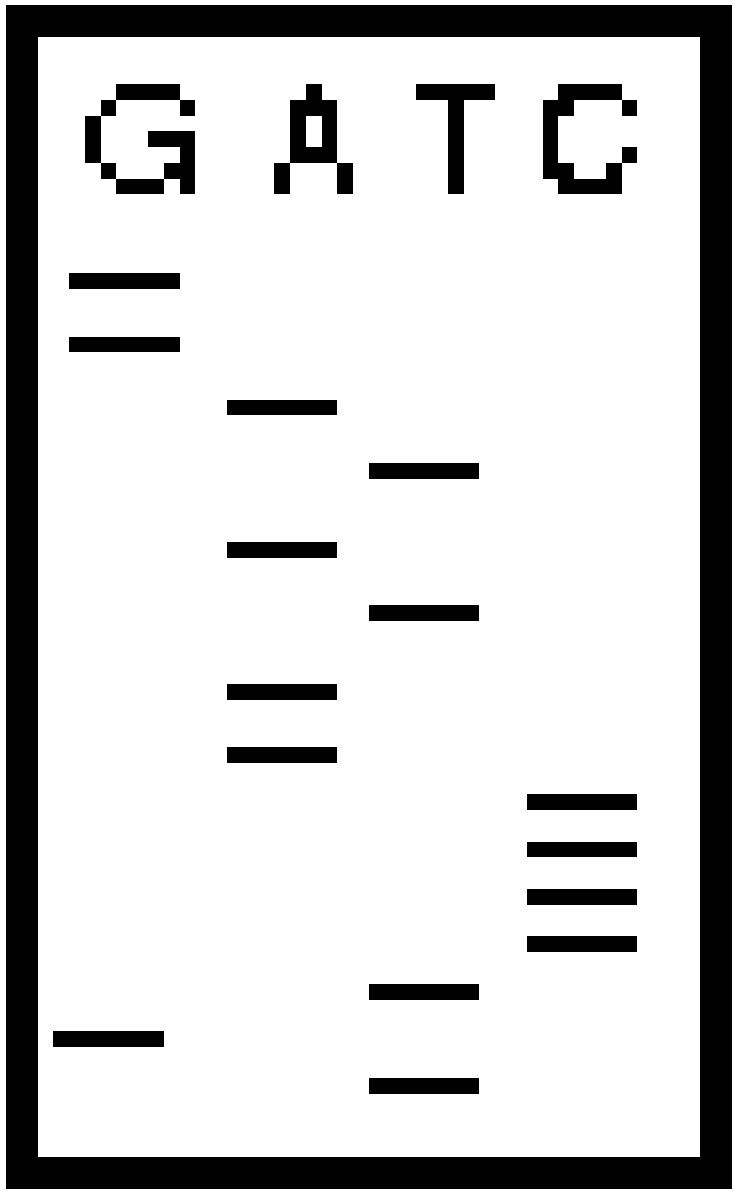


Vigilant et al. 1989
PNAS 86:9350-9354

Radio-labeled ddNTPs (4 rxns)

Sequence (5’ to 3’)

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



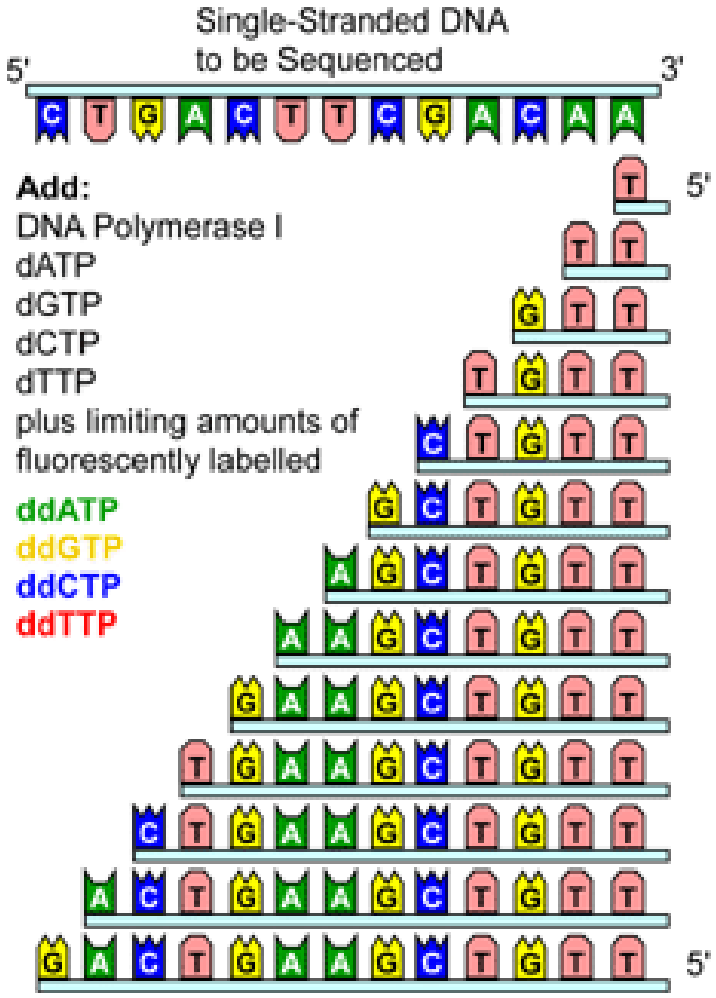
Short products

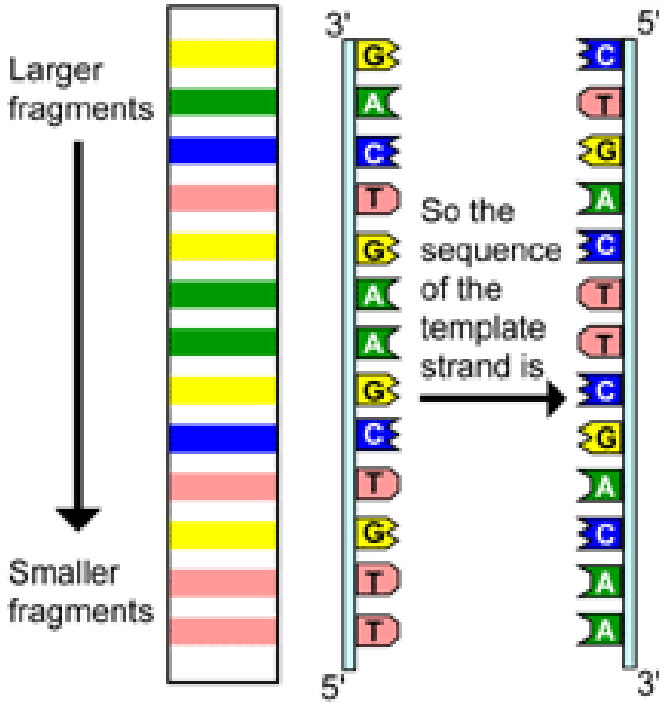
Long products

Manual vs automatic sequencing

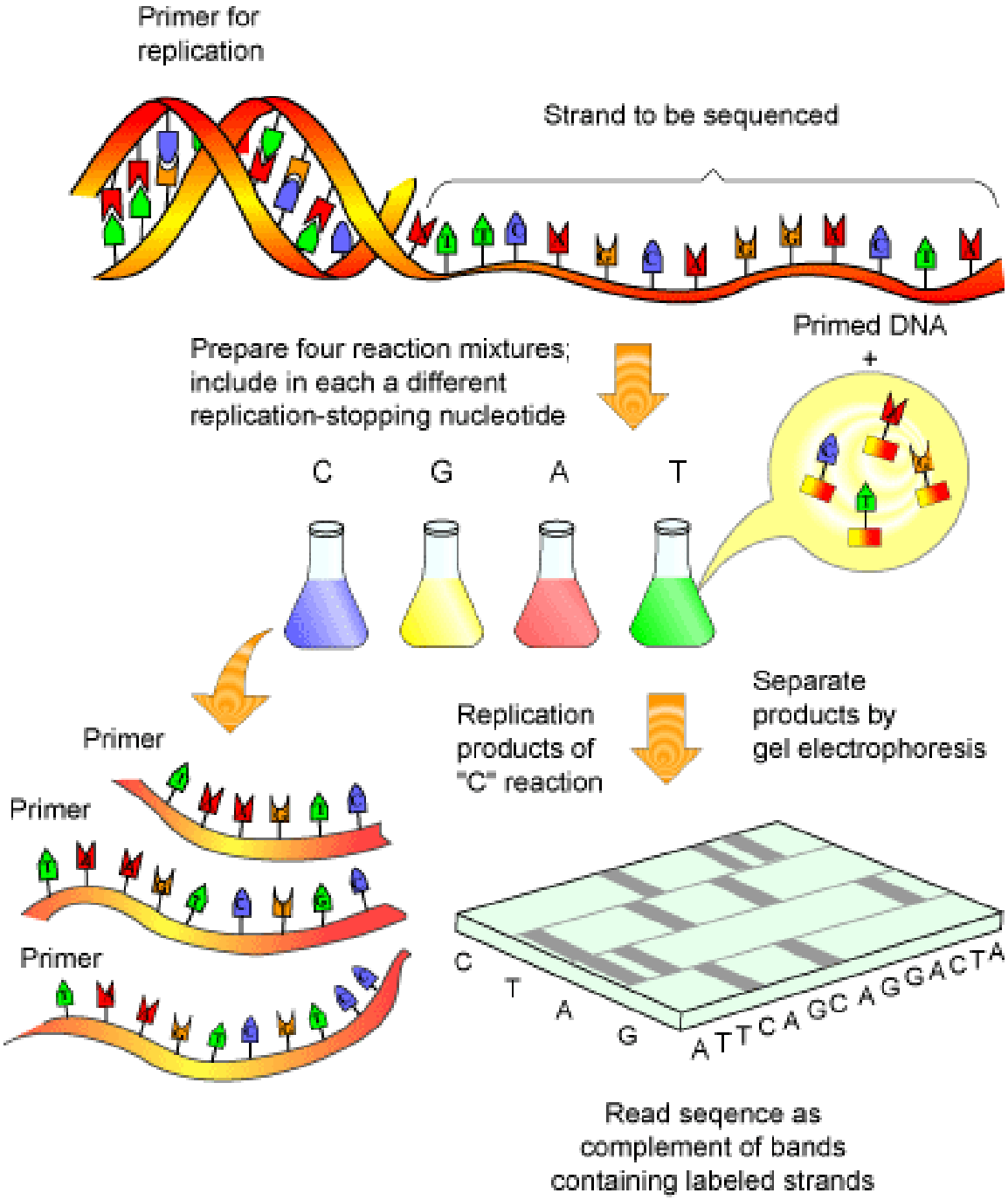
- Manual sequencing has basically died out.
- It needs four lanes, radioactive gels, and a technician in one day from one gel can get four sets of four lanes, with maybe 300 base pairs of data from each template.
- Everyone now uses “automatic sequencing” – the downside is no one lab can afford the machine, so it is done in a central facility (eg. Keck center).
- Most automated DNA sequencers can load robotically and operate around the clock for weeks with minimal labor.

Dye deoxy terminators

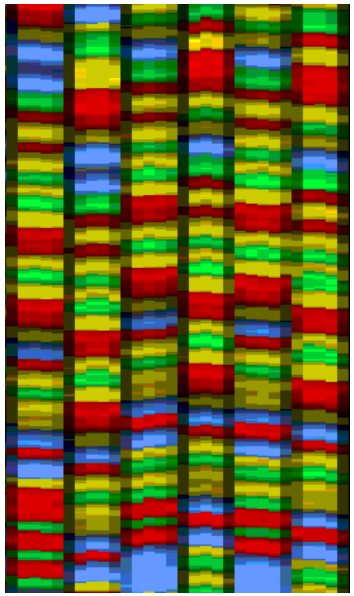




One tube.
One gel lane or capilliary

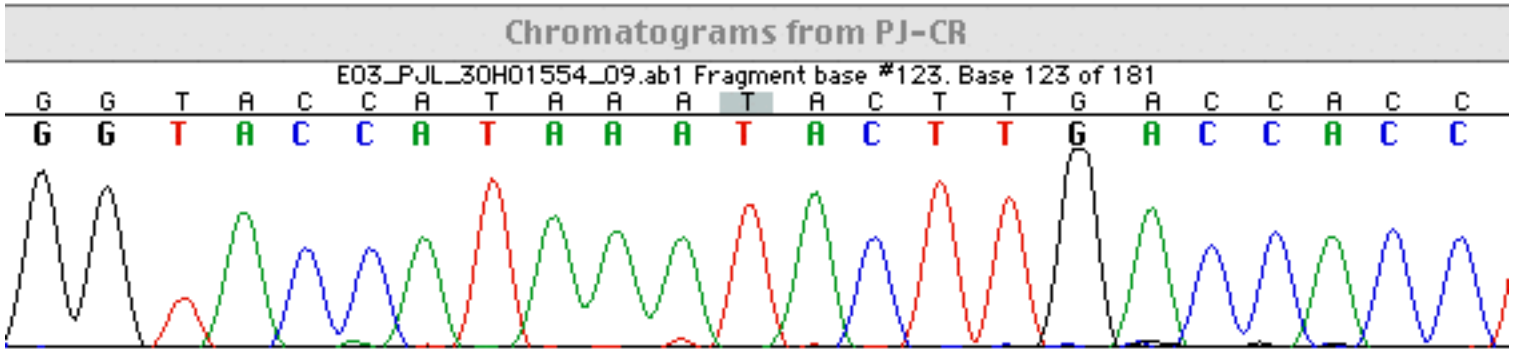

Primer for replication
Strand to be sequenced
Primed DNA
+
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
C
G
A
T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
Read seqence as complement of bands containing labeled strands

DNA sequence output from ABI 377 (a gel-based sequencer)



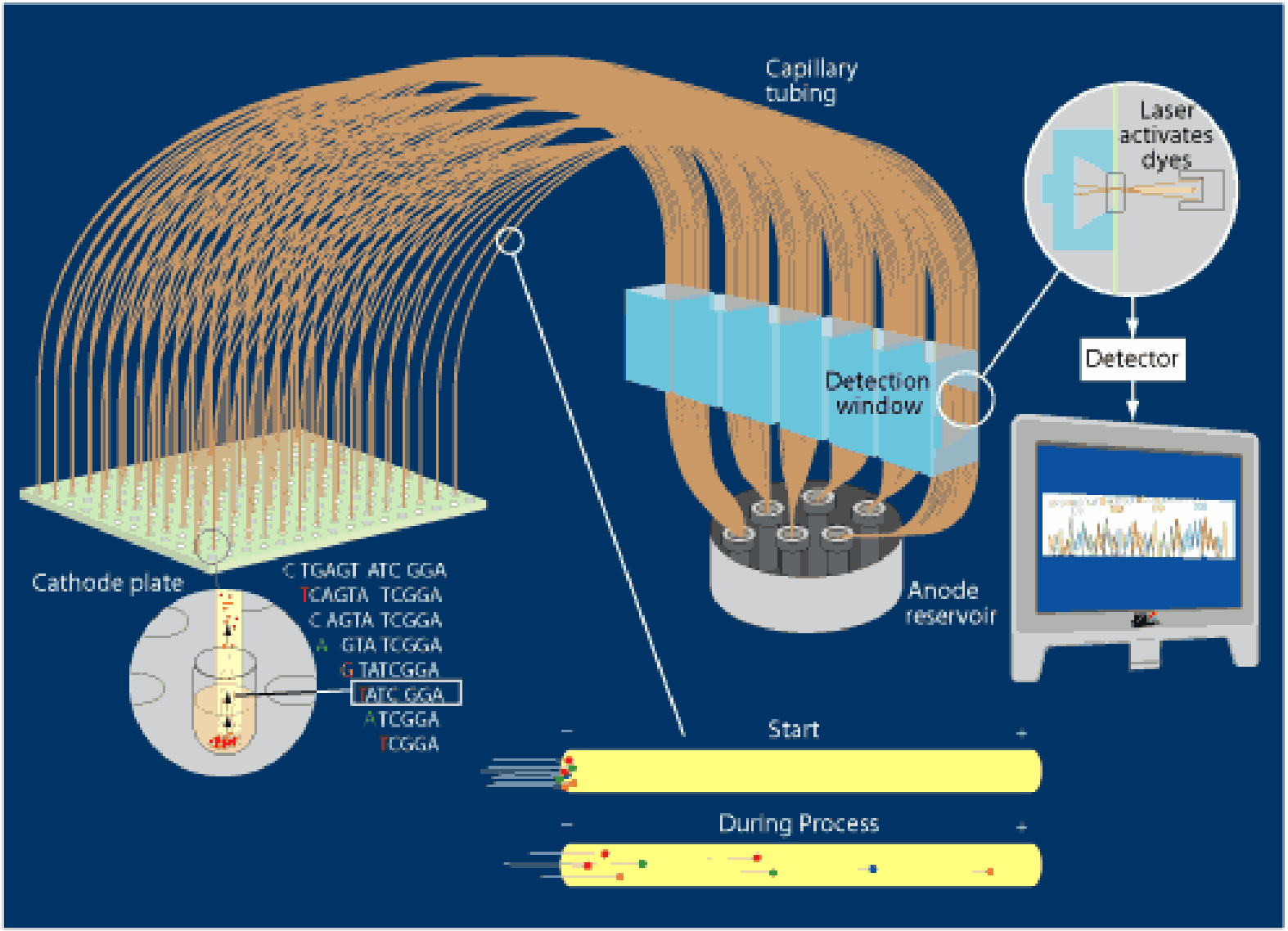
1. **Trace files (dye signals) are analyzed and bases called to create chromatograms.**

2. **Chromatograms from opposite strands are reconciled with software to create double-stranded sequence data.**



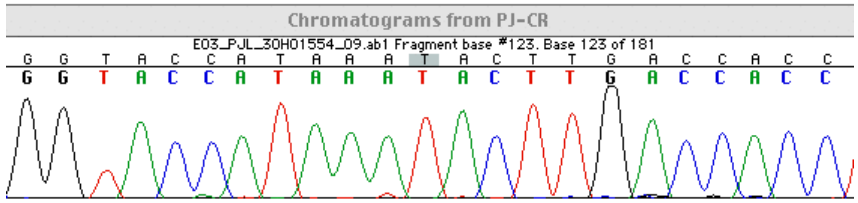
Robotic 96 capillary machine: ABI 3730 xl

Capillary sequencing – no gels



Databases

- DNA sequence and chromatogram data



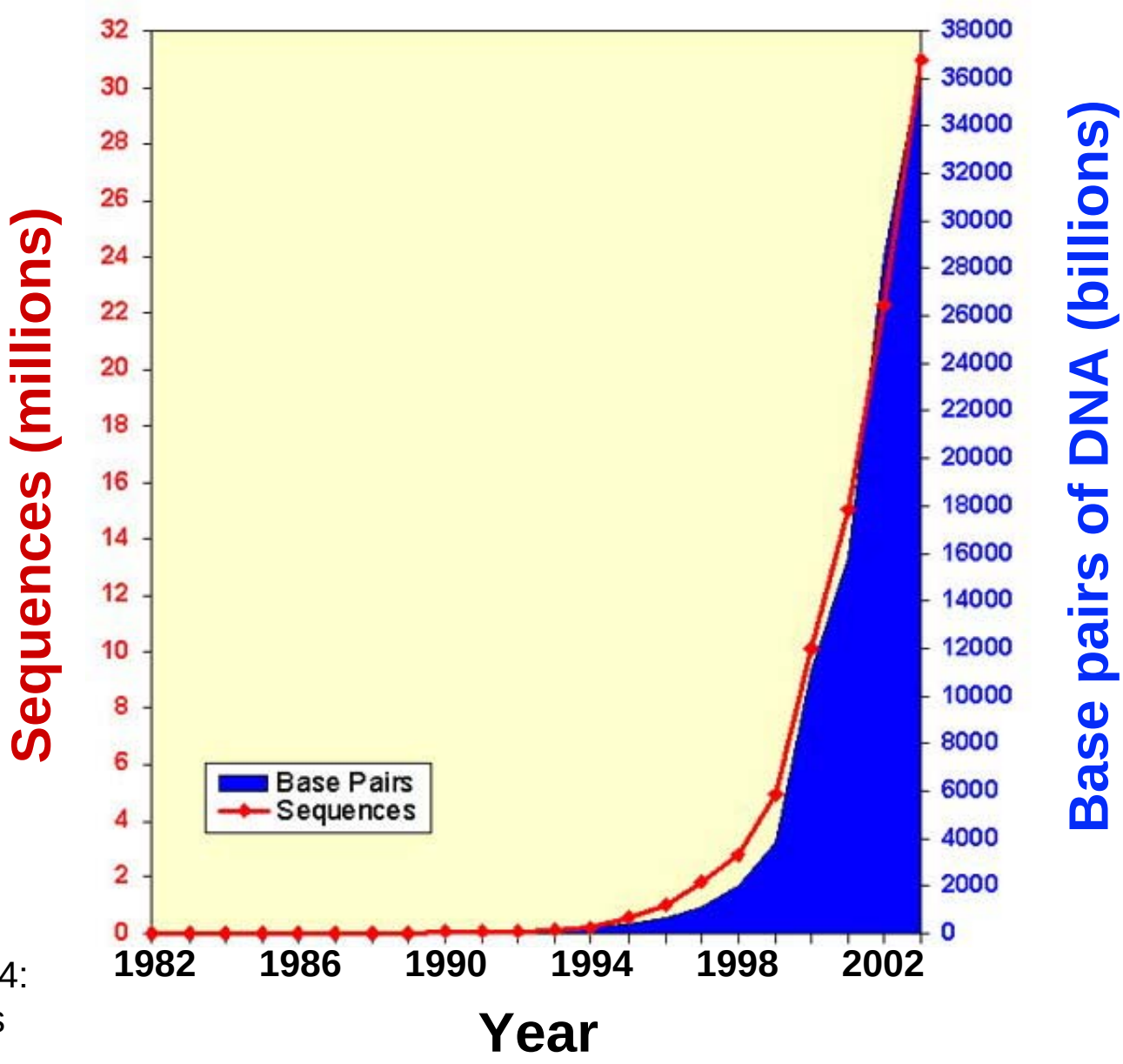
- All is deposited in “public repositories”

- The analysis of DNA sequences is the basis of the field of bioinformatics.

What is bioinformatics?

- Interface of biology and computers
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- Genomics is the analysis of genomes. The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

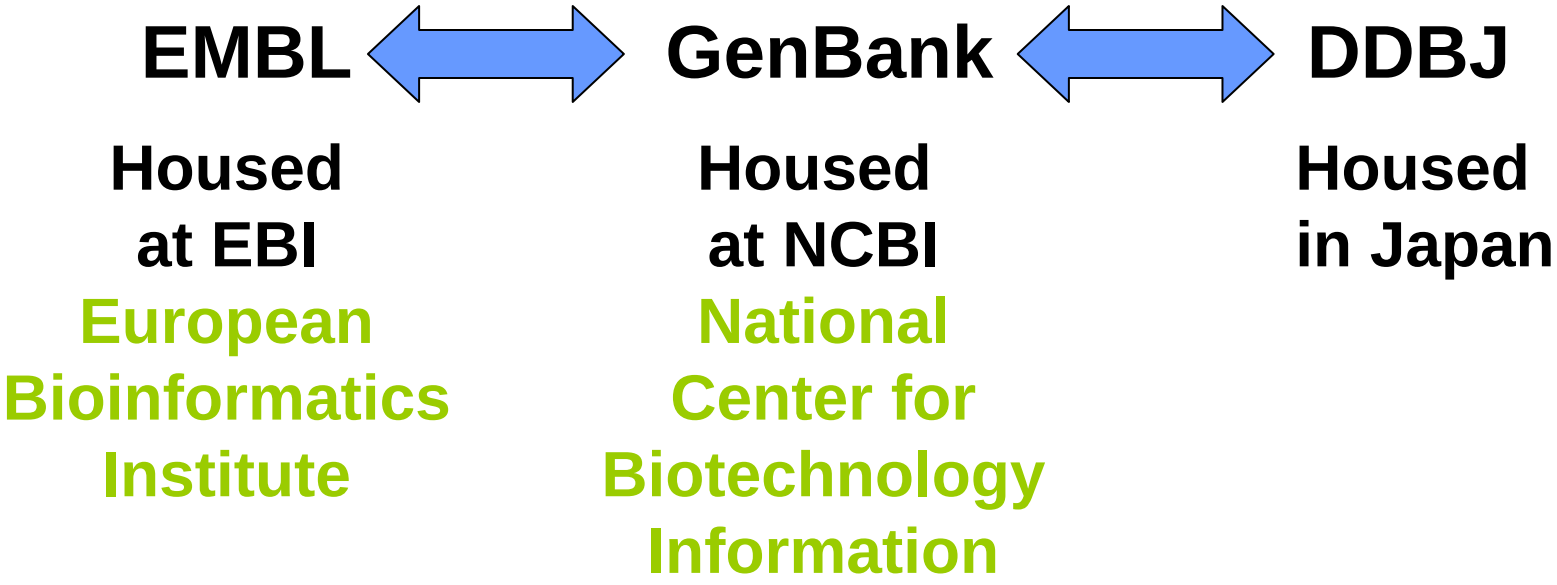
Growth of GenBank



Updated 8-12-04:
>40b base pairs

Fig. 2.1
Page 17

There are three major public DNA databases



>100,000 species are represented in GenBank

all species	128,941
viruses	6,137
bacteria	31,262
archaea	2,100
eukaryota	87,147

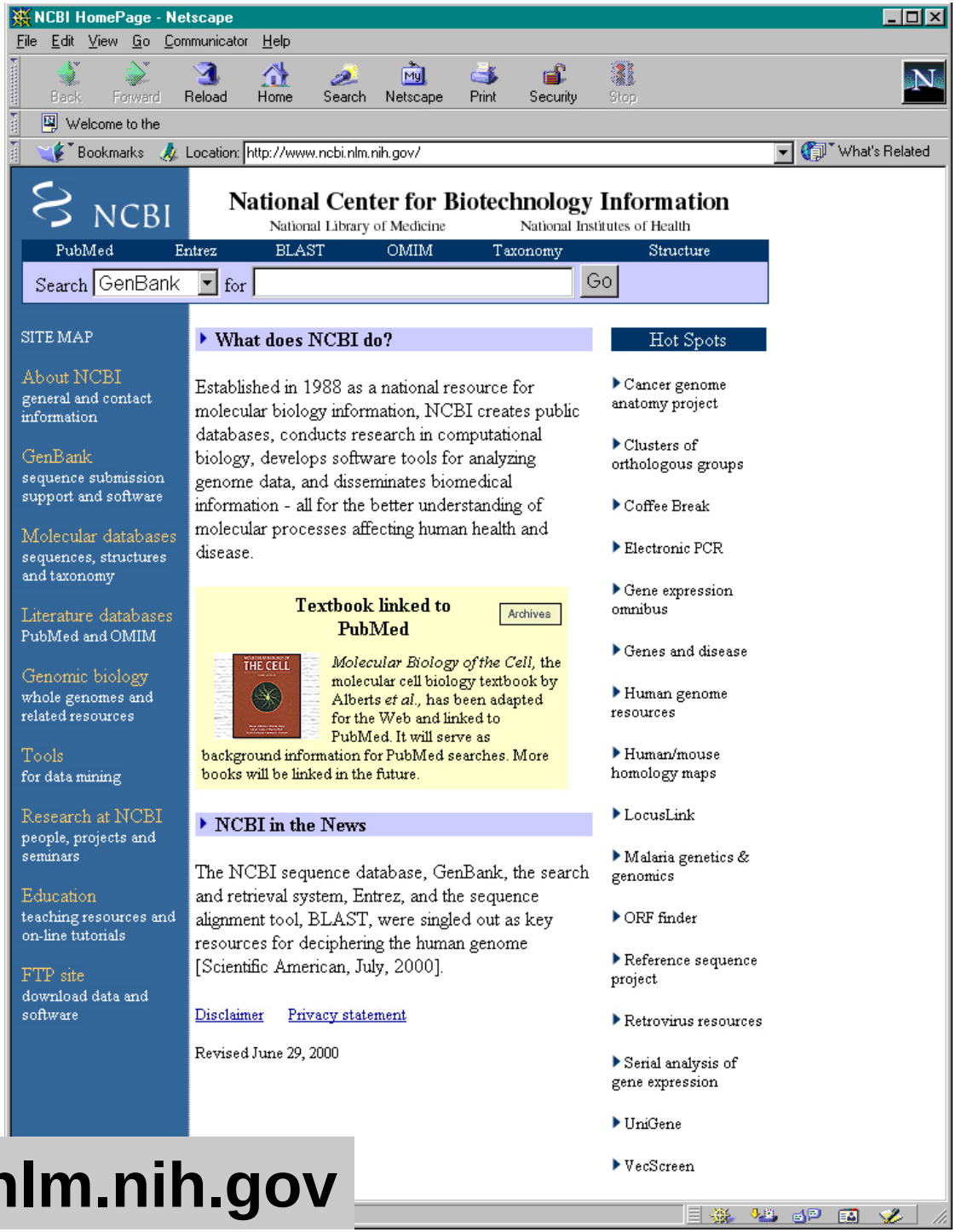
Table 2-1

The most sequenced organisms in GenBank

Homo sapiens	11.2 billion bases
Mus musculus	7.5b
Rattus norvegicus	5.7b
Danio rerio	2.1b
Bos taurus	1.9b
Zea mays	1.4b
Oryza sativa (japonica)	1.2b
Xenopus tropicalis	0.9b
Canis familiaris	0.8b
Drosophila melanogaster	0.7b

National Center for Biotechnology Information (NCBI)

www.ncbi.nlm.nih.gov



www.ncbi.nlm.nih.gov

Fig. 2.5
Page 25

The genome factory

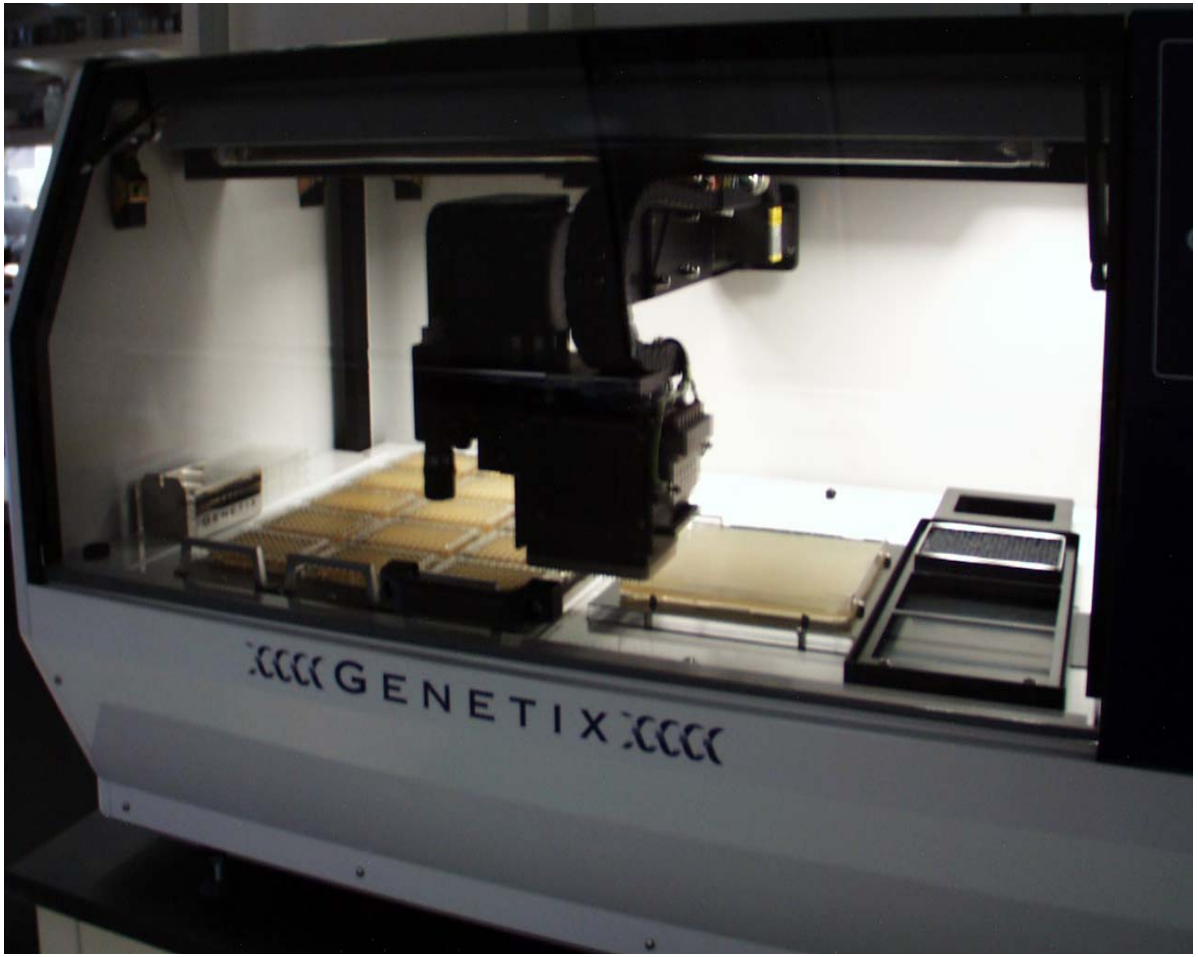
There are a few centers around the world that have a “factory” big enough to do shotgun sequence of a large eukaryotic genome:

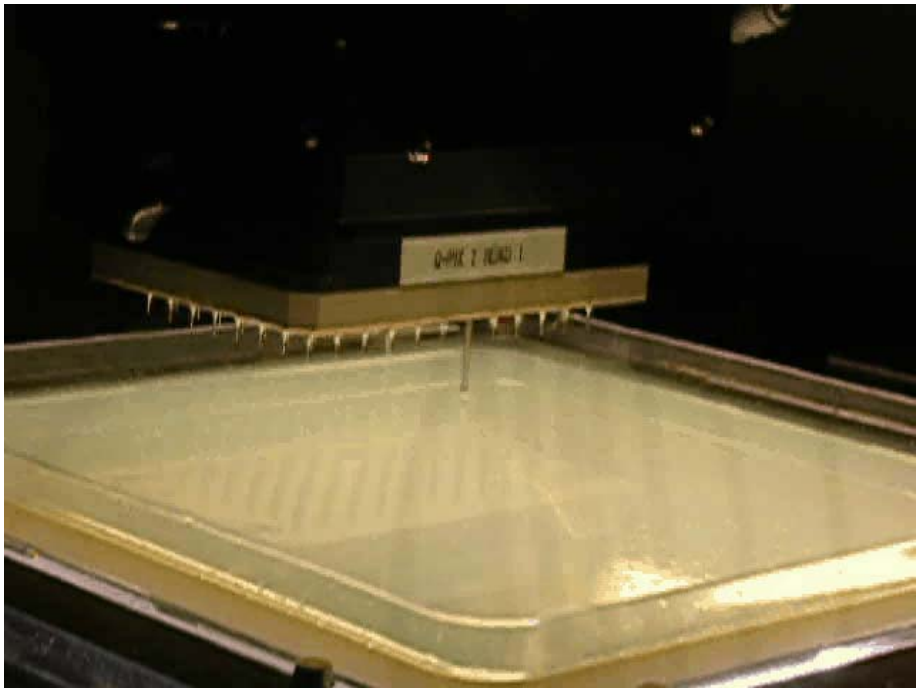
Broad Institute, MIT
Baylor College of Medicine, Houston
Washington University, St Louis
DoE Joint Genomics Institute, Walnut Creek, CA

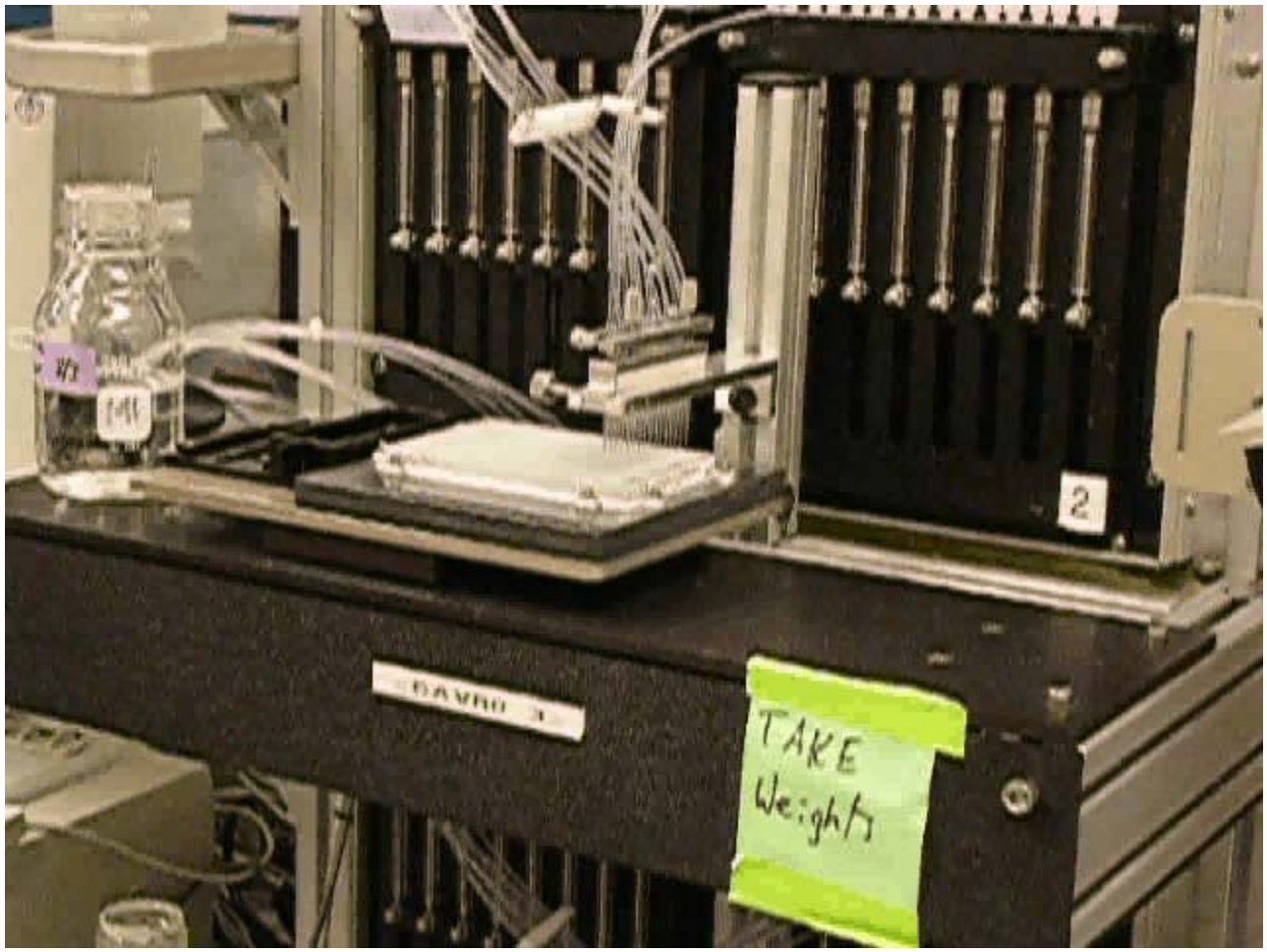
Sanger Centre, Cambridge
Beijing Genomics Institute, Chinese Academy of Sciences

Pictures from JGI

Qpix robot – picks colonies






2
TAKE
Weights

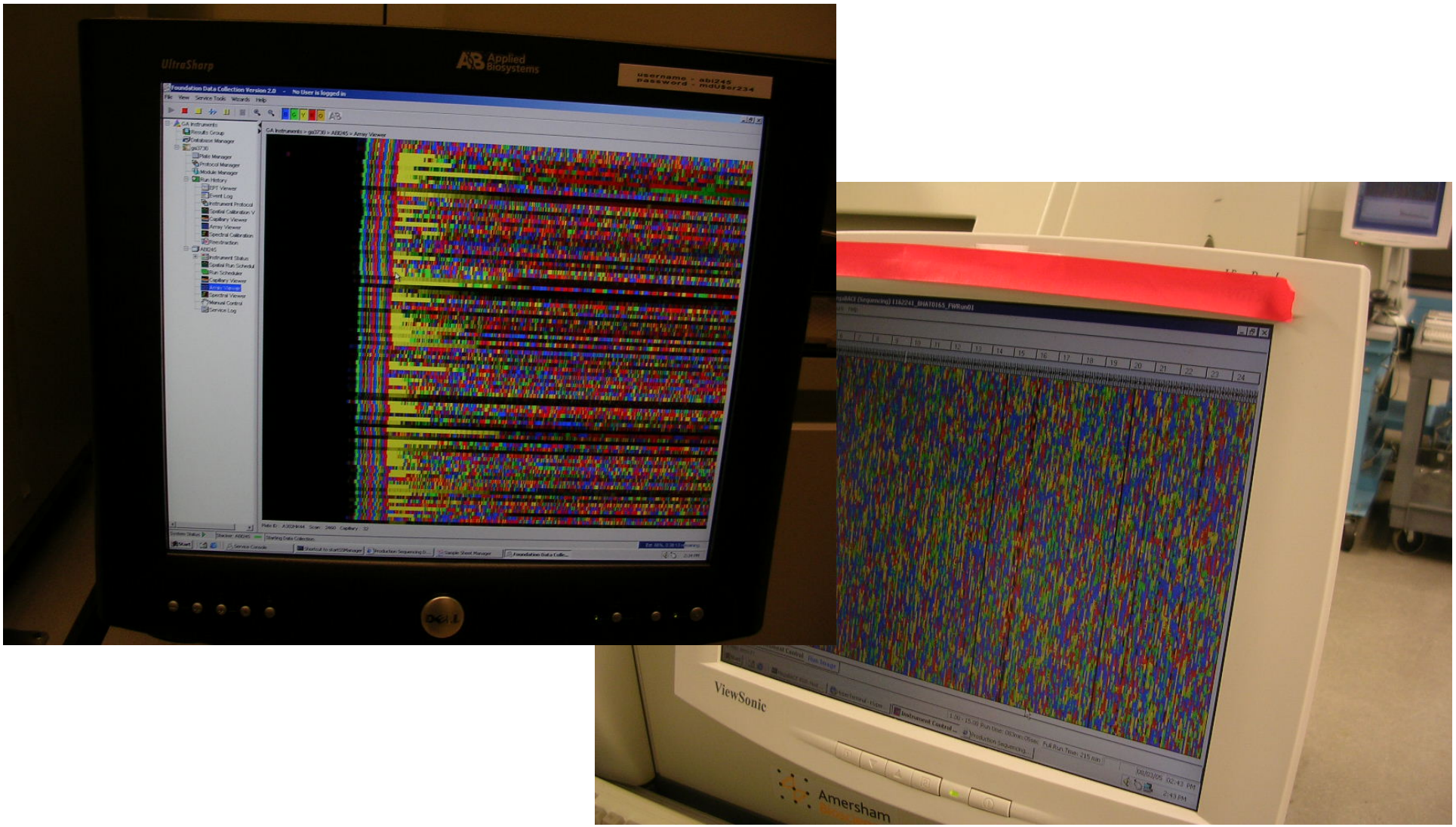
Biomek – miniprep robot


LIGHT CURTAIN WILL STOP ROBOT

PCR – 384 x 4 x 48 x 3

About 150 sequencers, at $200,000 each…

Sequence analysis



Bioinformatics

Armies of programmers and large supercomputers are necessary to assemble and annotate the sequence

The past

- Most of these genome factories are dying out and being replaced by machines that can do massively-parallel sequencing – 1.6 million well plates instead of 96 well or 384 well plates.

- The cost will go down enough to allow the differences between humans, or plants, to be detected by whole-genome **resequencing**

Current genomic technology

Roche “454” sequencer
~$500k
Equivalent to about 10,000 “old” sequencers

Illumina GA2 sequencer
~$500k
Equivalent to about 100,000 “old” sequencers

And then what?

- Fastest moving field in biology
- Keep up with it!