Systems Microbiology

Monday Oct 23 - Ch 15 -Brock

Genomics

- DNA sequencing technology
- Genome sequencing technology
- Current statistics
- Basics of genome sequence analysis

Human genome sequence analysis reached a first stage of completion in the summer of 2000 summer by the New York Times:

Scanned magazine and newspaper articles about J. Craig Venter
and Francis Collins removed due to copyright restrictions.

The Race to Sequence

Cartoon image of J. Craig Venter and Francis Collins racing to the finishremoved due to copyright restrictions.

- Celera - J. Craig Venter
- NHGRI - Francis Collins
- G5
  - Sanger Centre
  - Whitehead Institute
  - JGI Walnut Creek
  - Washington University
  - Baylor

Human Genome Project

- HGP drove innovation in biotechnology
- 2 major technological benefits
  - stimulated development of *high throughput methods* -- the assembly line (or dis-assembly line) meets biological research
  - reliance on *computational tools* for data mining and visualization of biological information

DNA Sequencing

- ✓ **DNA sequencing = determining the nucleotide sequence of DNA.**
- ✓ **Developed by Frederick Sanger in the 1970s.**

First whole DNA genome sequenced was a virus, PhiX174 in 1977 - 5386 bp.

First demo of Sanger dideoxynucleotide nucleotide sequencing technique.

Photographs removed due to copyright restrictions.

1980: Walter Gilbert (Biol. Labs) & Frederick Sanger (MRC Labs)

Manual Dideoxy DNA sequencing-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**
    2. **Primer annealed to template DNA**
    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.**

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

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

7. **Each of the four reaction mixtures produces a population of DNA molecules with DNA chains terminating at all possible positions.**

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

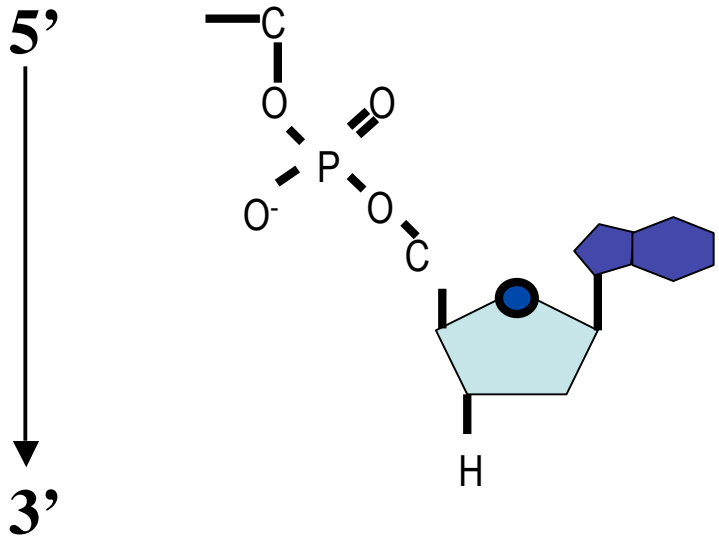
8. **Extension products in each of the four reaction mixutes also end with a different radio-labeled ddNTP (depending on the base).**

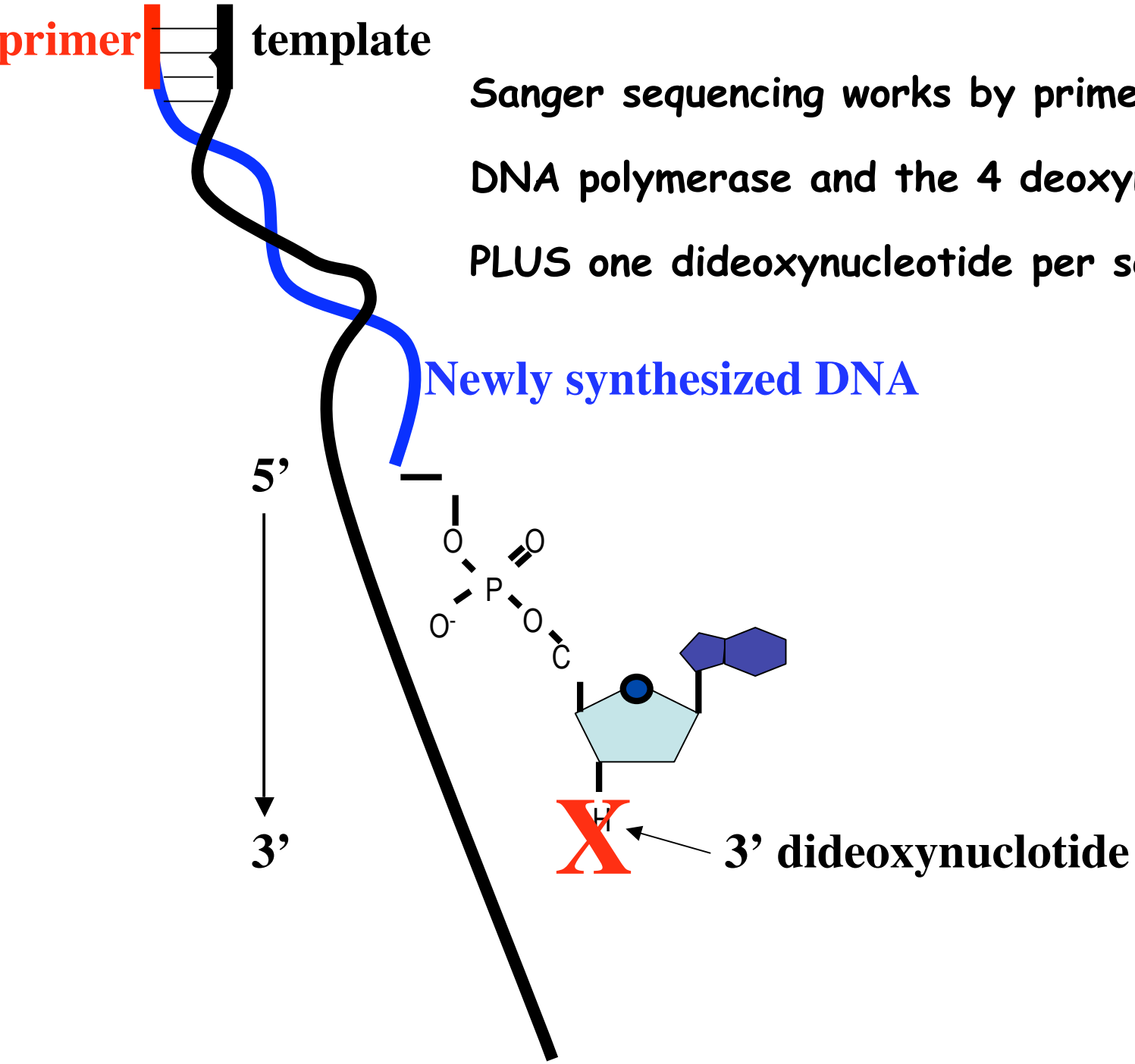
9. **Next, each reaction mixture is electrophoresed in a separate lane (4 lanes) at high voltage on a polyacrylamide gel.**

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

11. **Location of "bands" in each of the four lanes indicate the size of the fragment terminating with a respective radio-labeled ddNTP.**

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





Sanger sequencing works by primer extension using

DNA polymerase and the 4 deoxynucleotide triphosphates

PLUS one dideoxynucleotide per sequencing rxn.

Images of polyacrylamide gels removed due to copyright restrictions.

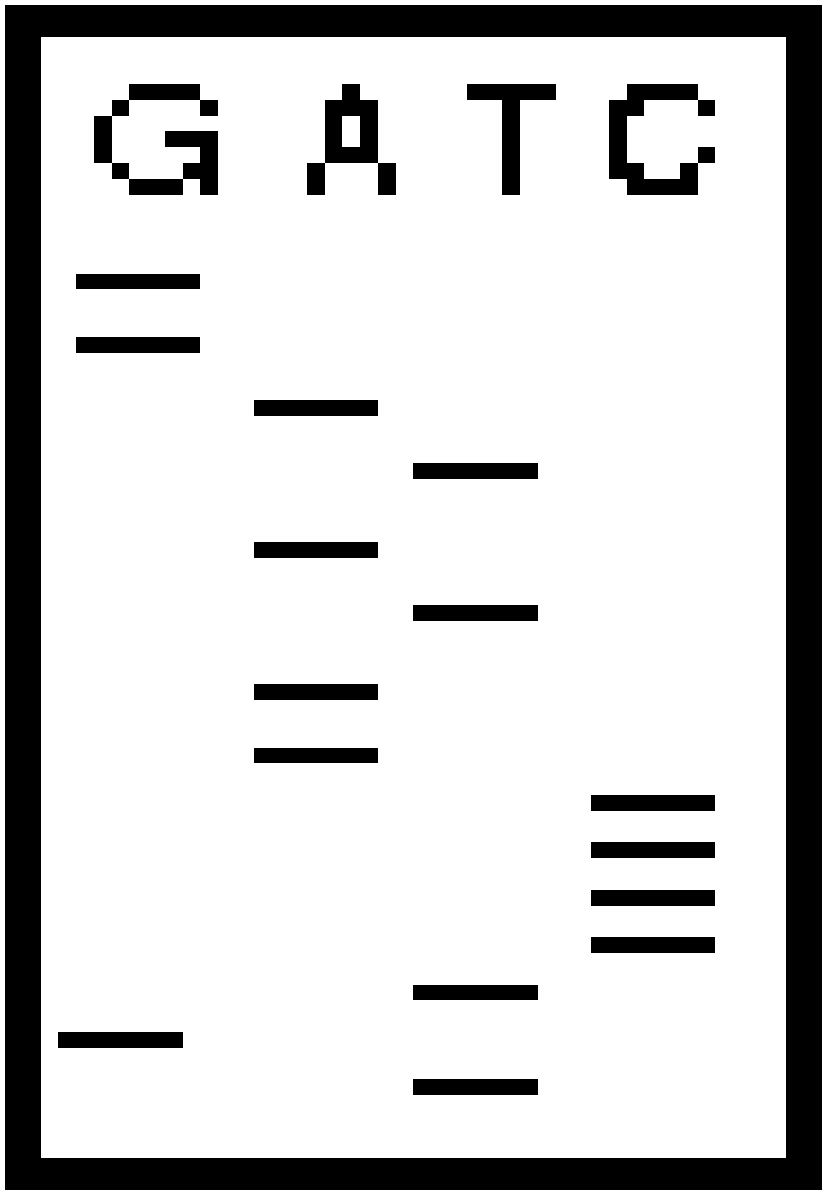
Vigilant et al. 1989
PNAS 86:9350-9354

(polyarcrylamide gel)

Radio-labeled ddNTPs (4 rxns)

Sequence (5′ to 3′)

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



Long products

Short products

Automated Dye-Terminator DNA Sequencing:

1. **Dideoxy DNA sequencing was time consuming, radioactive, and throughput was low, typically ~300 bp per run.**

2. **Automated DNA sequencing employs the same general procedure, but uses ddNTPs labeled with fluorescent dyes.**

3. **Combine 4 dyes in one reaction tube and electrophores in one lane on a polyacrylamide gel or capillary containing polyacrylamide.**

4. **UV laser detects dyes and reads the sequence.**

5. **Sequence data is displayed as colored peaks (chromatograms) that correspond to the position of each nucleotide in the sequence.**

6. **Throughput is high, up to 1,200 bp per reaction and 96 reactions every 3 hours with capillary sequencers.**

7. **Most automated DNA sequencers can load robotically and operate around the clock for weeks with minimal labor.**

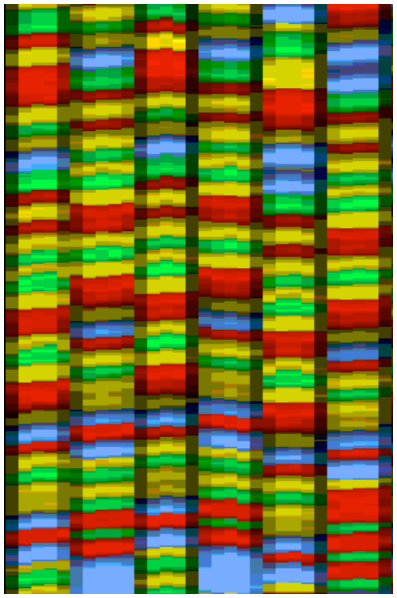
Applied Biosystems PRISM 377
(Gel, 34-96 lanes)

Applied Biosystems PRISM 3700
(Capillary, 96 capillaries)

Photographs of equipment removed due to copyright restrictions.

Applied Biosystems PRISM 3100
(Capillary, 16 capillaries)

"virtual autorad" - real-time DNA sequence output from ABI 377



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.**

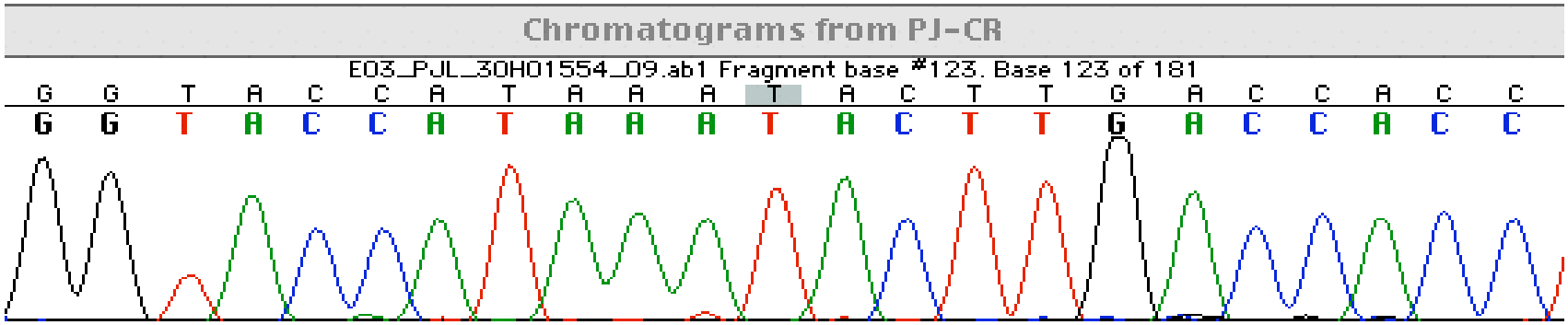


Table and graph showing the growth of GenBank removed due to copyright restrictions.

Fraser, Eisen, and Salzman **Shotgun genome sequencing**

Nature **406**, 799-803 (17 August 2000)

Images removed due to copyright restrictions.

Small insert cloning/seqing

Blue white selection of insert containing clones based on lacZ (beta galactosidase) activity

Image removed due to copyright restrictions.
See Figure 15-1bc in Madigan, Michael, and John Martinko. *Brock Biology of Microorganisms*. 11th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006. ISBN: 0131443291.

~~Gel readings resulting~~ from the sequencing of clones are overlapped - using sequence identities - to obtain large segments

Fluorescent Sanger dideoxy seqing rxns, & capillary electrophoresis !

Resulting "contigs" are combined to assembly the whole genome

By overlapping individual readings, a genome may be covered several times, therefore each base is confirmed by multiple readings in both directions.

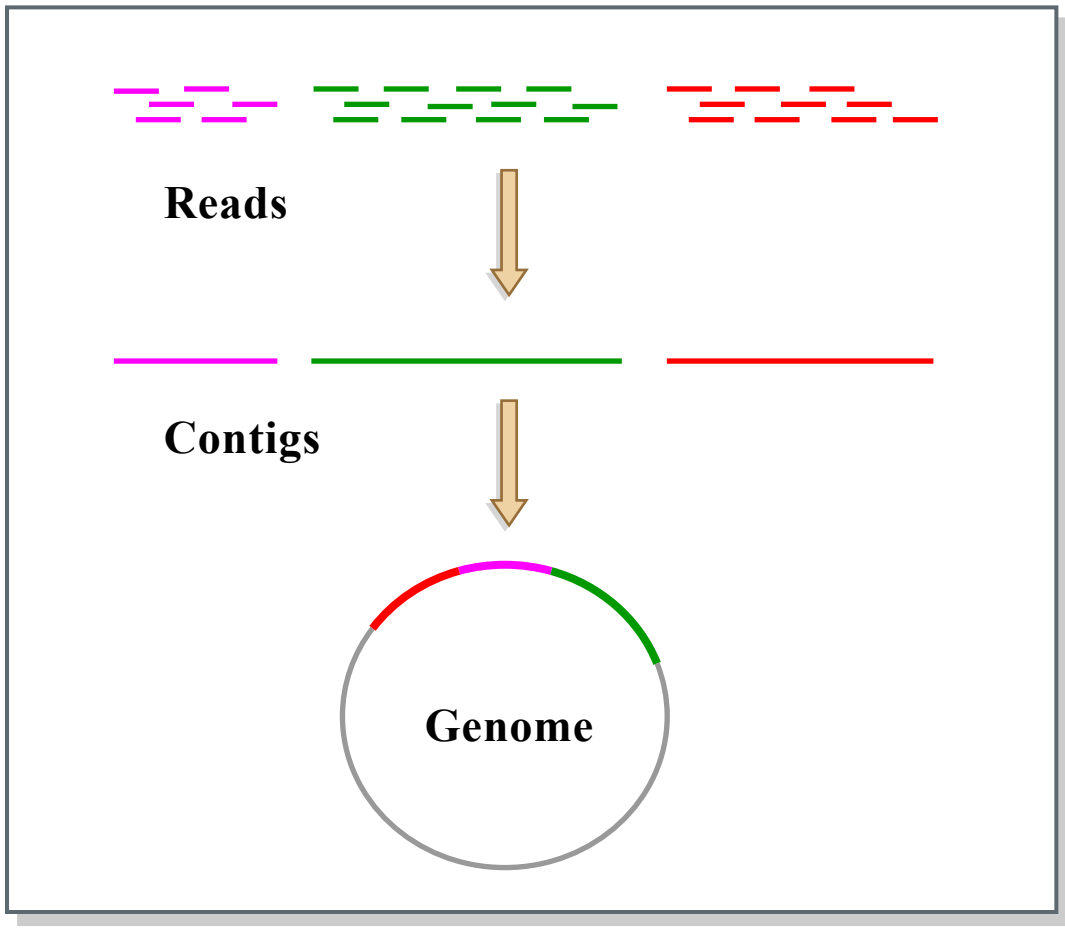


Figure by MIT OCW.

(typically, shotgun seqing done to 8X min.)

Shotgun sequencing: A Bottom Up, Random Strategy

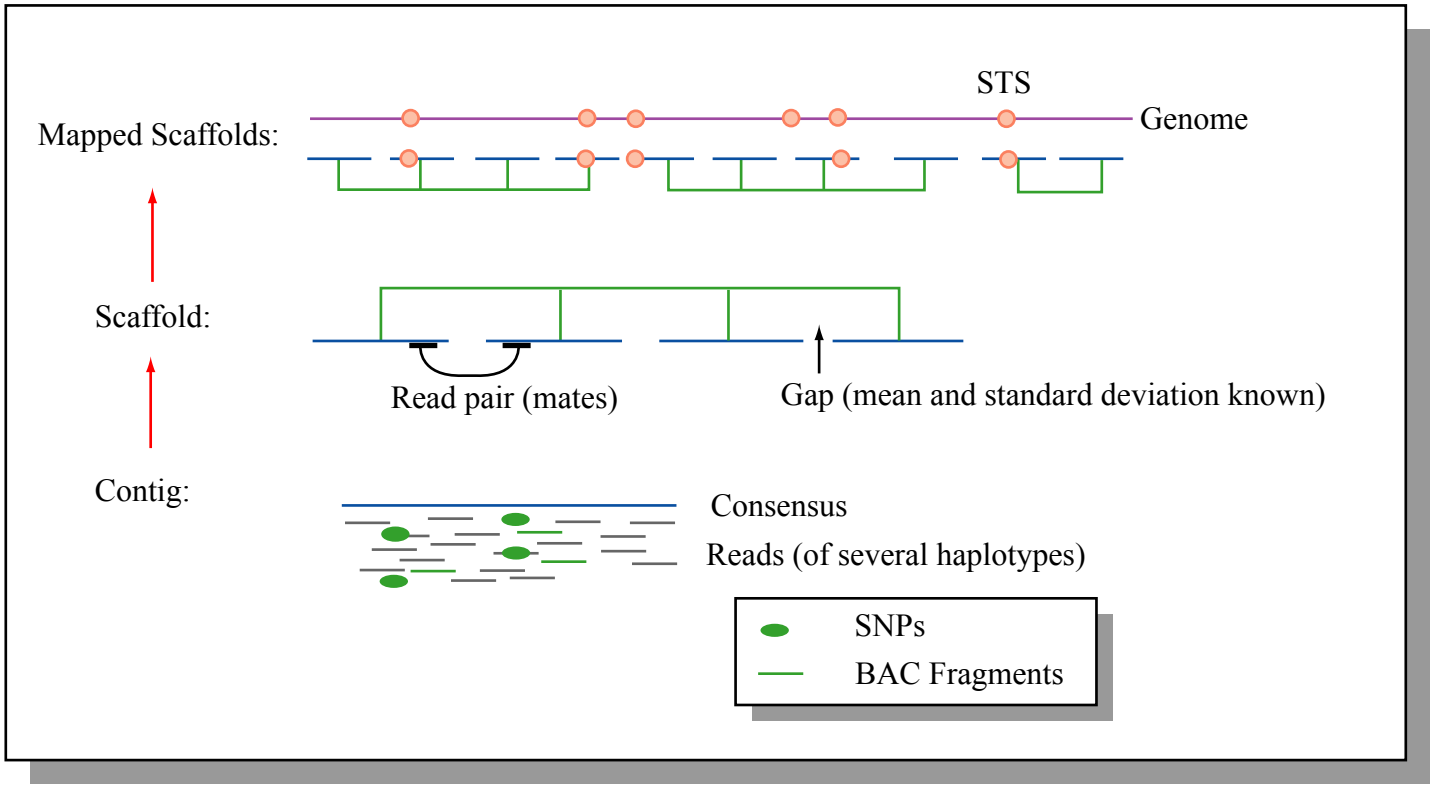


Figure by MIT OCW.

The key to the strategy is shotgun cloning of random (not mapped) fragments of a **defined length** (2kb or 10kb) so that **scaffolds** can be assembled from **terminal sequences** of each cloned insert.

DNA Sequence Assembly

Computer screenshot showing DNA sequencing removed due to copyright restrictions.

Large insert cloning/seqing

Image removed due to copyright restrictions.
See Figure 15-2 in Madigan, Michael, and John Martinko. *Brock Biology of Microorganisms.* 11th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006. ISBN: 0131443291.

BAC sequencing: A Top down Up Strategy Using Maps

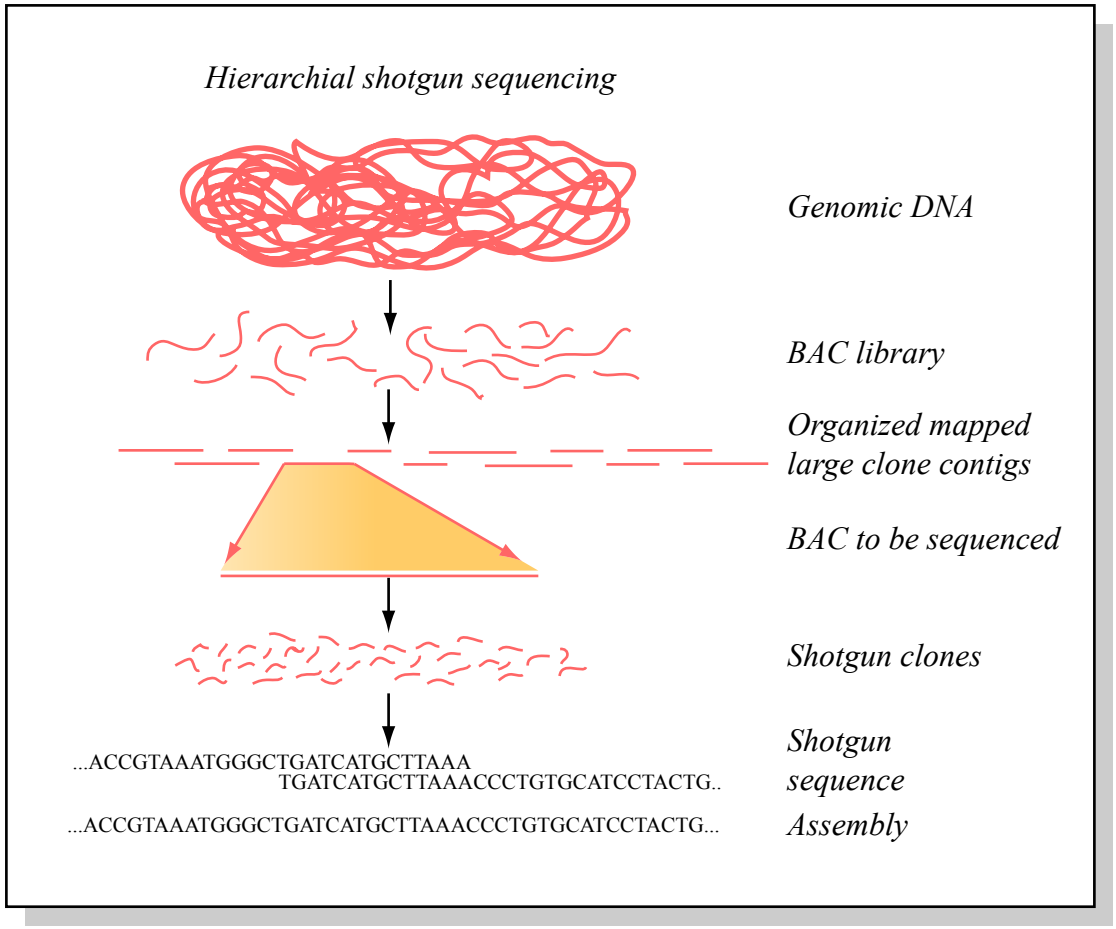


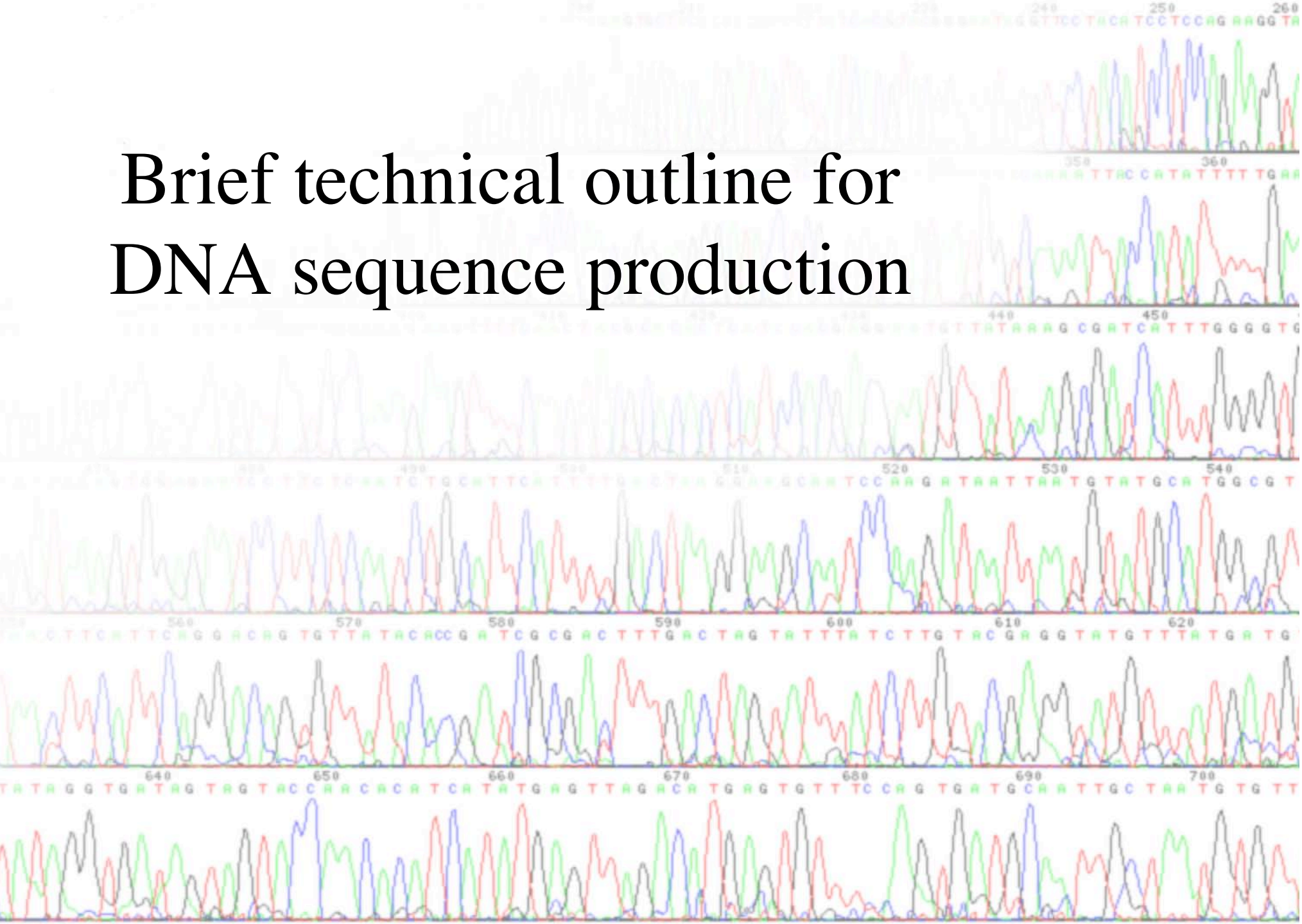
Figure by MIT OCW.

BAC = Bacterial Artificial Chromosome (F-Factor based)

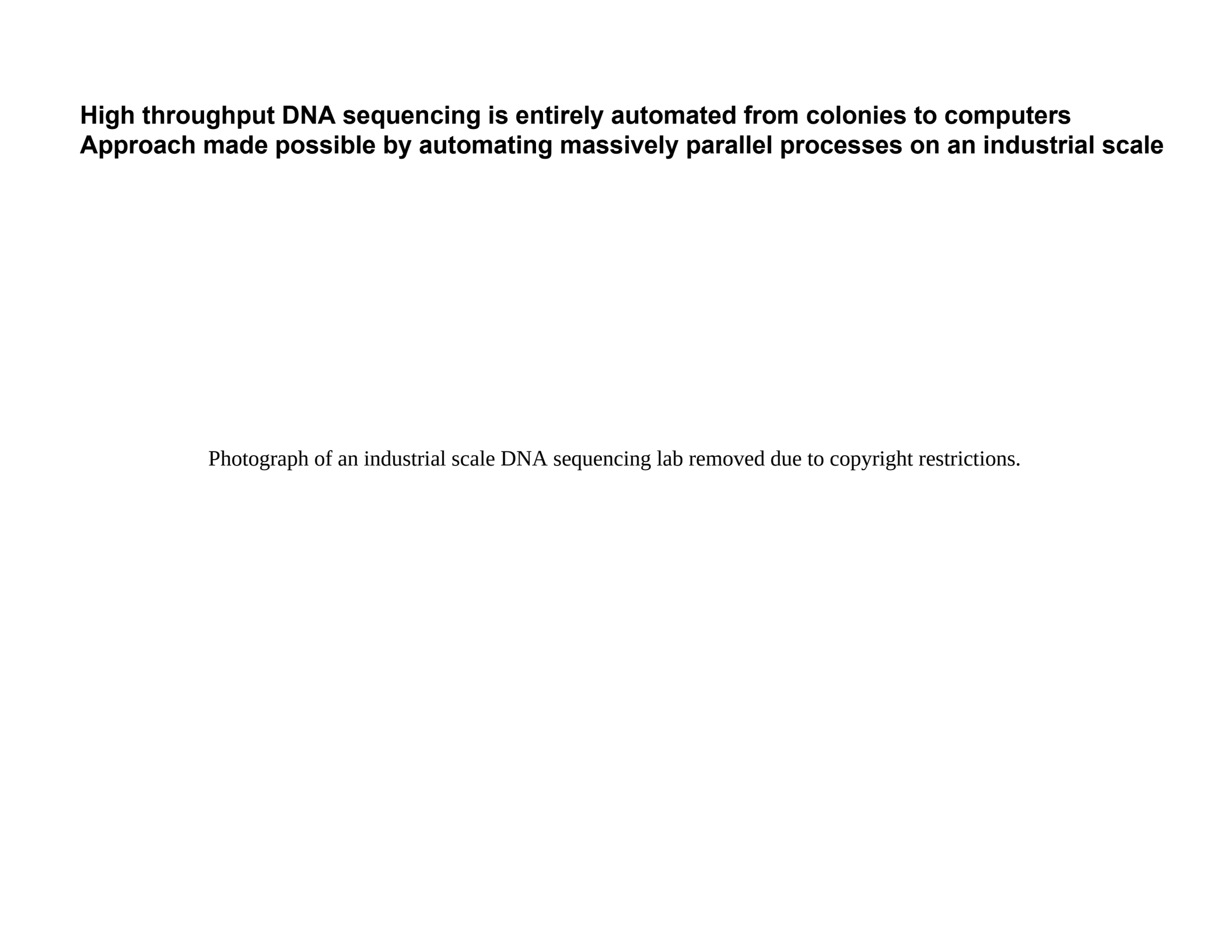
Image removed due to copyright restrictions.

The NIH consortium strategy was dependent on multiple clone overlaps

Brief technical outline for DNA sequence production



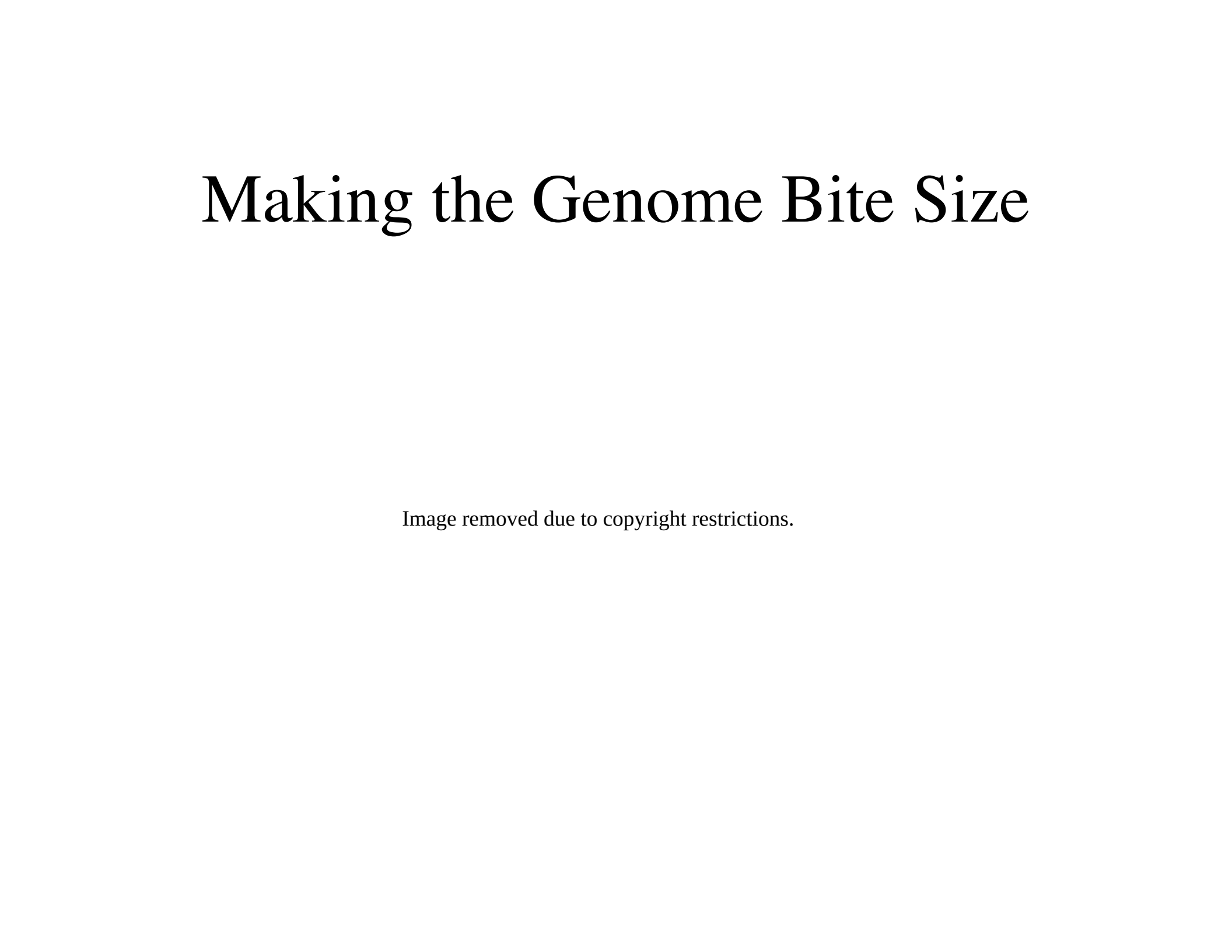
High throughput DNA sequencing is entirely automated from colonies to computers
Approach made possible by automating massively parallel processes on an industrial scale



Sequence Production Pipeline

- Pick & archive library clones
- Template production
- dideoxy terminator reactions
- Post-reaction processing
- Capillary sequencing
- Sequence data analysis - bioinformatics

Making the Genome Bite Size



Making a DNA fragment library

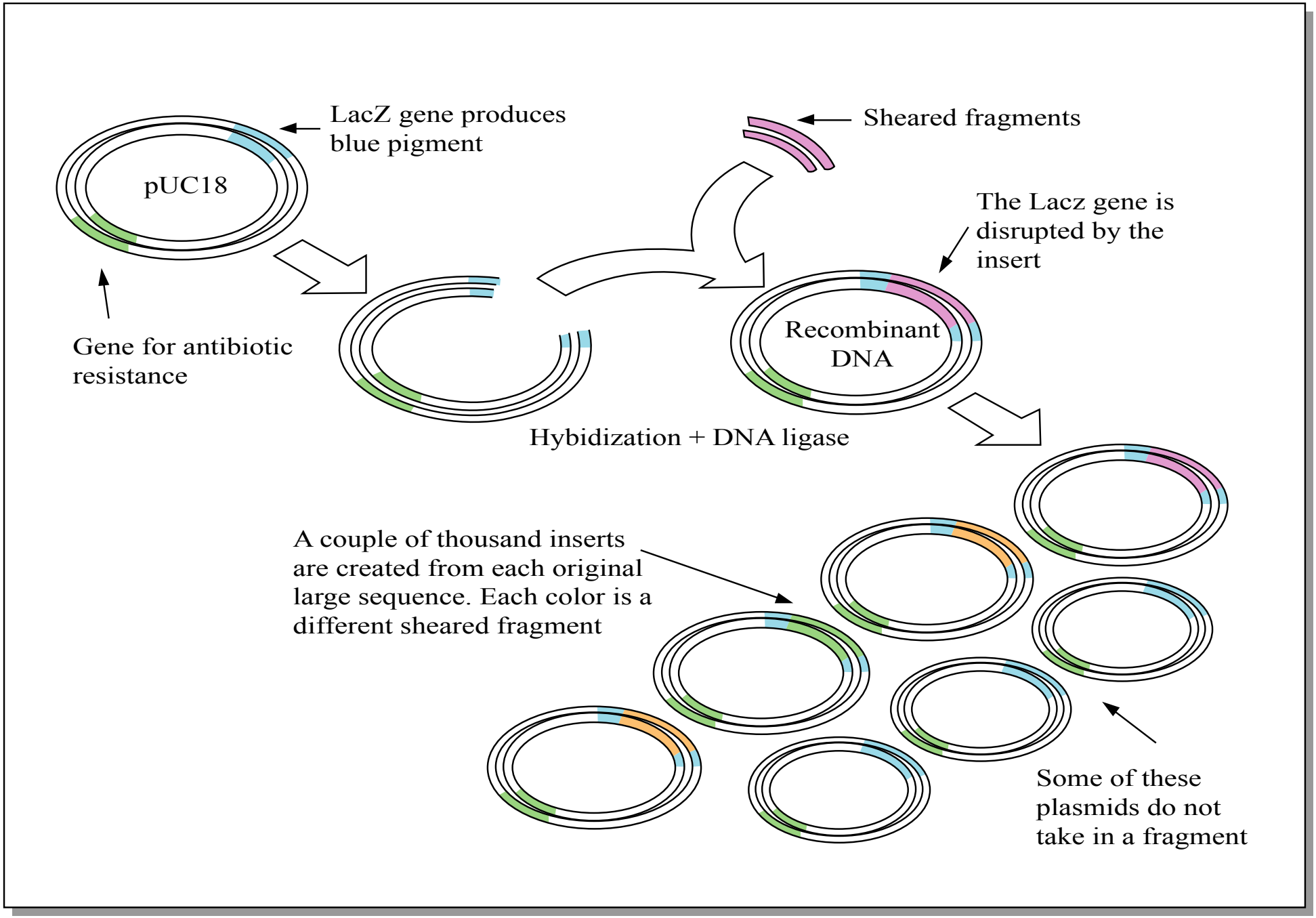


Figure by MIT OCW.

Plating a DNA Fragment Library

Plating photographs removed due to copyright restrictions.

Picking & Archiving Clones

- QPixII from Genetix, Ltd (UK)
- 4500 colonies per hour
- 300,000 colonies; 67 hrs; 384-well plates = 781

Various laboratory photographs removed due to copyright restrictions.

454 Corporation

The first commercial,
massively parallel,
DNA sequencing technology

Images removed due to copyright restrictions.

Deposit DNA Beads into PicoTiterPlate™

Images of DNA beads on PicoTiterPlates removed due to copyright restrictions.

Sequencing Instrument

Images removed due to copyright restrictions.

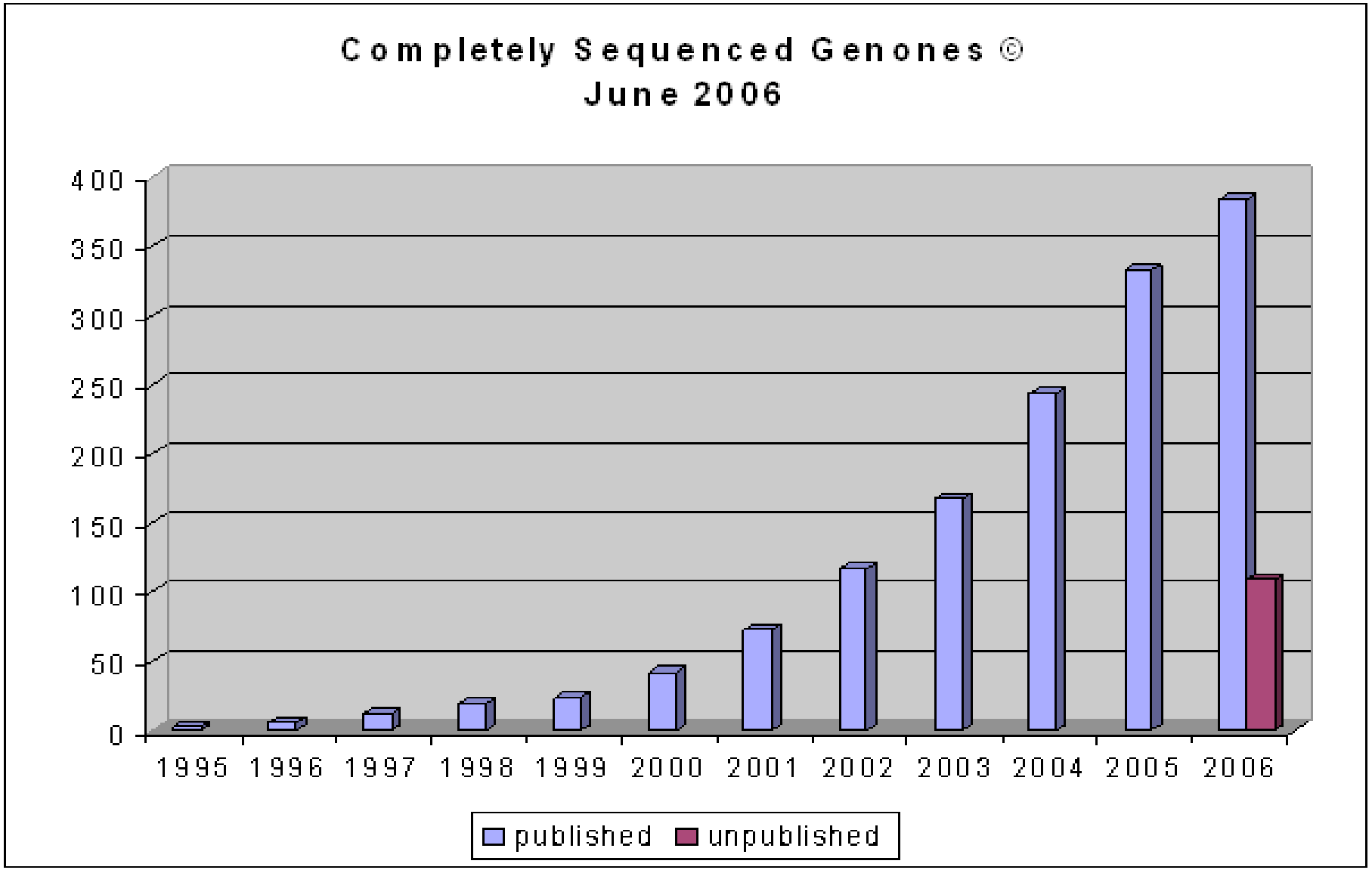
454 DNA sequencing technology

- 300,000 DNA fragments sequenced in parallel
- Average read length 100 - 110 nt
- 30 million nt per run
- One 454 run = 4 hours
- $5,000 reagent cost

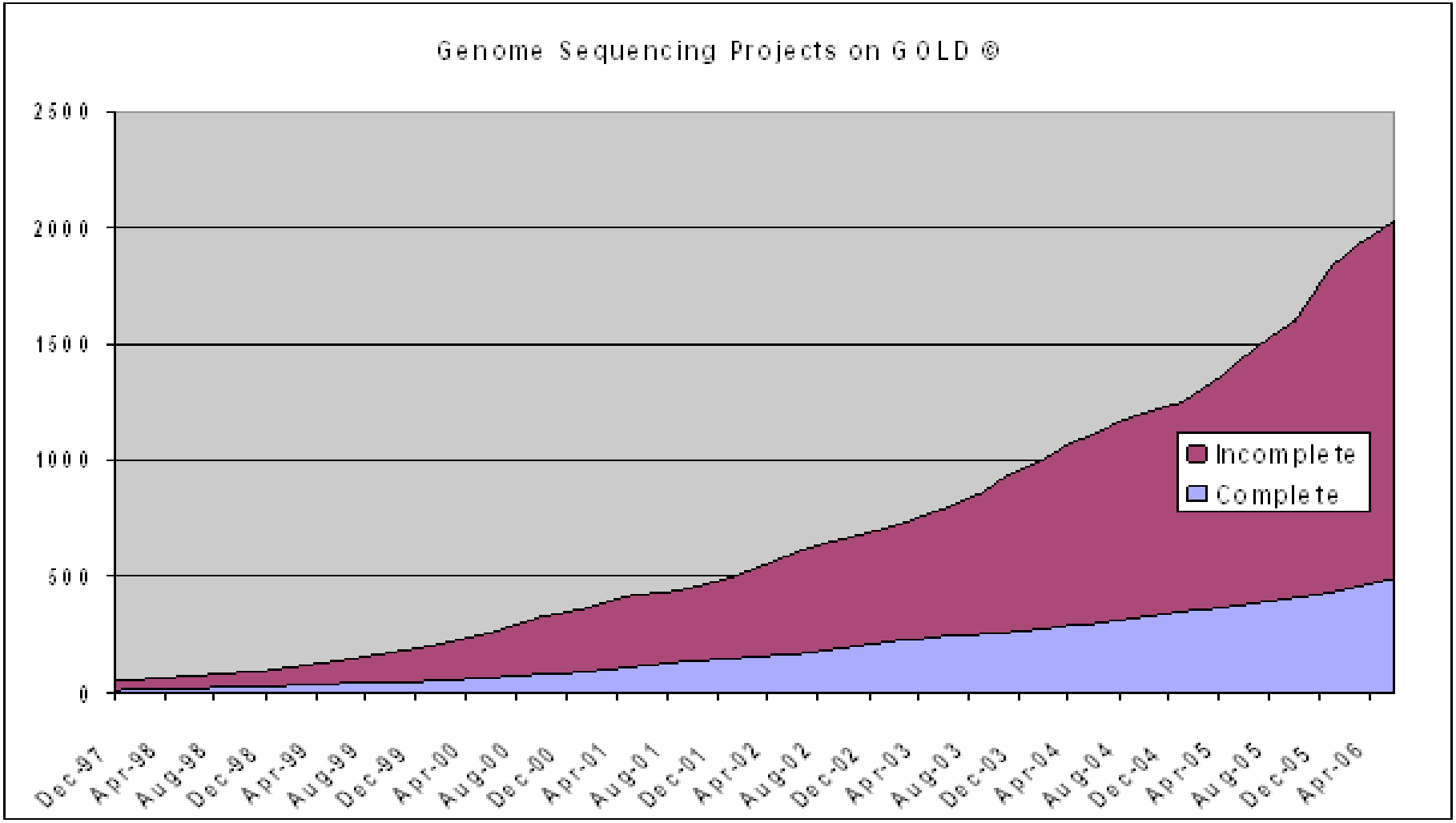
Company	Format	Read length	Throughput
Sequencing			
Amersham Biosciences (Uppsala, Sweden)	Capillary electrophoresis	550-1,000 bases	2.8 megabases/day (384-capillary instrument)
Applied Biosystems (Foster City, CA)	Capillary electrophoresis	500-950 bases (depending on column length)	2 megabases/day (production-scale system)
Lynx Therapeutics (Hayward, CA)	Massively parallel bead arrays	20 bases	8 megabases/day
Pyrosequencing (Uppsala, Sweden)	Real-time sequencing by synthesis	40-50bases/well (96- or 384-well plate)	110-691 kilobases/day
Resequencing			
454 Life Sciences (Branford, CT)	Massively parallel microfluidic, solid-surface sequencing	100 bases currently	60 megabases/day
Affymetrix (Santa Clara, CA)	High-density microarrays	30 kilobases/array currently; 500 kilobases/array projected	3 megabases/day
Perlegen (Mountain View, CA)	Whole-wafer Affymetrix microarray	15 megabase/wafer	300 megabases/day

Sequencing and resequencing technologies currently available

http://www.genomesonline.org/Gold_statistics.html

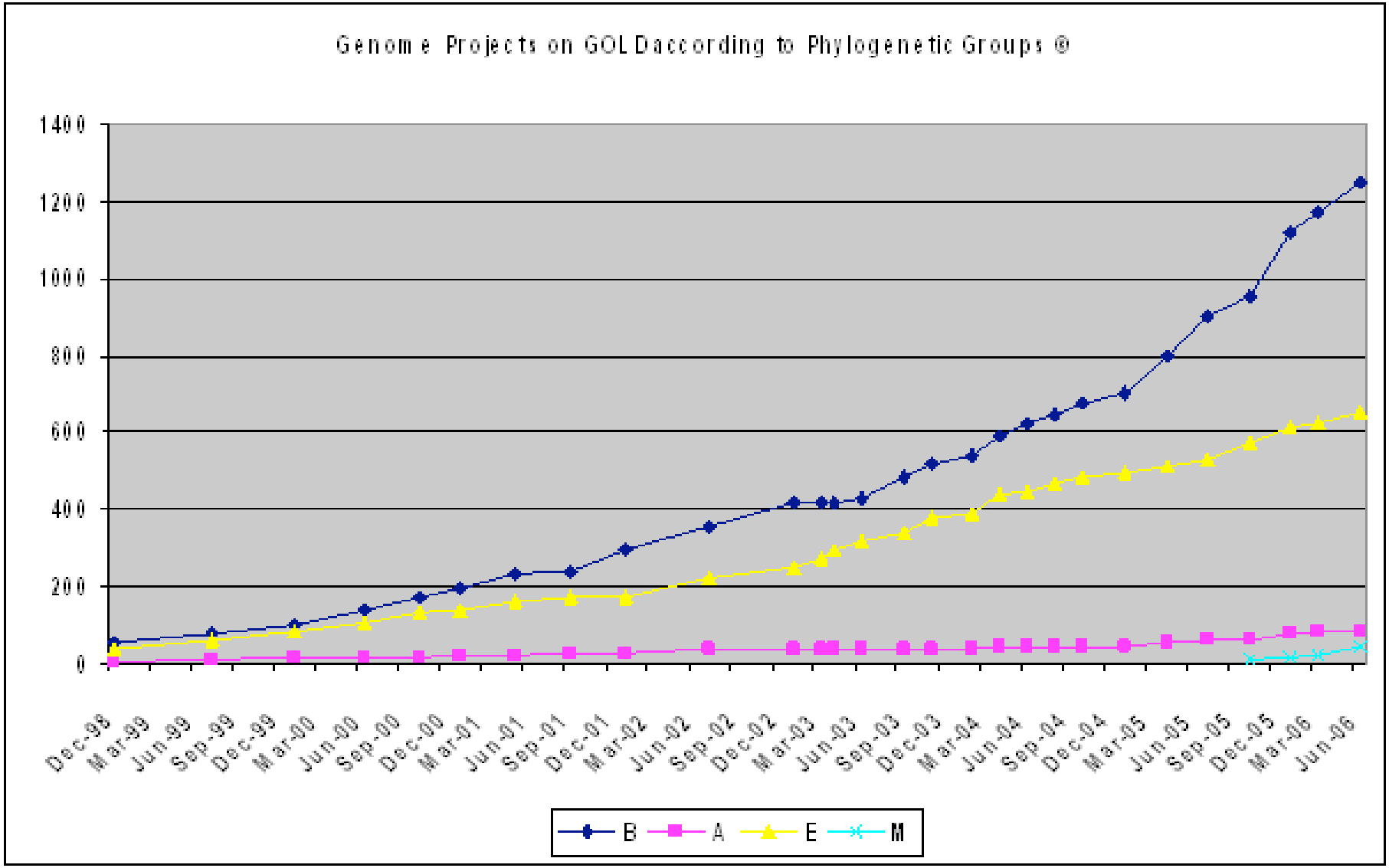


http://www.genomesonline.org/Gold_statistics.html



Courtesy of Genomes Online. Used with permission

http://www.genomesonline.org/Gold_statistics.html



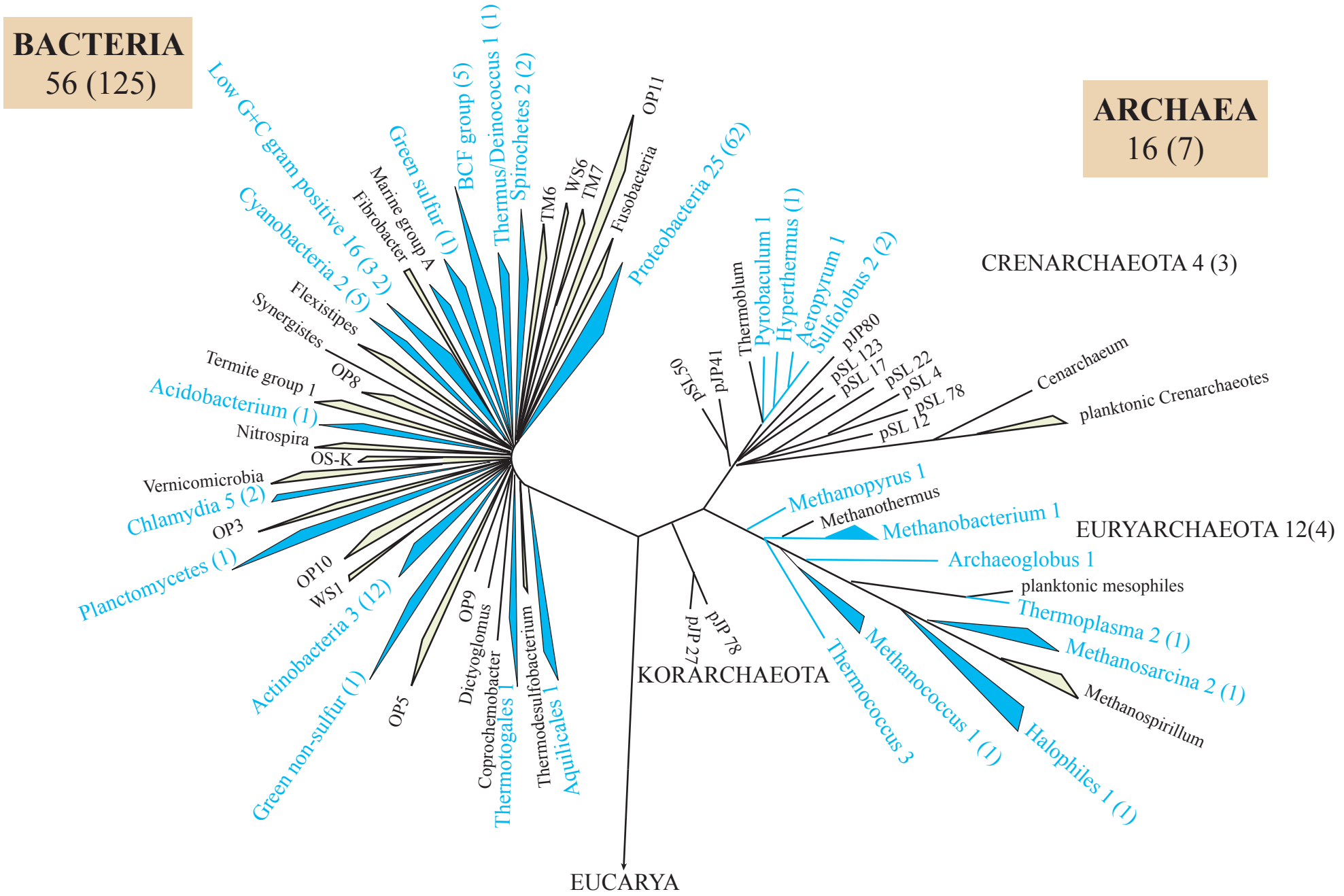
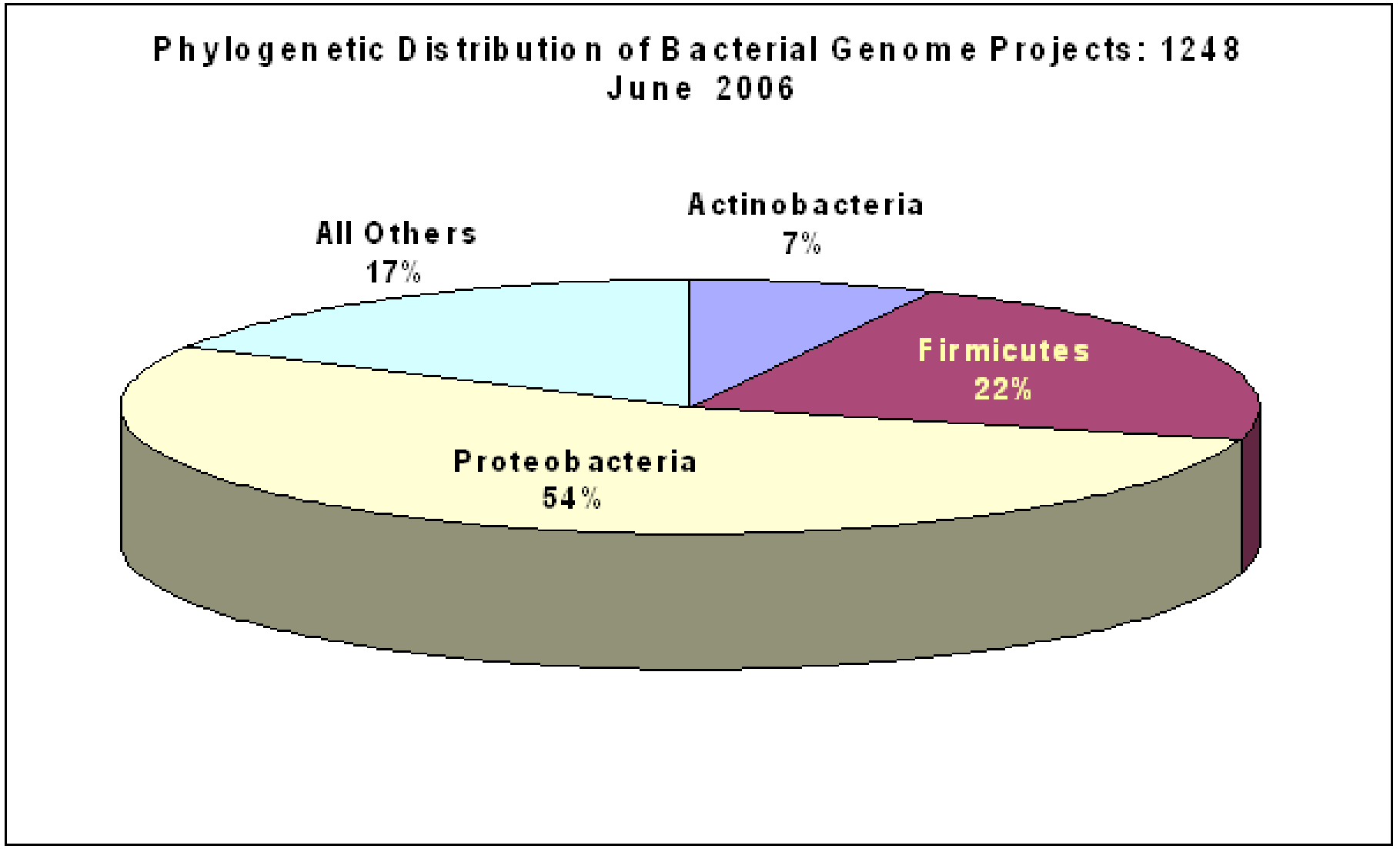
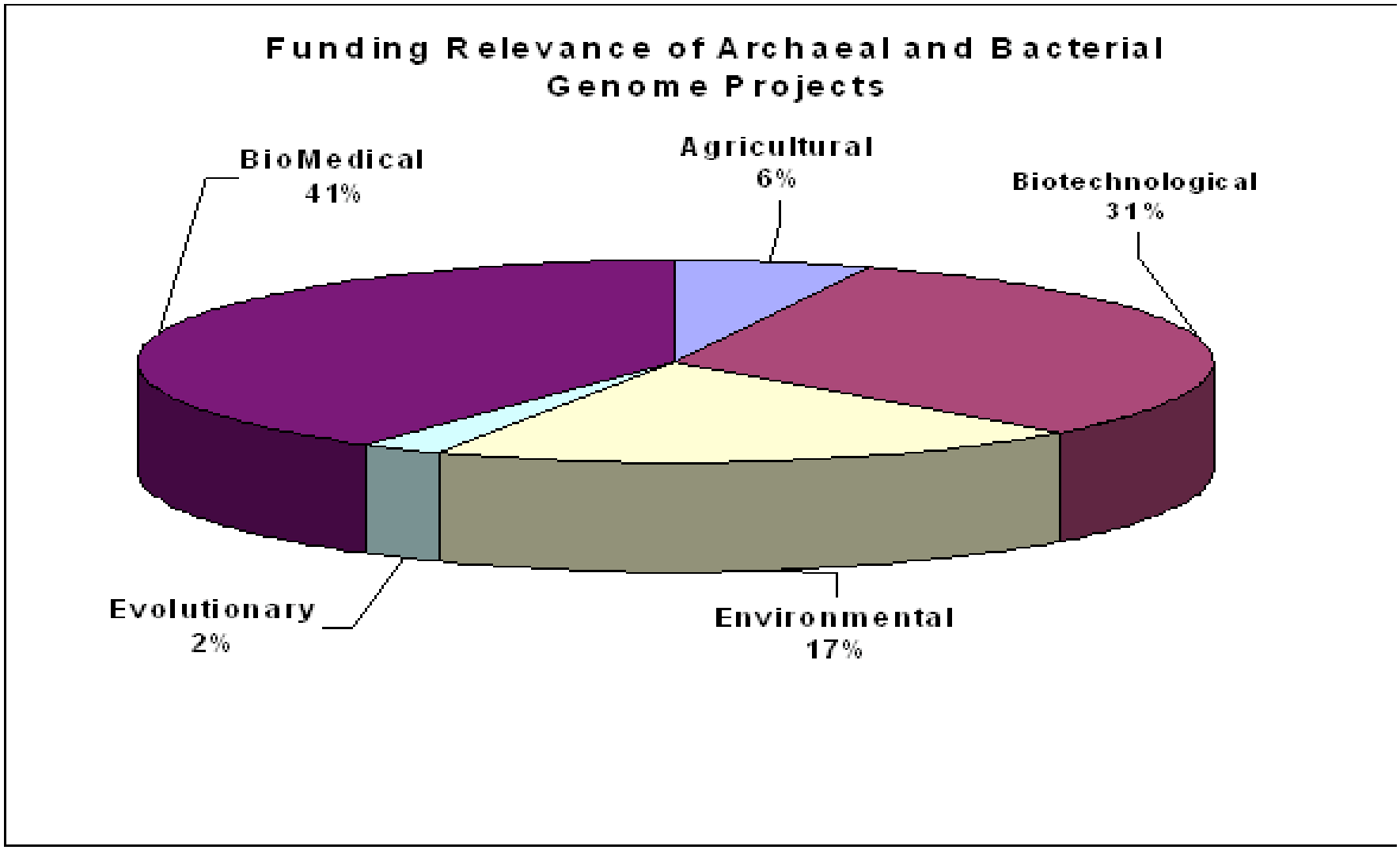


Figure by MIT OCW.

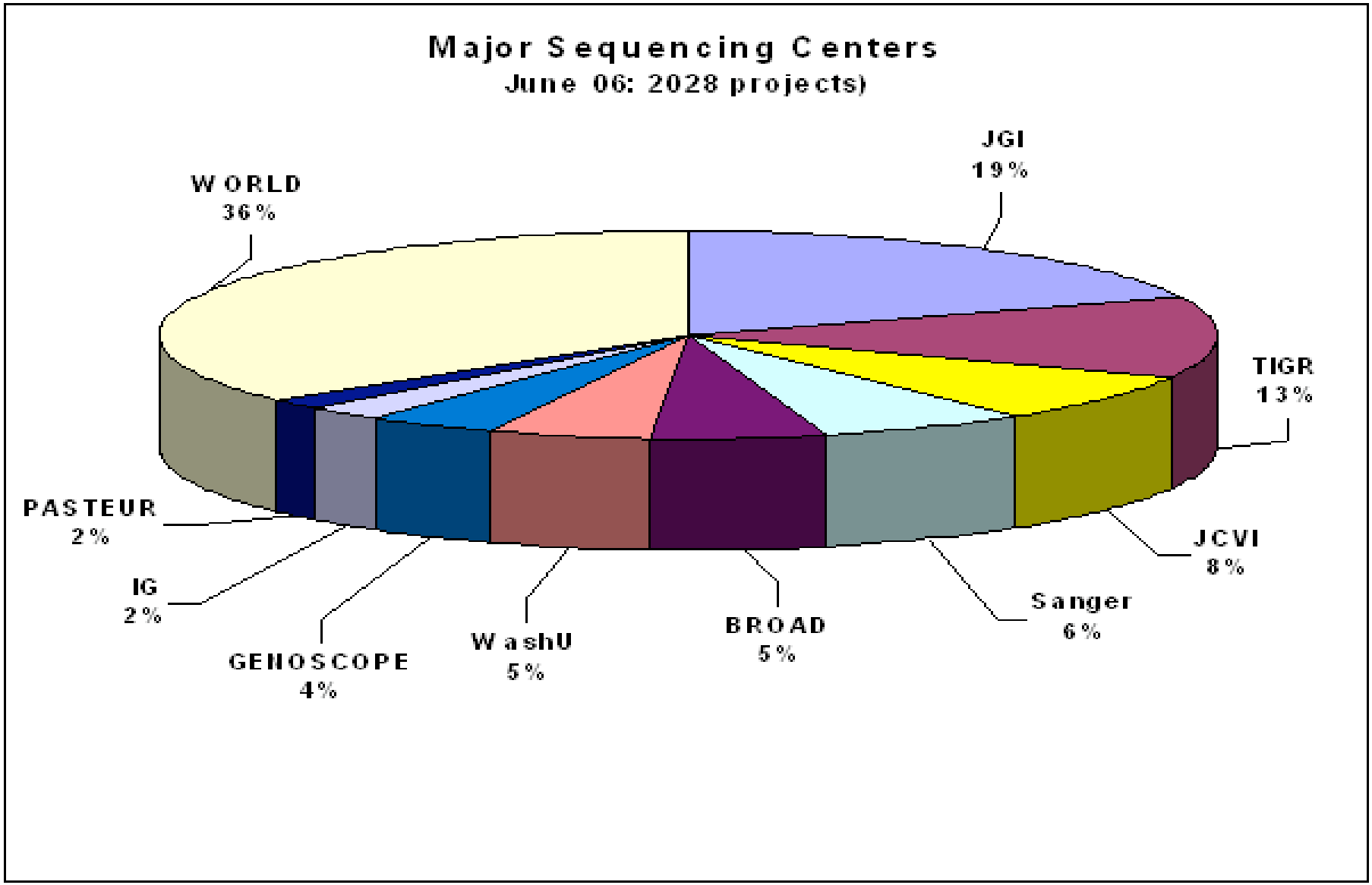


Courtesy of Genomes Online. Used with permission.

http://www.genomesonline.org/Gold_statistics.html



http://www.genomesonline.org/Gold_statistics.html



Tables removed due to copyright restrictions.
See Tables 15-1 and 15-3 in Madigan, Michael, and John Martinko. *Brock Biology of Microorganisms*. 11th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006. ISBN: 0131443291.

Genome Streamlining in a Cosmopolitan Oceanic Bacterium
Giovanonni et al, Science 309:1242 (2005)

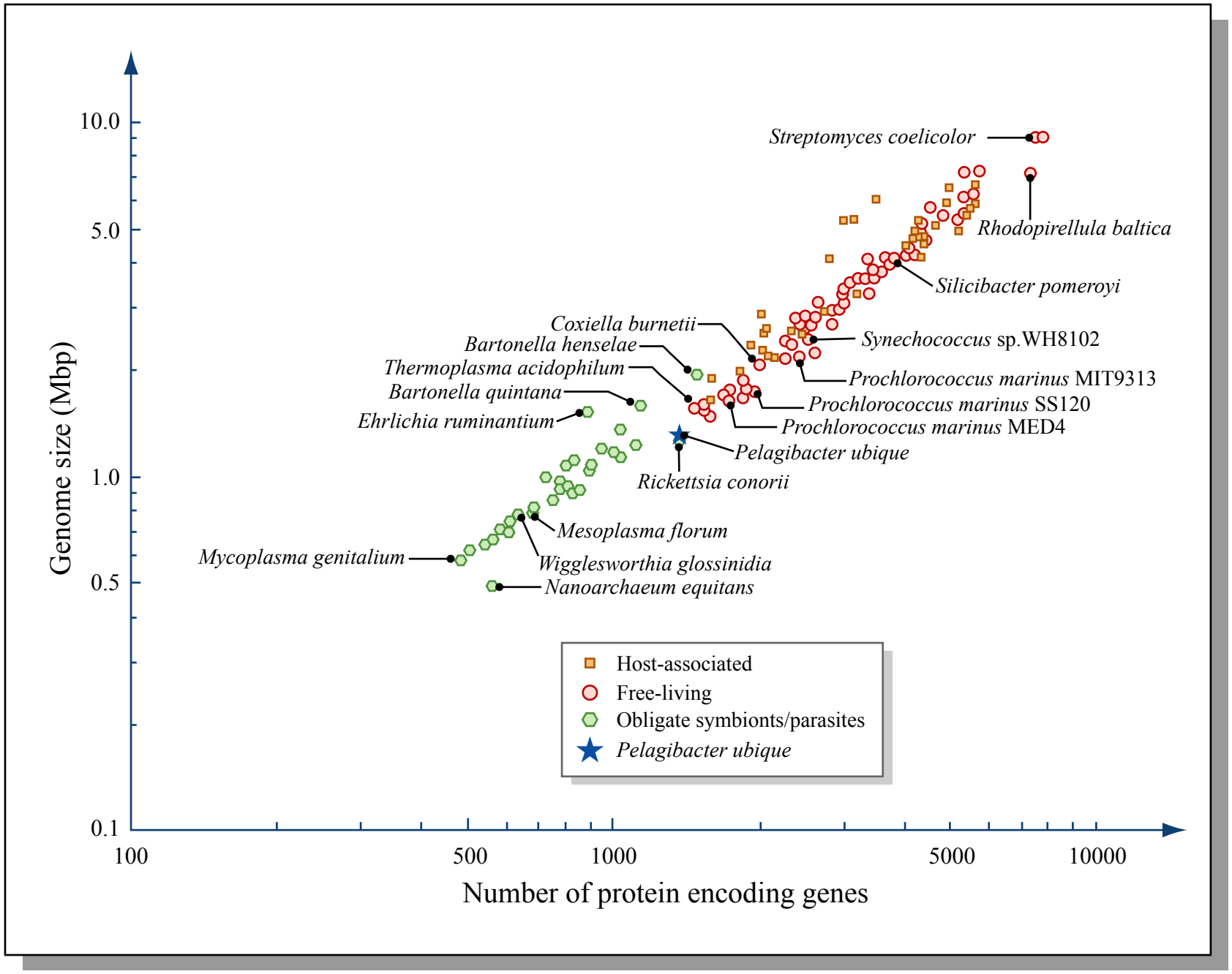


Figure by MIT OCW.

The 160-Kilobase Genome of the Bacterial Endosymbiont *Carsonella*

Atsushi Nakabachi,[1,2]* Atsushi Yamashita,[3]† Hidehiro Toh,[3,4]† Hajime Ishikawa,[5] Helen E. Dunbar,[2] Nancy A. Moran,[2] Masahira Hattori[4,7]*

Science 314:267
(Oct 13, 2006)

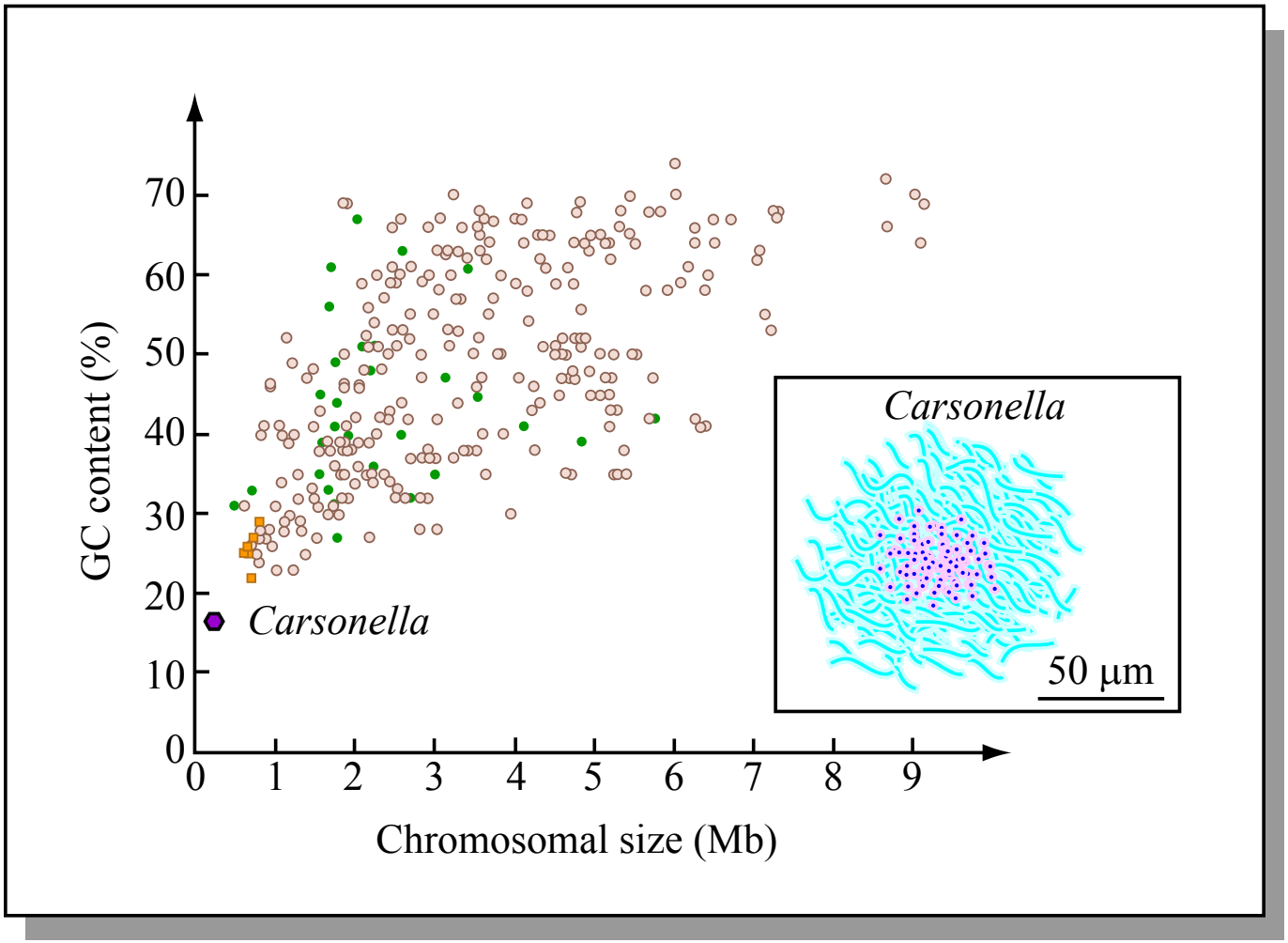


Figure by MIT OCW.