How to Sequence a whole genome

Ameer Effat M. Elfarash

Dept. of Genetics
Fac. of Agriculture, Assiut Univ.
aelfarash@aun.edu.eg

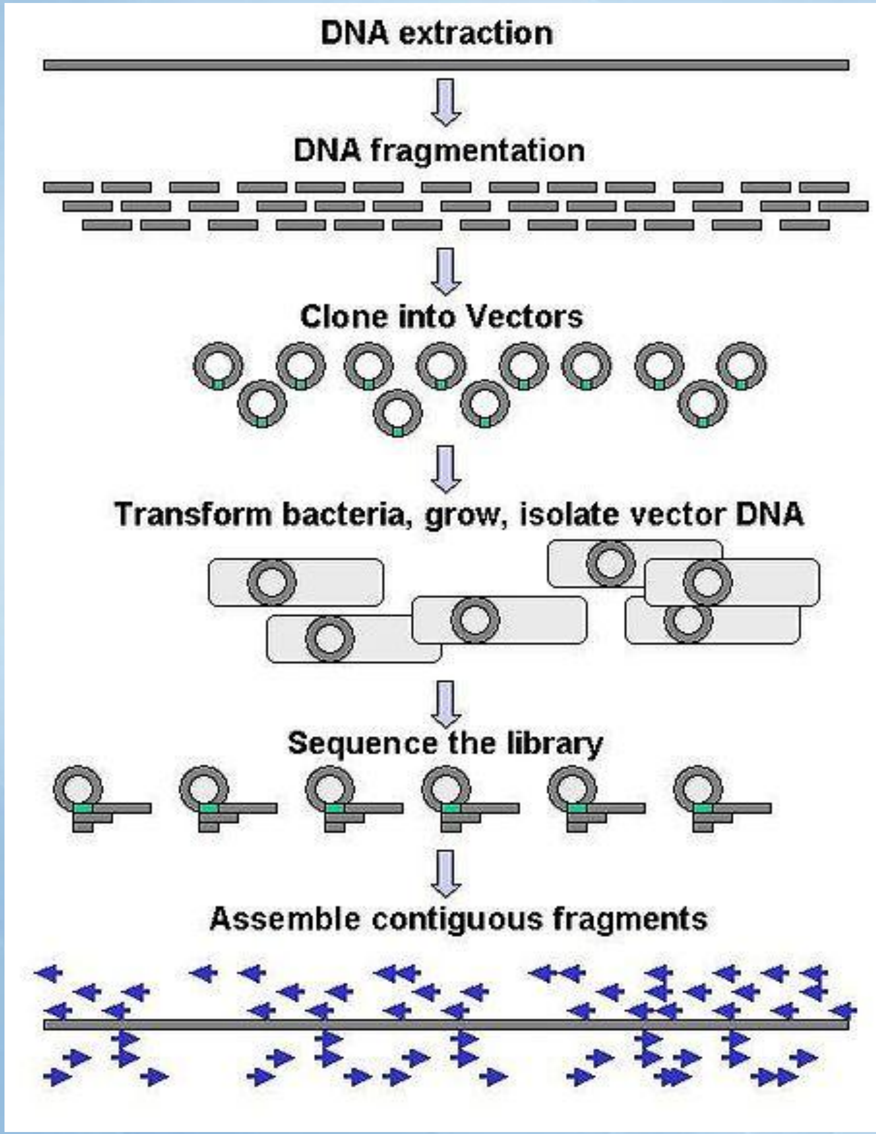
Why do we want to know the sequence of an entire genome??

To know all the genes – then proteins, then pathways…

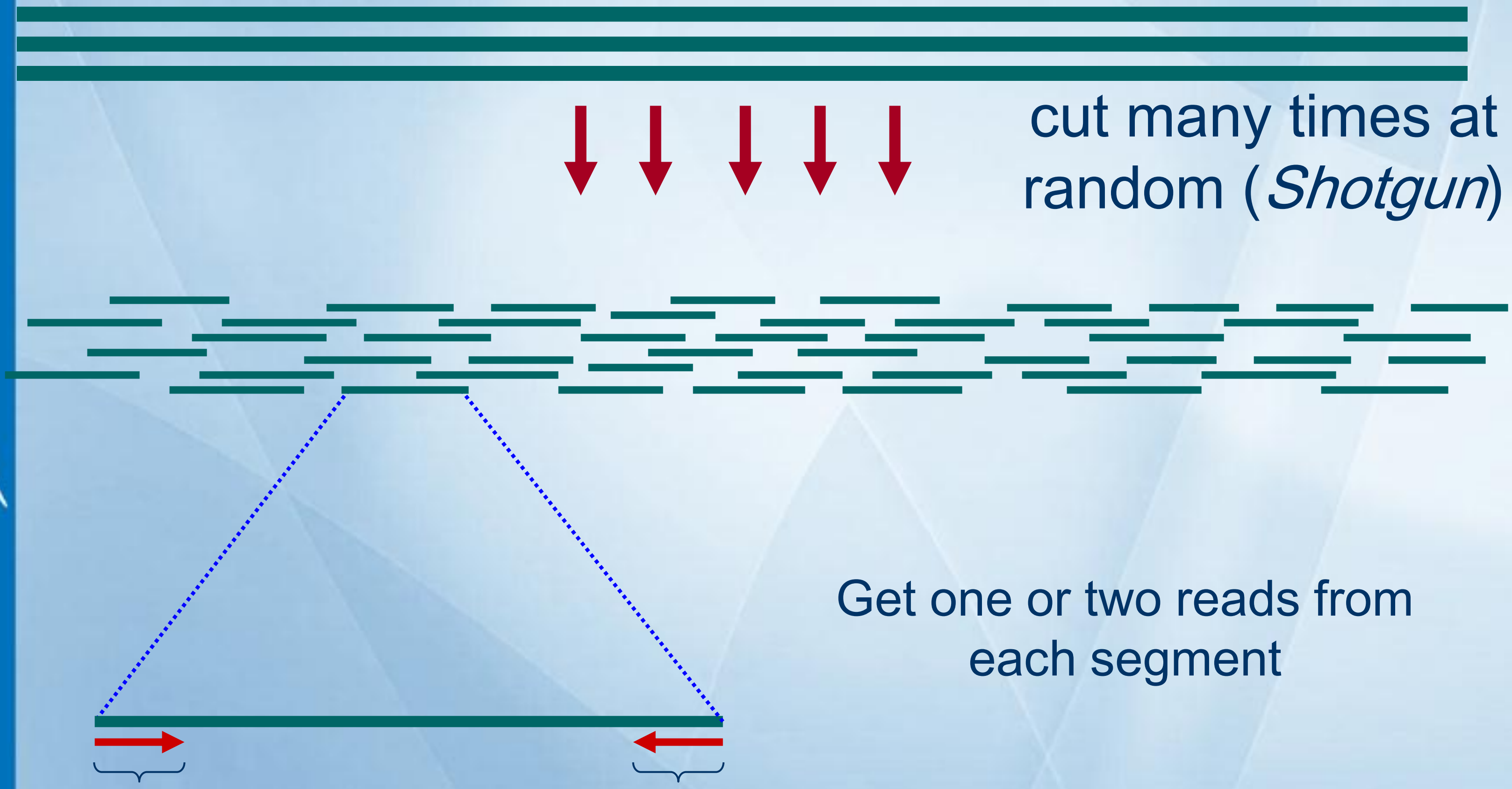
We can understand:

- the biochemistry of the organism
- diseases
- Regulation

Genome Sequencing (Shotgun)



Method to sequence longer regions



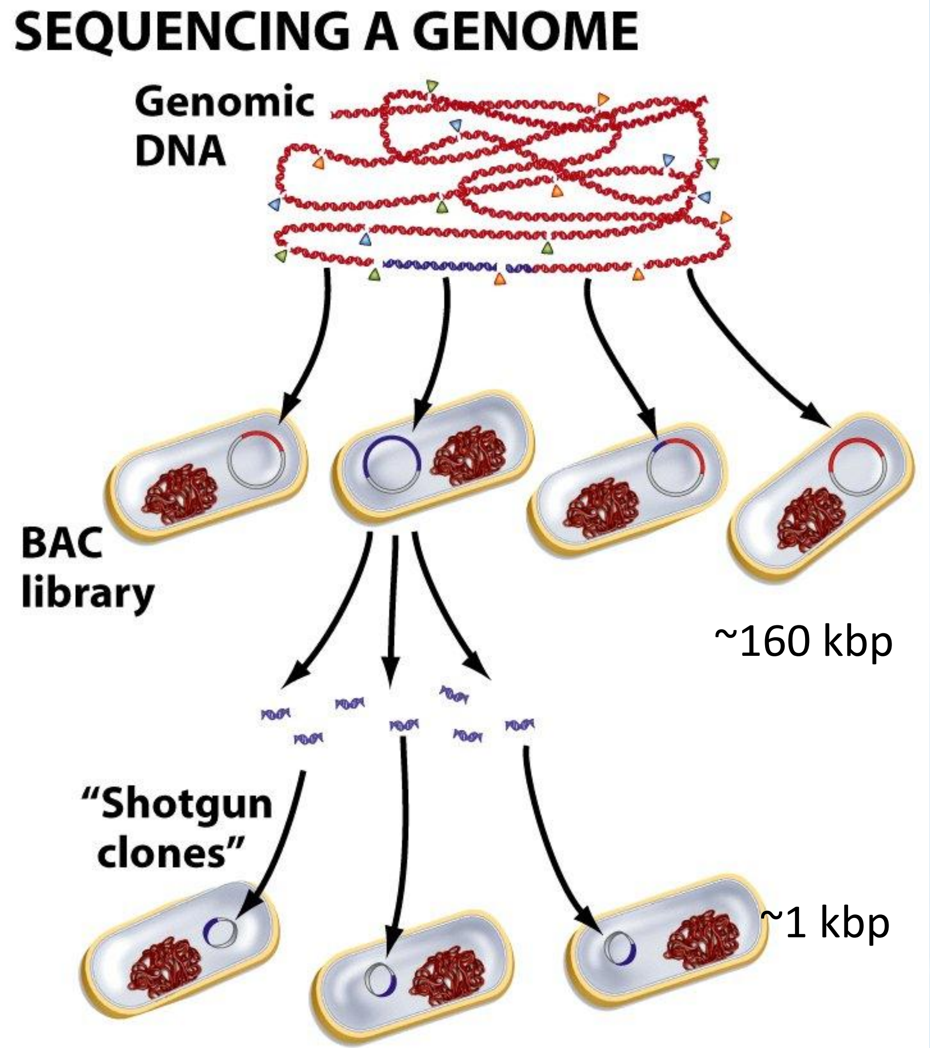
genomic segments

cut many times at random (*Shotgun*)

Get one or two reads from each segment

~500 bp

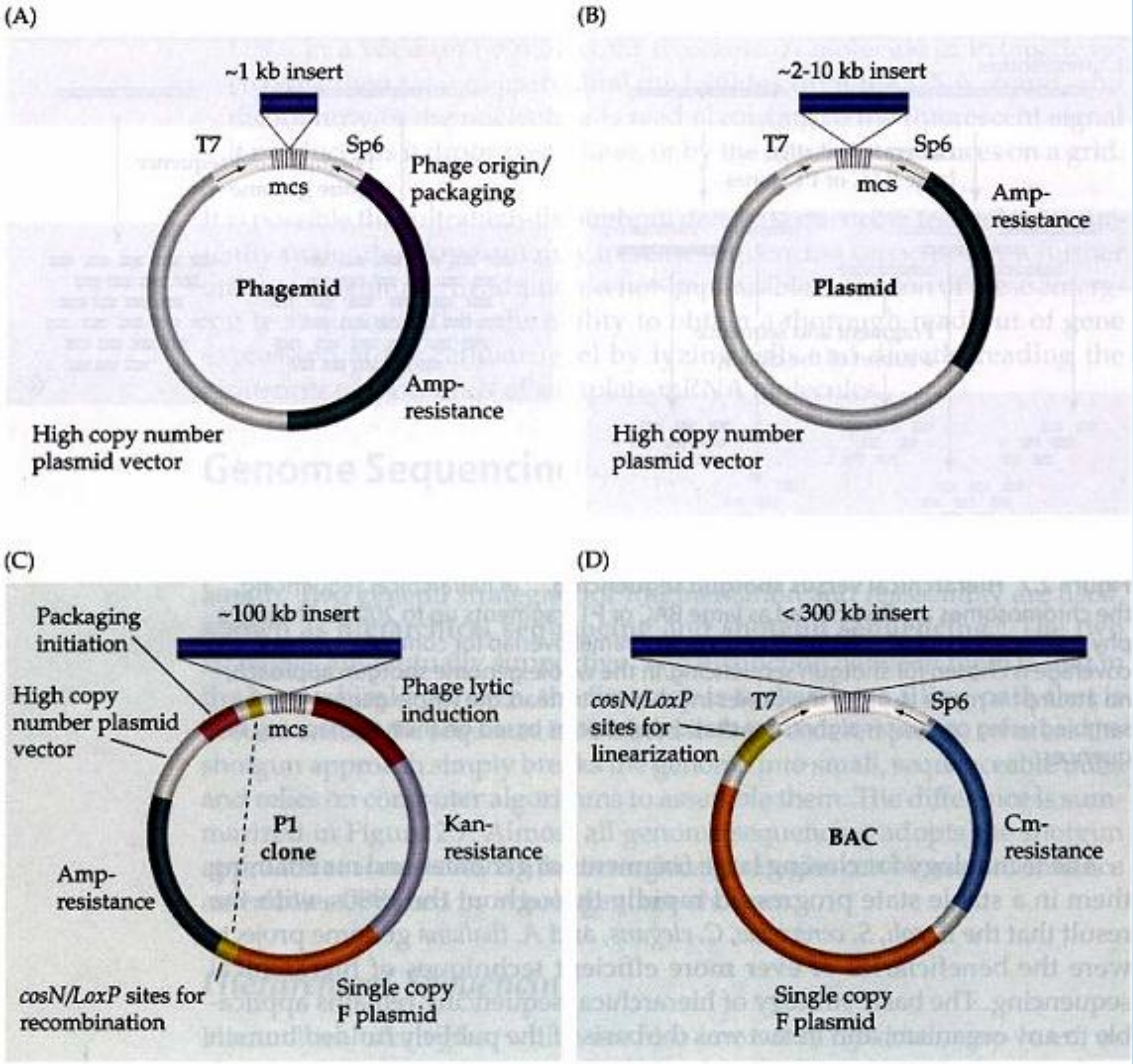
~500 bp


SEQUENCING A GENOME
Genomic DNA
BAC library
"Shotgun clones"

~160 kbp

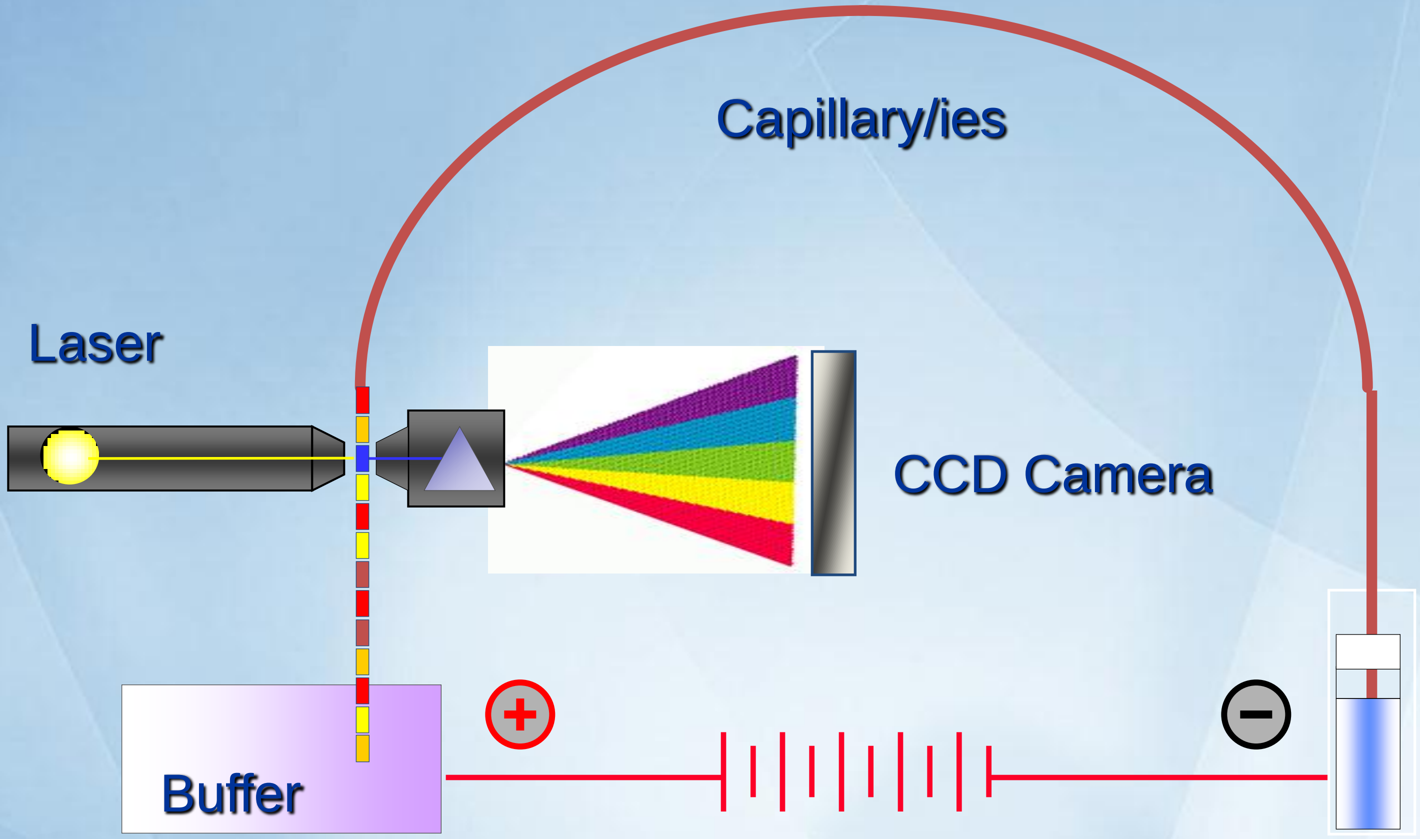
~1 kbp

Cloning vectors for genome sequencing



Sizes of inserts in sequencing vectors

Vector	Size (approx.)
P1	100 Kb
YAC	300 -1500 Kb
BAC	70 - 300 Kb
Cosmid	~ 40 Kb
Plasmid	2 -10 Kb
M13 or Phagmid	~ 1 Kb

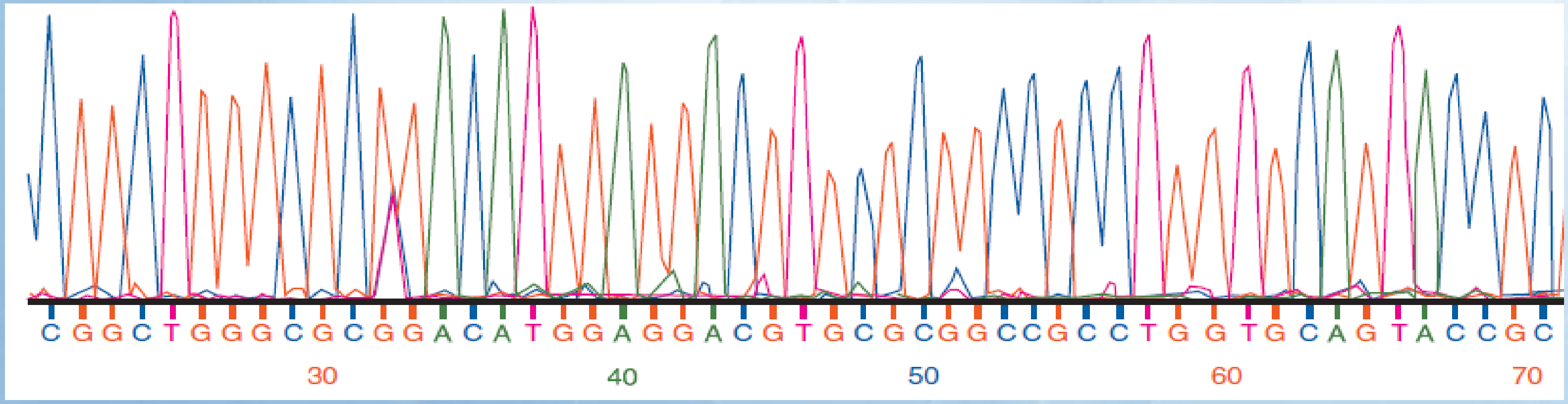


Capillary/ies

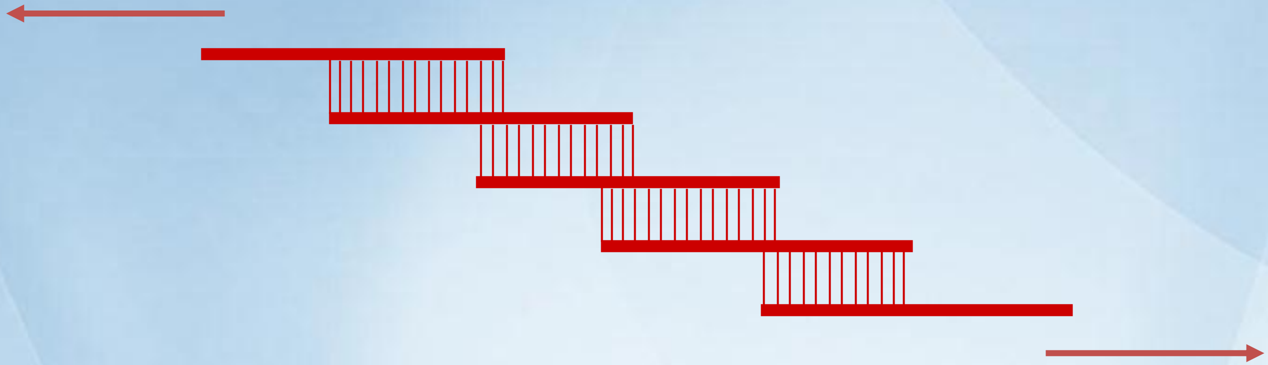
Laser

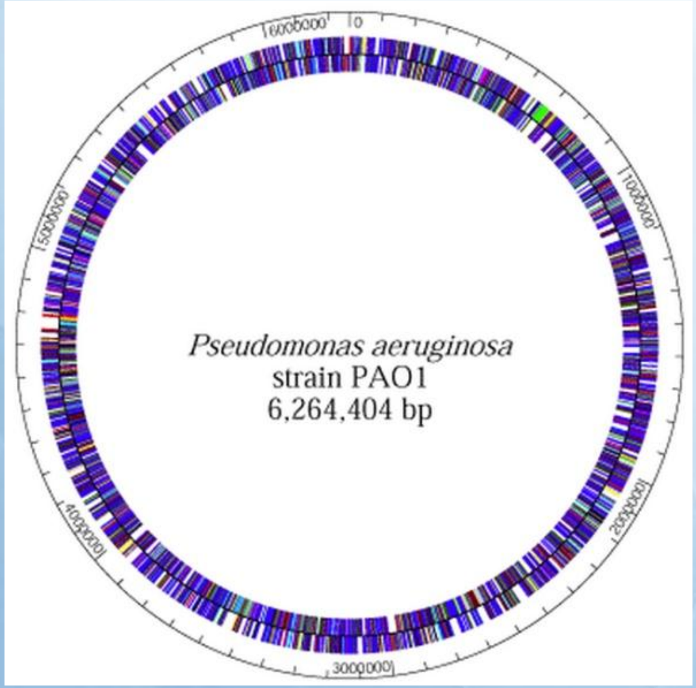
CCD Camera

Buffer



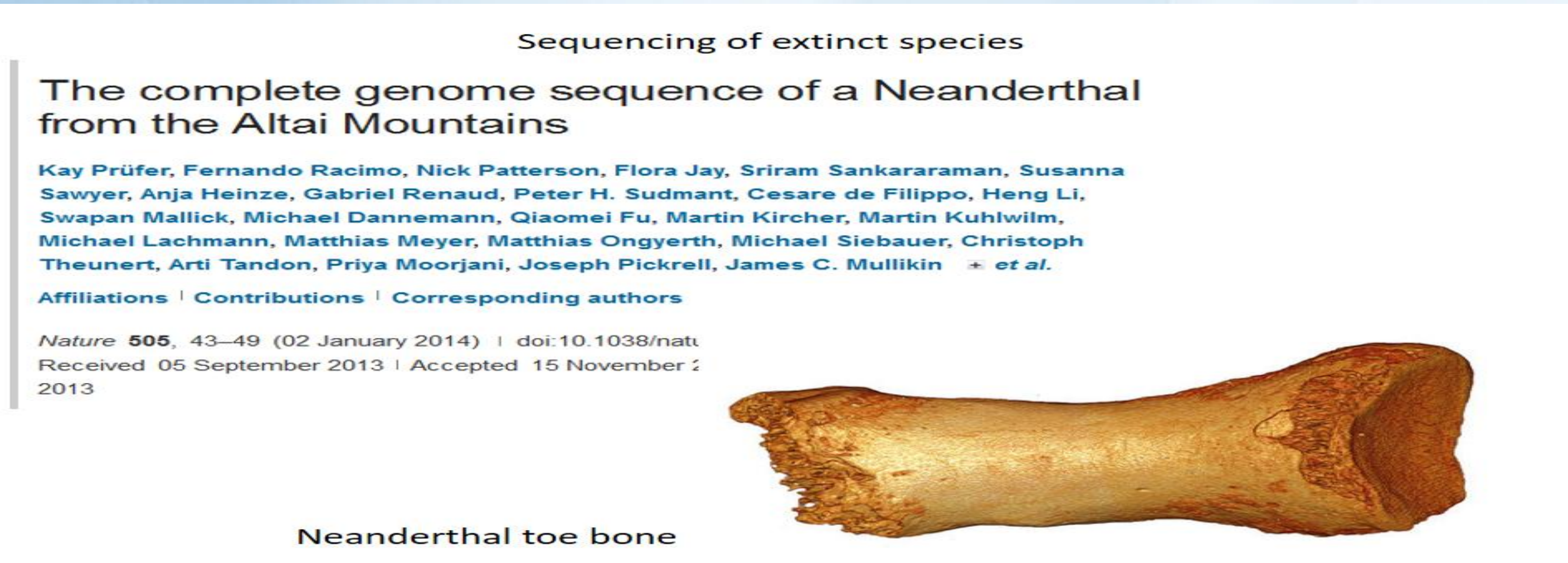
Reconstructing the Sequence (Fragment Assembly)





Genomes to Date

- **69 higher animals + other model animals**
- **55 insects and lower metazoans**
- **39 plants**
- **563 fungi**
- **Over 200 protist species and subspecies**
- **Over 20 000 bacteria species and subspecies**
  - Microbial communities in oceans, desserts, hot springs, inside bodies



Sequencing of extinct species

The complete genome sequence of a Neanderthal from the Altai Mountains

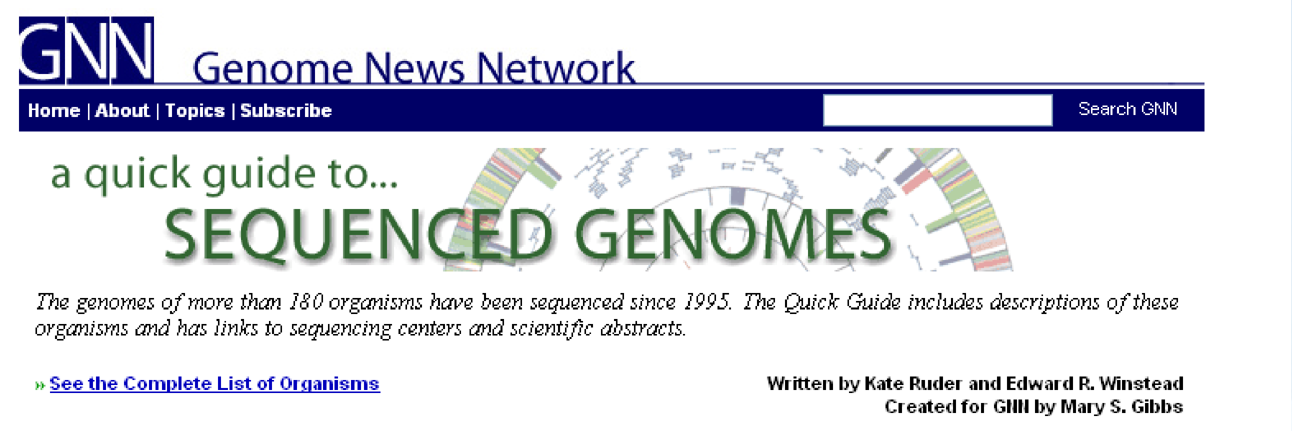
Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **505**, 43–49 (02 January 2014) | doi:10.1038/natu
Received 05 September 2013 | Accepted 15 November
2013

Neanderthal toe bone

Where can I find genome sequences?



Genome News Network

a quick guide to...

SEQUENCED GENOMES

The genomes of more than 180 organisms have been sequenced since 1995. The Quick Guide includes descriptions of these organisms and has links to sequencing centers and scientific abstracts.

» See the Complete List of Organisms

Written by Kate Ruder and Edward R. Winstead
Created for GNN by Mary S. Gibbs

http://www.genomenewsnetwork.org/

Websites „genome browsers“ (include annotations of genes)
Ensembl genome browser
UCSC genome browser
NCBI genome browser

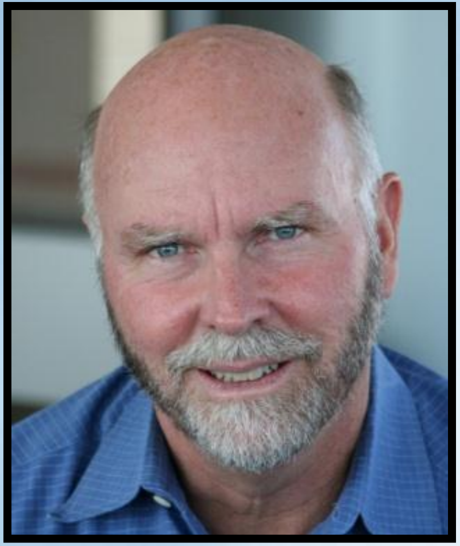
The Human Genome Project

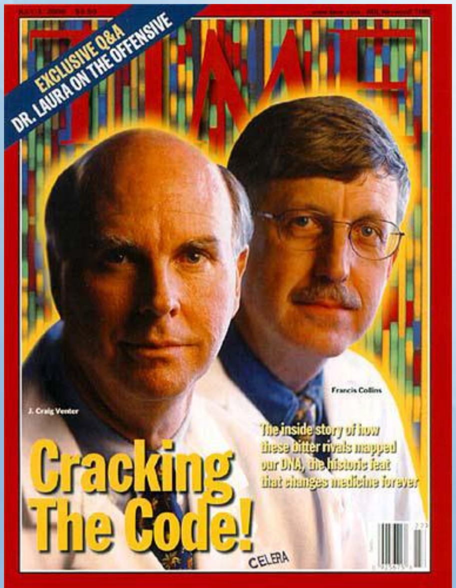
- The Mission of the HGP: To understand the human genome and the role it plays in both health and disease.
- Initiated 1990 -Completion originally planned for 2005 (expected to take 15 years)
- $3-billion

movie

Finished sequence anticipated Spring, 2003, after 50 years of Watson and Crick publication (Nature 171: 737-738, April 25, 1953)

VENTER'S SHOTGUN
Human genome sequencing



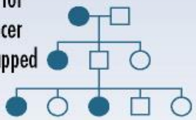


1990

Human Genome Project (HGP) launched in the U.S.

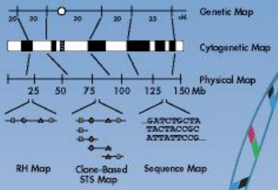
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE

First gene for breast cancer (BRCA1) mapped



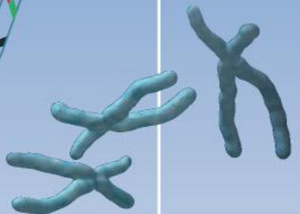
1991

First U.S. Genome Centers established



1992

Second-generation human genetic map developed



Rapid data release guidelines established by NIH and DOE

1993

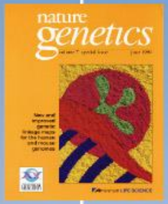
New five-year plan for the HGP in the U.S. published

Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)

The Wellcome Trust

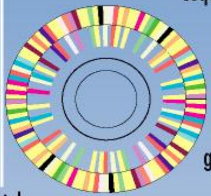
1994

HGP's human genetic mapping goal achieved



1995

HGP's human physical mapping goal achieved



First bacterial genome (*H. influenzae*) sequenced

U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996

First human gene map established

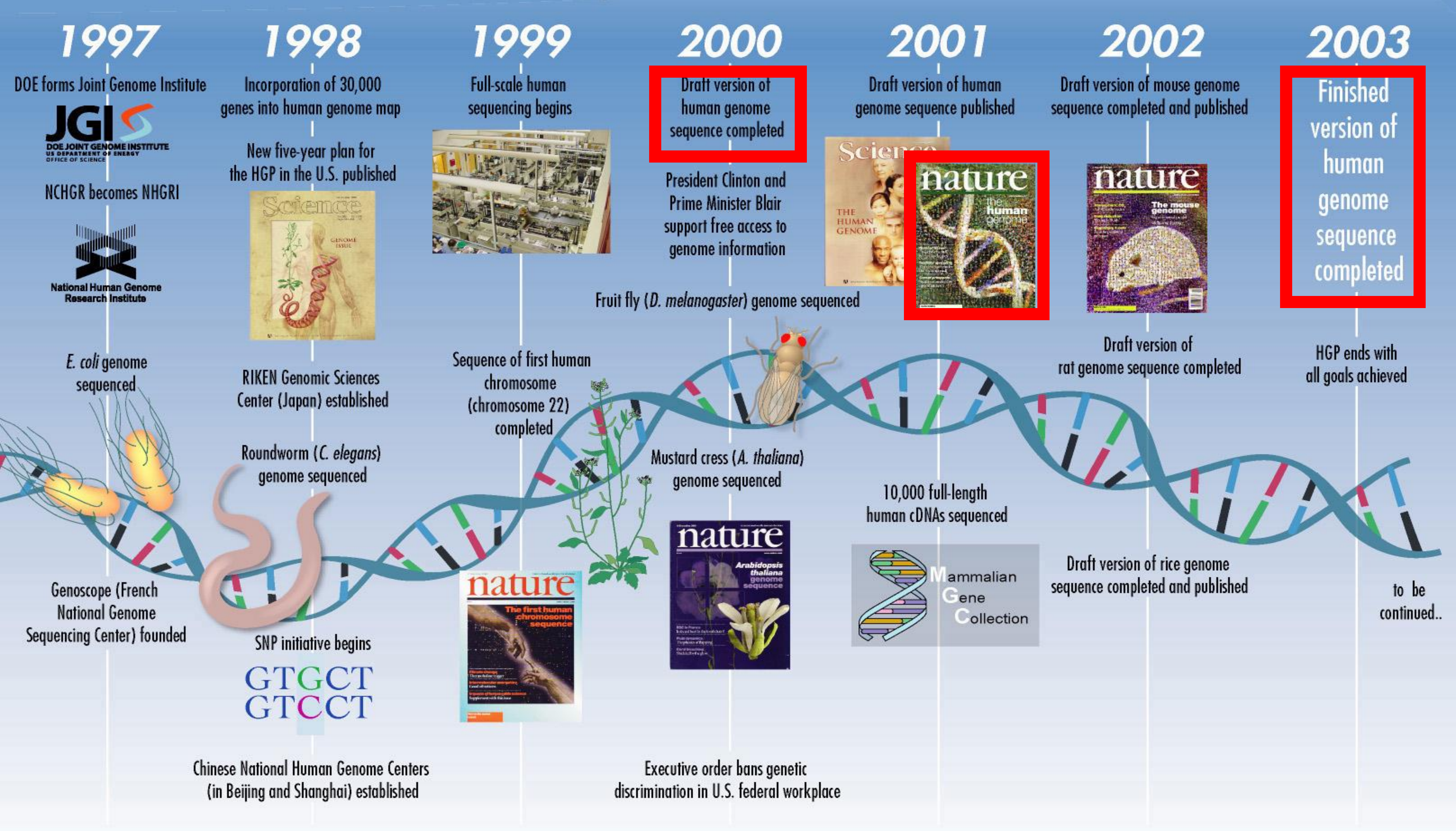
Pilot projects for human genome sequencing begin in U.S.

First archael genome sequenced

Yeast (*S. cerevisiae*) genome sequenced

HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established



1997

DOE forms Joint Genome Institute

NCHGR becomes NHGRI

E. coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded

1998

Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the U.S. published

RIKEN Genomic Sciences Center (Japan) established

Roundworm (*C. elegans*) genome sequenced

SNP initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers (in Beijing and Shanghai) established

1999

Full-scale human sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000

Draft version of human genome sequence completed

President Clinton and Prime Minister Blair support free access to genome information

Fruit fly (*D. melanogaster*) genome sequenced

Mustard cress (*A. thaliana*) genome sequenced

Executive order bans genetic discrimination in U.S. federal workplace

2001

Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002

Draft version of mouse genome sequence completed and published

Draft version of rat genome sequence completed

Draft version of rice genome sequence completed and published

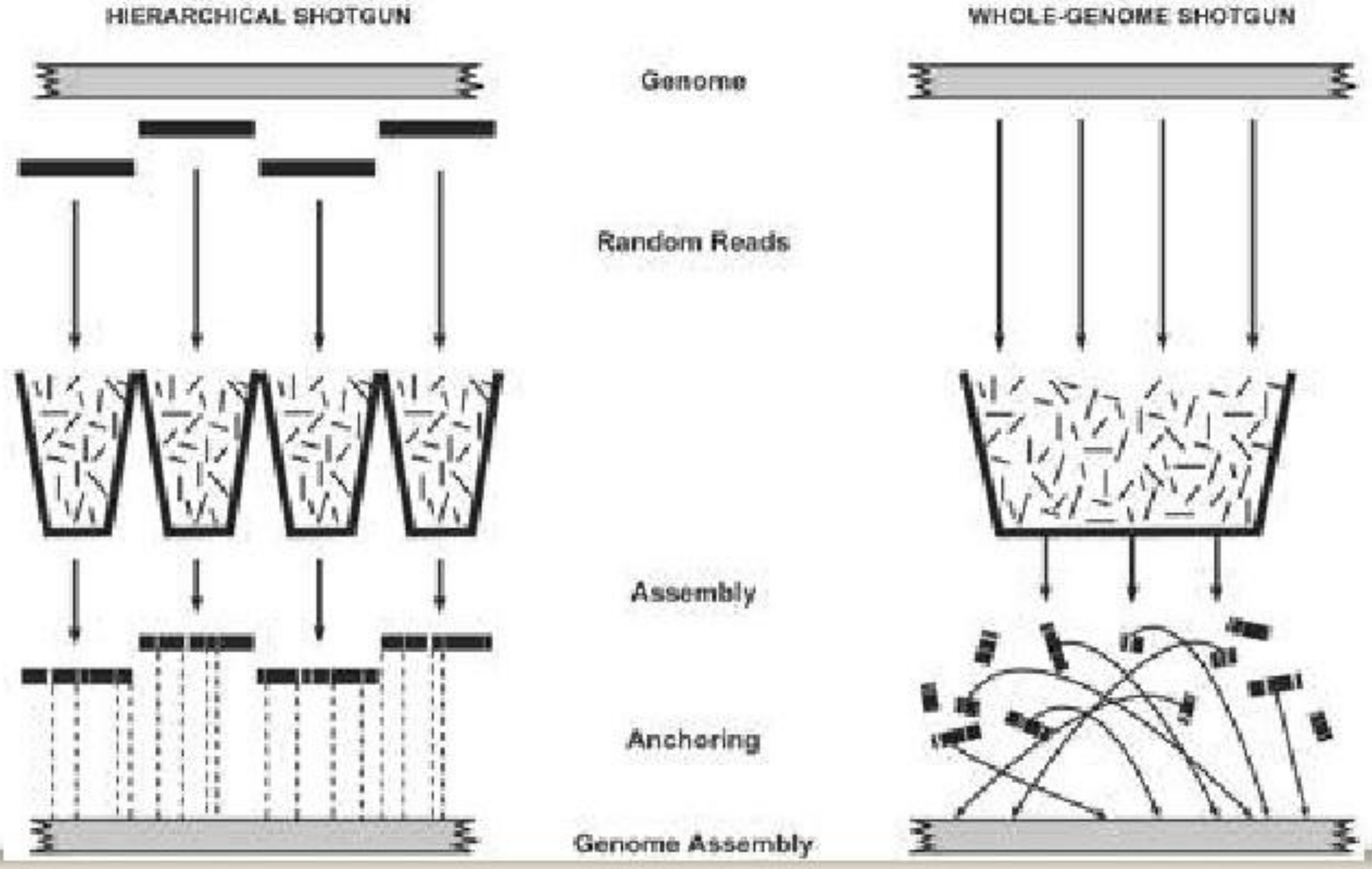
2003

Finished version of human genome sequence completed

HGP ends with all goals achieved

to be continued..

Hierarchical *vs.* Whole Genome



"All the News
That's Fit to Print"

The New York Times

National Edition

Southern California: Mostly sunny with light winds. Highs ranging from the 70's along the beaches to over 100 in the deserts. Tonight, mainly clear, low 65-70. Weather map, Page A24.

VOL. CXLIX . . . . No. 51,432 TUESDAY, JUNE 27, 2000 Printed in California ONE DOLLAR

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part of our national culture."

The court said in an opinion by Chief Justice William H. Rehnquist that because the 1966 Miranda decision "announced a constitutional rule," a statute by which Congress had sought to overrule the decision was itself unconstitutional.

Miranda had appeared to be in jeopardy, both because of that long-ignored but recently rediscovered law, by which Congress had tried to overrule Miranda 32 years ago, and because of the court's perceived hostility to the original decision.

The chief justice said, though, that the 1968 law, which replaced the Miranda warnings with a case-by-case test of whether a confession was voluntary, could be upheld only if the Supreme Court decided to overturn Miranda. But with Miranda having "become embedded in routine police practice" without causing any measurable difficulty for prosecutors, there was no justification for doing so, he said. [Excerpts, Page A18.]

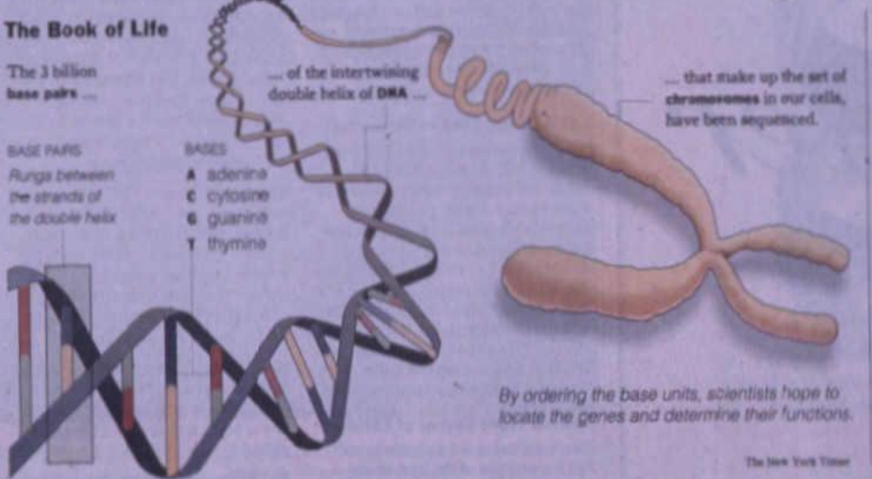
Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption that a confession was not voluntary unless preceded by the warnings was not required by the Constitution.

The decision today — only 14 pages long, in Chief Justice Rehnquist's typically spare style — brought an abrupt end to one of the odder episodes in the court's recent history, an intense and strangely delayed re-fighting of a previous generation's battle over the rights of criminal suspects. Miranda v. Arizona was a hallmark of the Warren Court, and Chief Justice Rehnquist, despite his record as an early and tenacious critic of the decision, evidently did not want its repudiation to be an imprint of his own tenure.

There was considerable drama in the courtroom today as the chief justice announced that he would deliver the decision in the case, Dickerson v. United States, No. 99-5525. The announcement meant that he was the majority opinion's author. Given his statements over more than 25 years about Miranda's lack of constitutional foundation, there was the

The Book of Life



By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

Science Times

A special issue

- Putting the genome to work.
- Some information has already paid research dividends.
- Two research methods, two results
- More articles, charts and photos of the genome effort.
- From Mendel to helix to genome.

Section D

Francis S. Collins, head of the Human Genome Project, right, with J. Craig Venter, head of Celera Genomics, after the announcement yesterday that they had finished the first survey of the human genome.

Paul Hosefros/The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

"Today we are learning the language in which God created life," President Clinton said at a White House ceremony attended by members of the two teams and, via satellite, Prime Minister Tony Blair of England. [Excerpts, Page D8.]

The teams' leaders, Dr. J. Craig Venter, president of Celera Genomics, and Dr. Francis S. Collins, director of the National Human Genome Research Institute, praised each other's contributions and signaled a spirit of cooperation from now on, even though the two efforts will remain firmly independent.

The human genome, the ancient script that has now been deciphered, consists of two sets of 23 giant DNA molecules, or chromosomes, with each set — one inherited from each parent — containing more than three billion chemical units.

The successful deciphering of this vast genetic archive attests to the extraordinary pace of biology's advance since 1953, when the structure of DNA was first discovered and presages an era of even brisker

A Pearl and a Hodgepodge: Human DNA

By NATALIE ANGIER

Collins, director of the National Human Genome Research Institute. "We only have to do this once, read-

Though scientists underscore the importance of their accomplishment by calling the genome a "portrait of

15 February 2001

nature

$10.00

www.nature.com

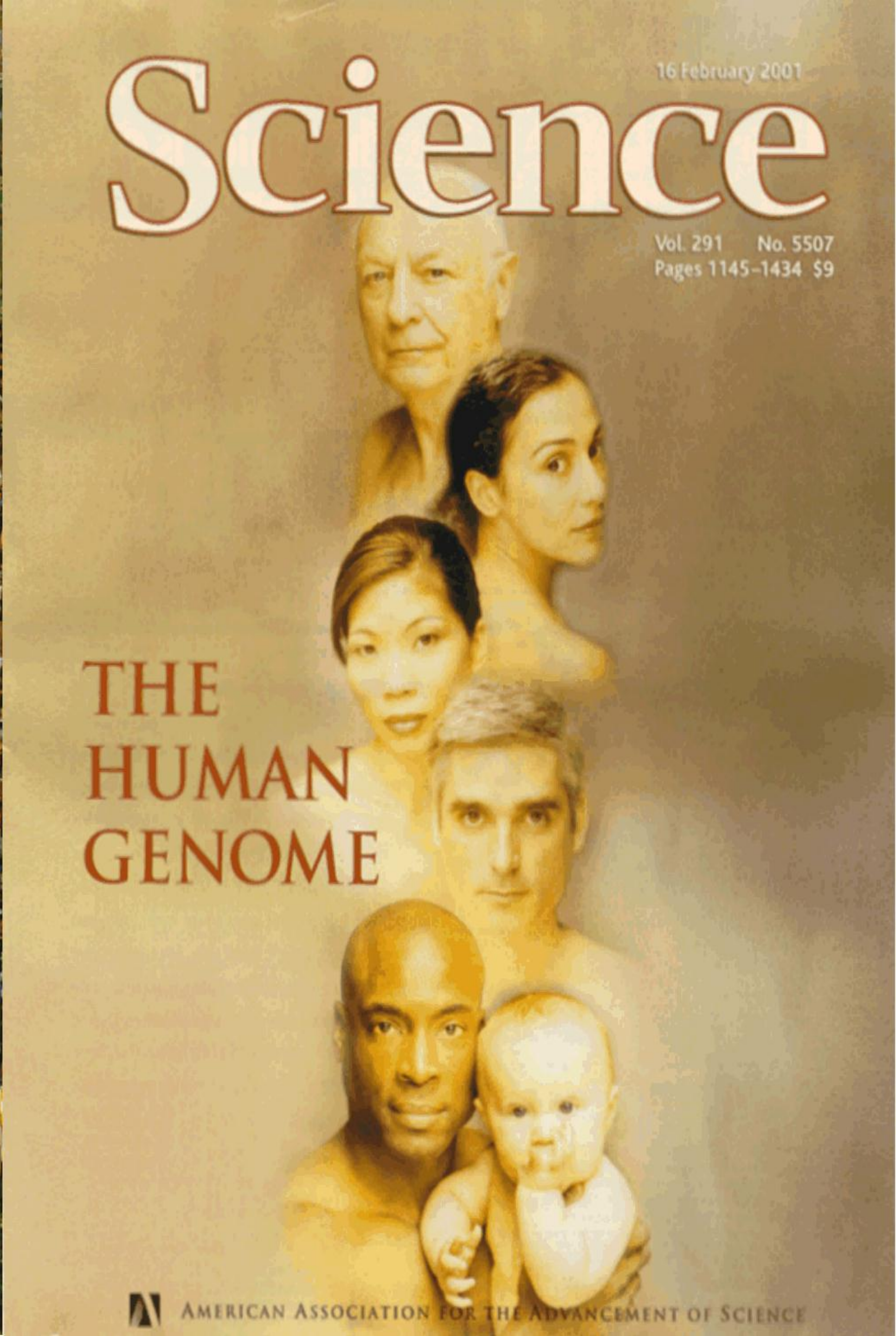
the
human
genome

Nuclear fission
Five-dimensional energy landscapes

Seafloor spreading
The view from under the Arctic ice

Career prospects
Sequence creates new opportunities

naturejobs
genomics special



16 February 2001

Science

Vol. 291 No. 5507
Pages 1145–1434 $9

THE
HUMAN
GENOME

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Human genome content

- Total length 3 billion bp ~ 30,000 genes (coding seq) functions of more than half of them are unknown

- 30,000 genes but > 500,000 known proteins (possibly exceed 2 million)

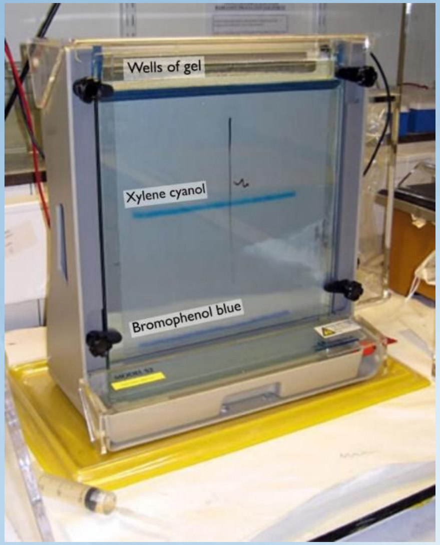
- Gene sequences < 5%
    - Exons ~ 1.5% (coding)
    - Introns ~ 3.5% (noncoding)

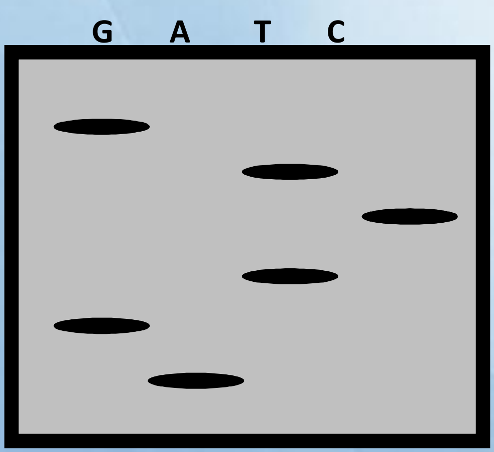
- Intergenic regions (junk) > 95%

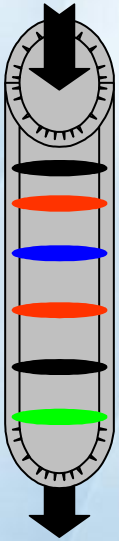
- The human genome is nearly the same (99.9%) in all people

- Almost half of all human proteins share similarities with other organisms

DNA sequencing







Next Generation Sequencing (NGS)

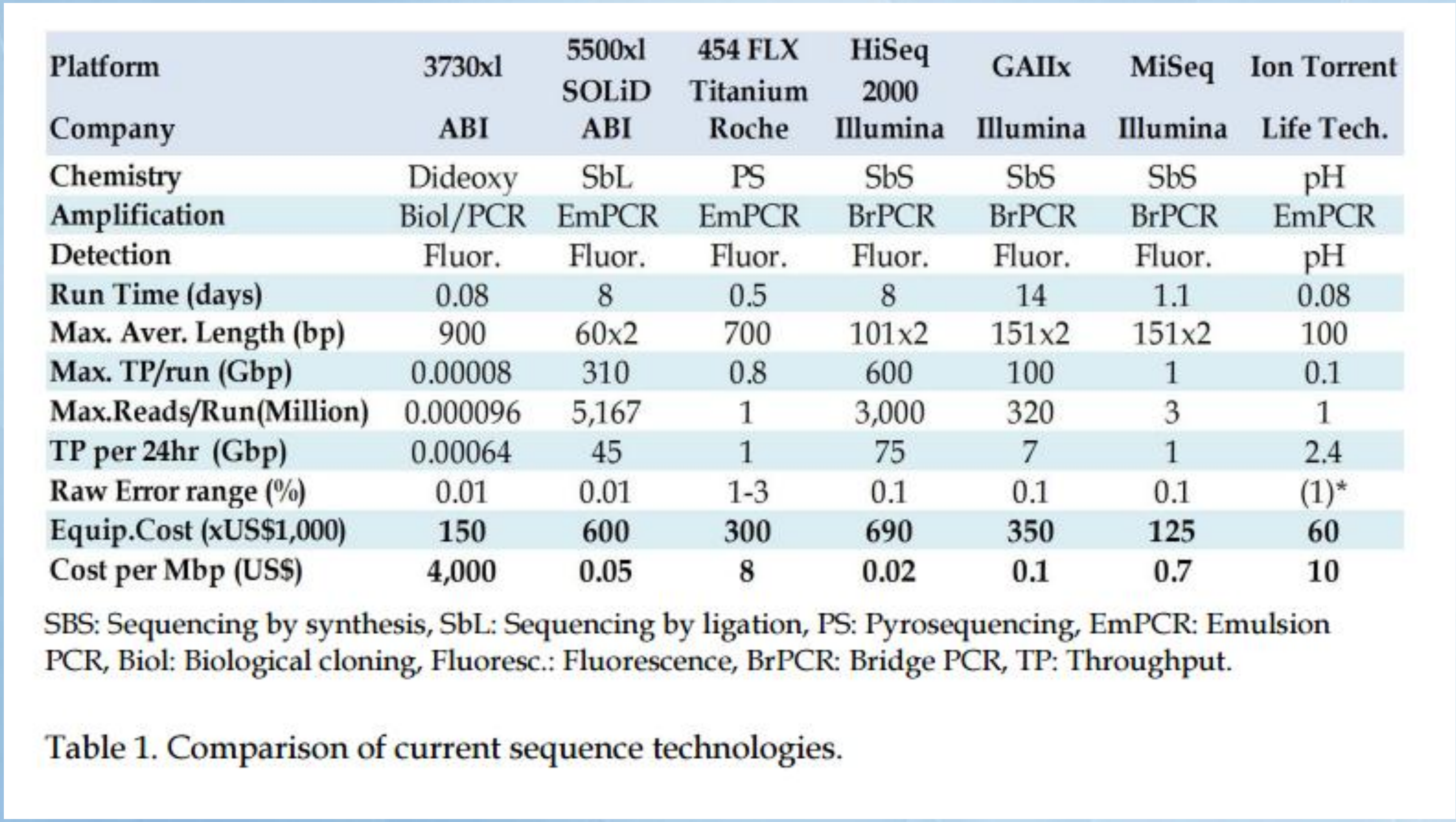
- Roche/454
  - (GS FLX+/GS Junior)
- Illumina Genome Analyzer
  - (HiSeq/MiSeq/NextSeq)
- Life Technologies
  - (3500 Genetic Analyzer, Ion Torrent Proton/PGM)
- Pacific Biosciences
  - (PACBIO RSII)
- Applied Biosystems
  - (SOLiD, 3730*xl* DNA Analyzer)

Sequencing Principles

- Sequencing by Synthesis
  - Sanger/Dideoxy chain termination (Life Technologies, Applied Biosystems)
  - Pyrosequencing (Roche/454)
  - Reversible terminator (Illumina)
  - Ion torrent (Life Technologies)
  - Zero Mode Waveguide (Pacific Biosciences)

- Sequencing by Oligo Ligation Detection
  - SOLiD (Applied Biosystems)

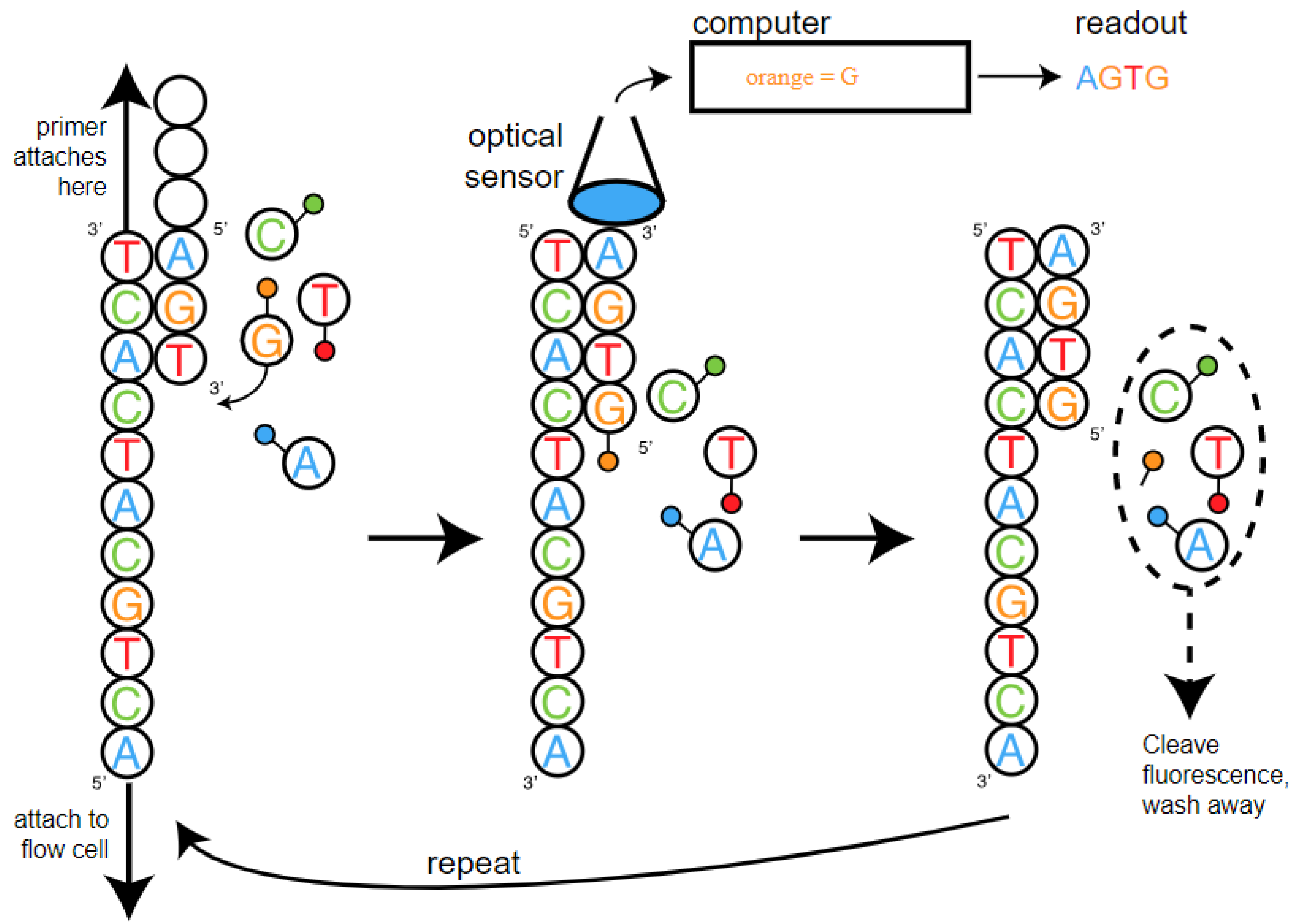
- Direct reading of DNA sequence
  - Nanopore sequencing 3rd generation sequencing
  - Electron microscope 3rd generation sequencing



Platform Company	3730xl ABI	5500xl SOLiD ABI	454 FLX Titanium Roche	HiSeq 2000 Illumina	GAIIx Illumina	MiSeq Illumina	Ion Torrent Life Tech.
Chemistry	Dideoxy	SbL	PS	SbS	SbS	SbS	pH
Amplification	Biol/PCR	EmPCR	EmPCR	BrPCR	BrPCR	BrPCR	EmPCR
Detection	Fluor.	Fluor.	Fluor.	Fluor.	Fluor.	Fluor.	pH
Run Time (days)	0.08	8	0.5	8	14	1.1	0.08
Max. Aver. Length (bp)	900	60x2	700	101x2	151x2	151x2	100
Max. TP/run (Gbp)	0.00008	310	0.8	600	100	1	0.1
Max.Reads/Run(Million)	0.000096	5,167	1	3,000	320	3	1
TP per 24hr (Gbp)	0.00064	45	1	75	7	1	2.4
Raw Error range (%)	0.01	0.01	1-3	0.1	0.1	0.1	(1)*
Equip.Cost (xUS$1,000)	**150**	**600**	**300**	**690**	**350**	**125**	**60**
Cost per Mbp (US$)	**4,000**	**0.05**	**8**	**0.02**	**0.1**	**0.7**	**10**

SBS: Sequencing by synthesis, SbL: Sequencing by ligation, PS: Pyrosequencing, EmPCR: Emulsion PCR, Biol: Biological cloning, Fluoresc.: Fluorescence, BrPCR: Bridge PCR, TP: Throughput.

Table 1. Comparison of current sequence technologies.



Sequencing by Synthesis. dNTP fluorescence is translated to a base call.

Phospholinked Fluorophores

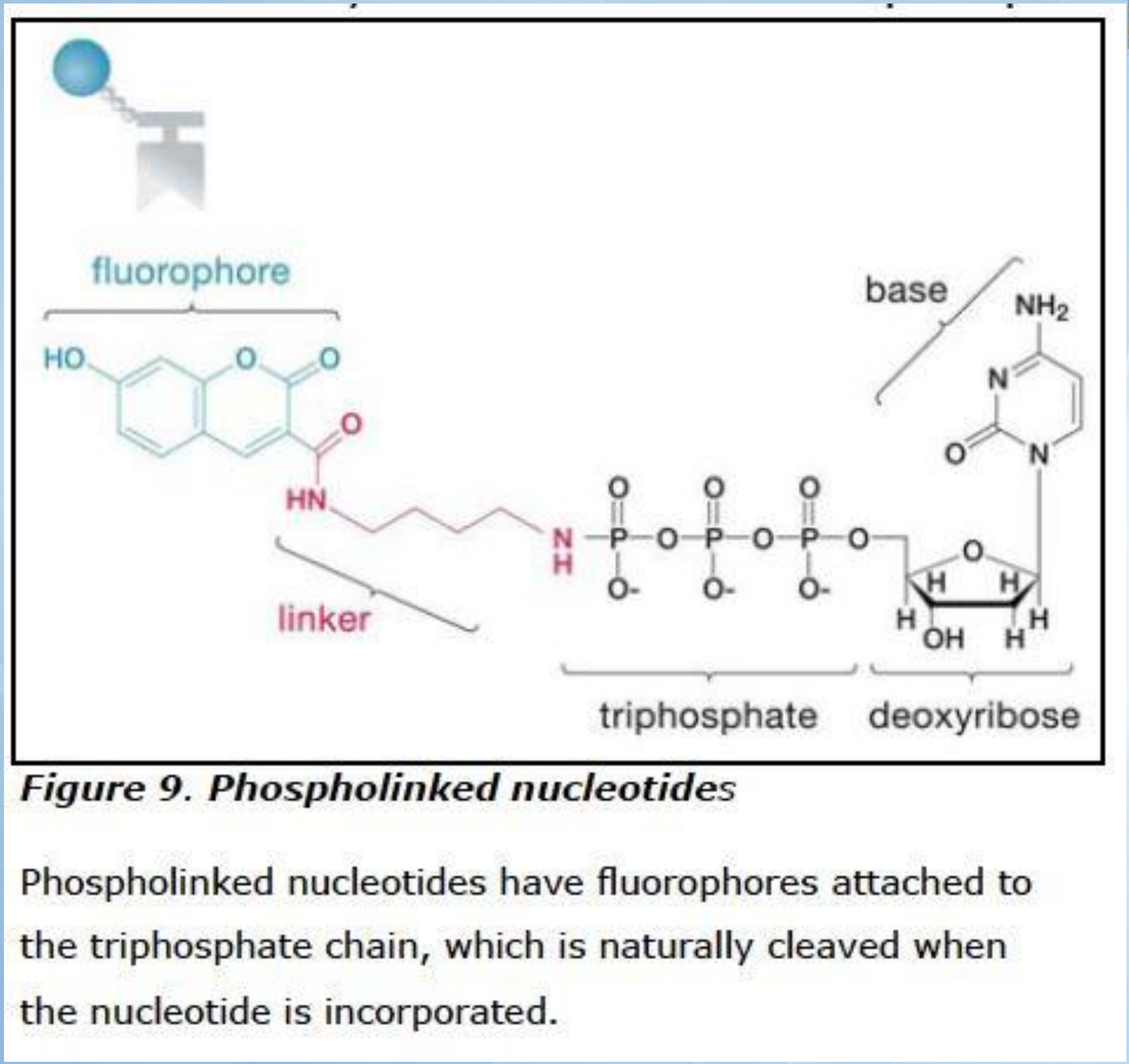
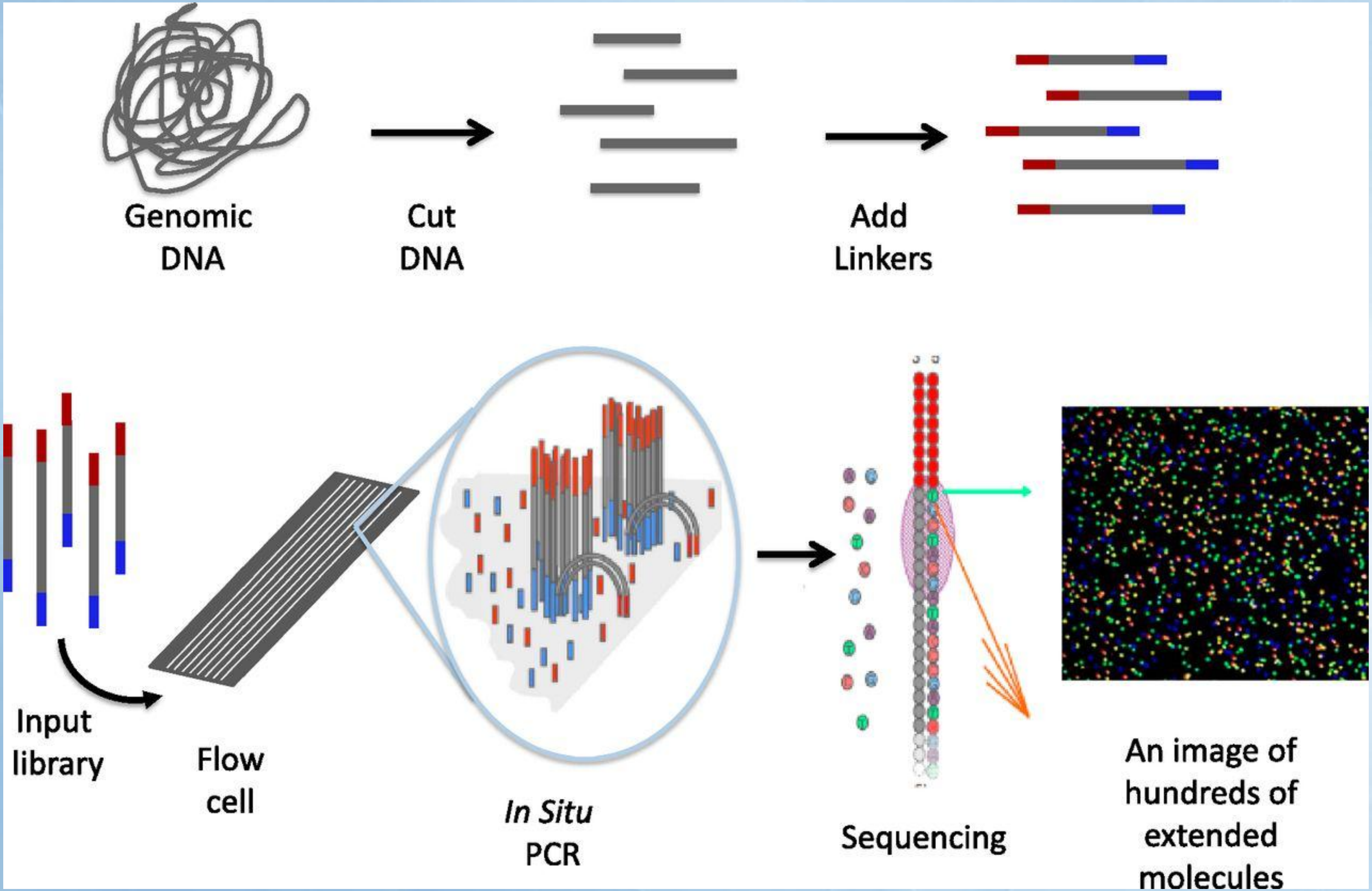


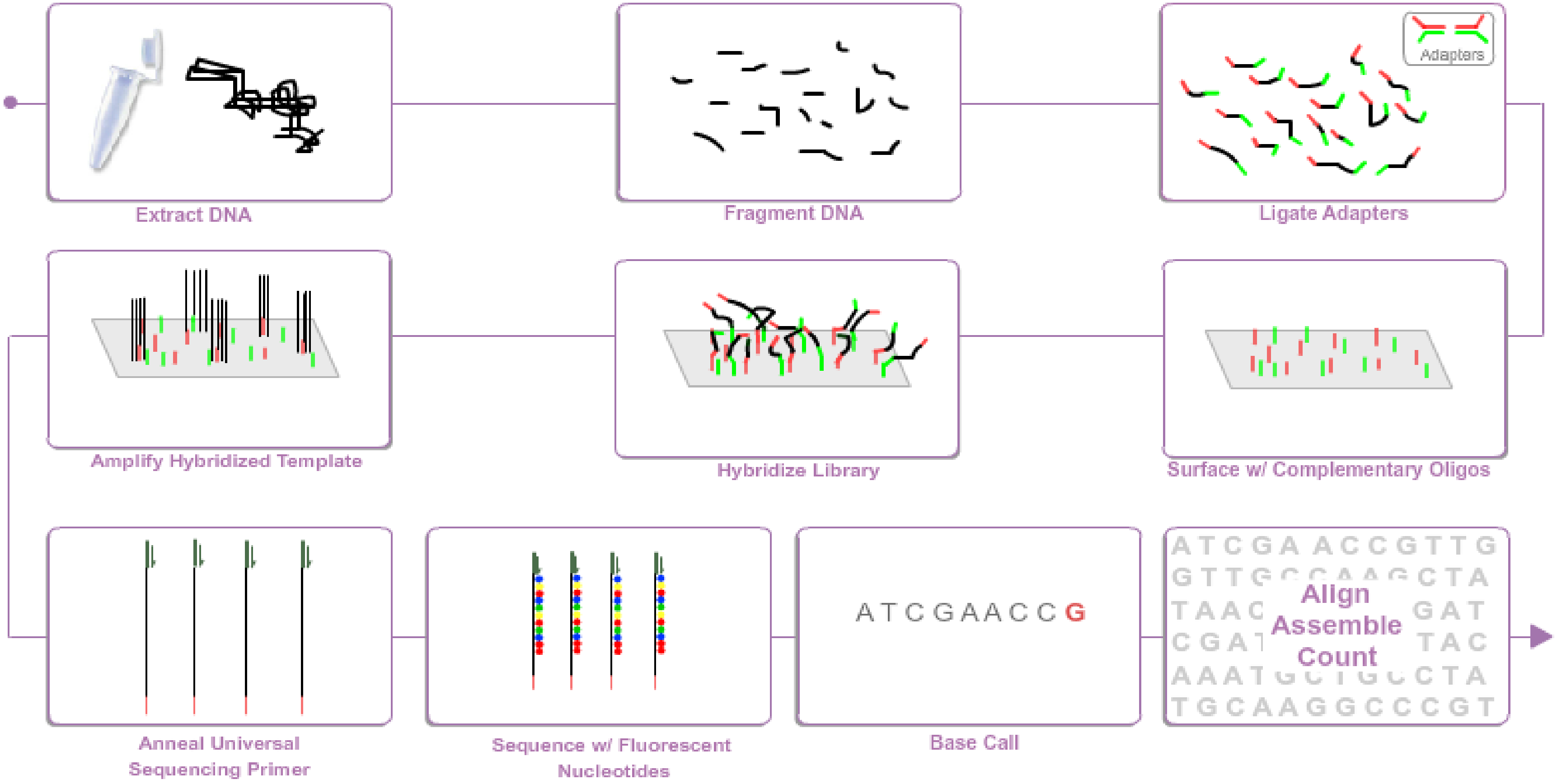
Figure 9. Phospholinked nucleotides

Phospholinked nucleotides have fluorophores attached to the triphosphate chain, which is naturally cleaved when the nucleotide is incorporated.

Illumina sequencing



Sequencing by Synthesis (SBS) Overview



Extract DNA

Fragment DNA

Ligate Adapters

Surface w/ Complementary Oligos

Hybridize Library

Amplify Hybridized Template

Anneal Universal Sequencing Primer

Sequence w/ Fluorescent Nucleotides

Base Call

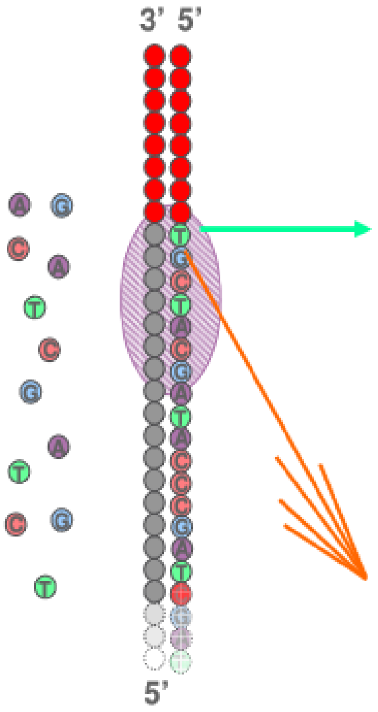
Align

Assemble

Count

Illumina sequencing can sequence billions of template strands simultaneously

A picture is taken every time a new base is added



Sequencing

36bp – 100bp

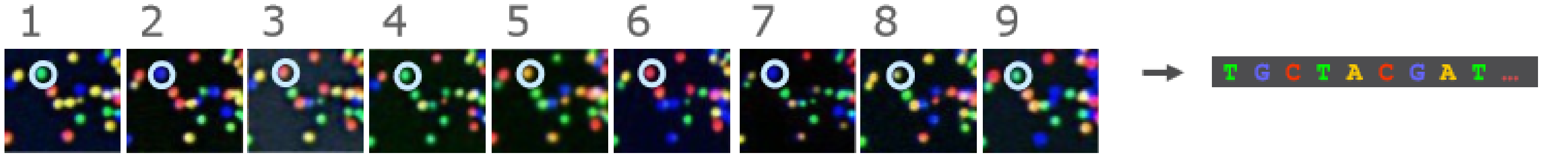
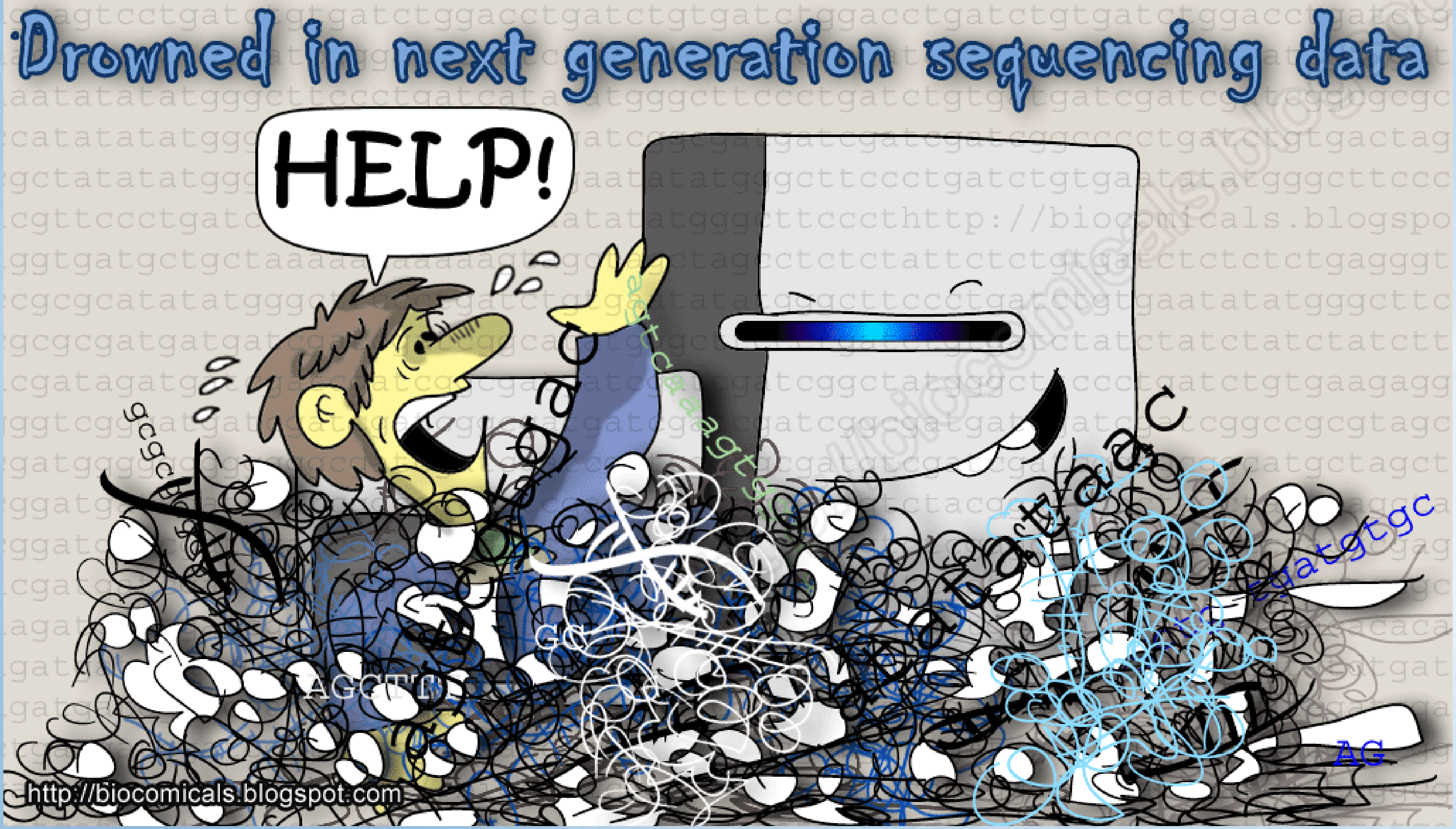


Image acquisition

Base calling


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com