Sequencing and Microbes

Jonathan A. Eisen

Sequencing and Microbes

- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

Sequencing and Microbes

- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

Sequencing and Microbes

- <u>Part 1: Four Eras of Sequencing and Microbes</u>
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

NOTE - New Eras Add On to Past Ones, Past Ones Do Not End

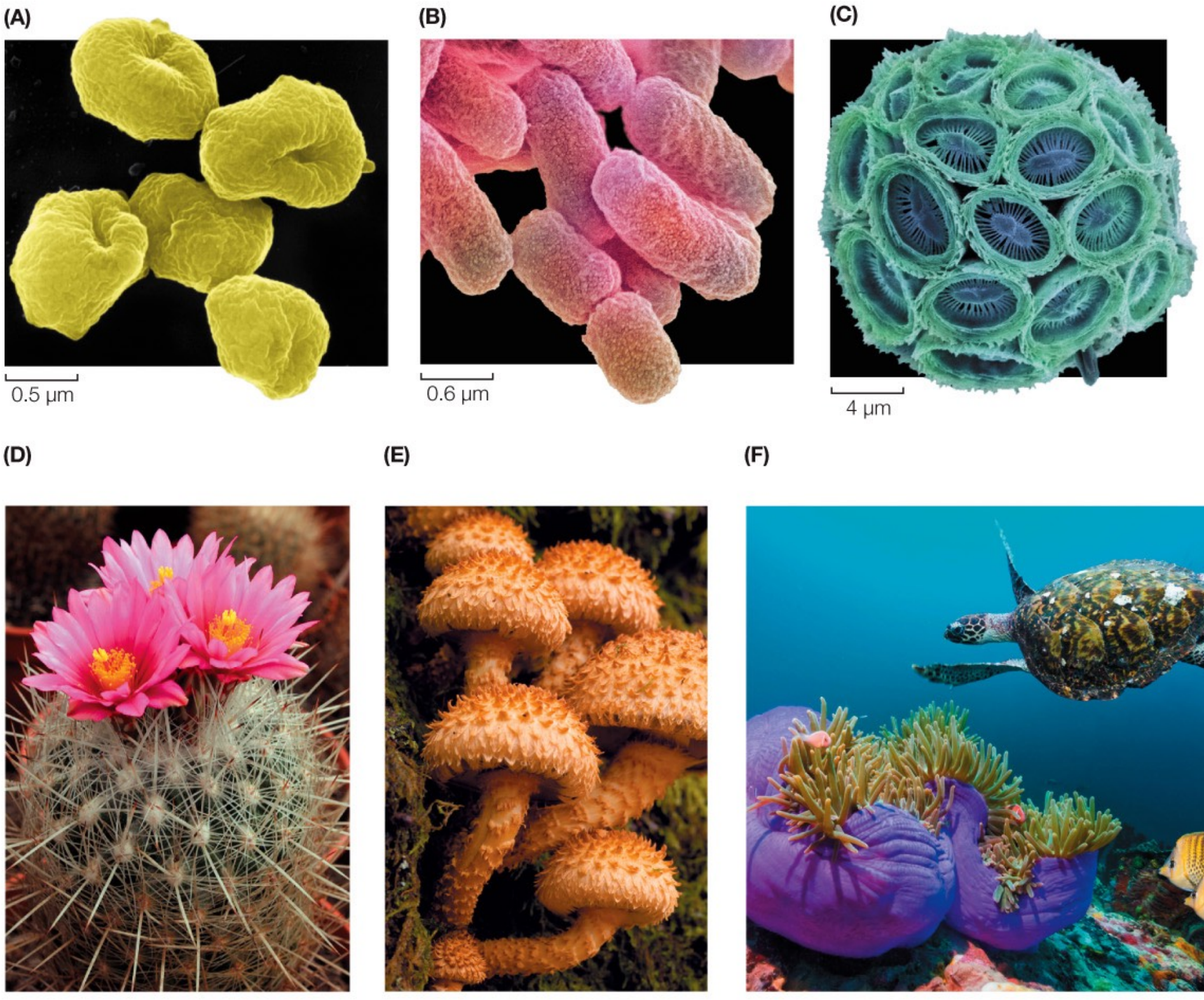
Sequencing and Microbes

- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

Carl Woese

http://mcb.illinois.edu/faculty/profile/1204

Diversity of Life of Earth



(a) © Eye of Science/SPL/Science Source; (b) © Science Photo Library RF/Photolibrary.com; (c) © Steve Gschmeissner/Science Source; (d) © Premaphotos/Alamy Stock Photo; (e) © GC Stock/Alamy Stock Photo; (f) © Steve Bloom Images/Alamy Stock Photo

LIFE: THE SCIENCE OF BIOLOGY 11e, Figure 1.1

Universal Traits

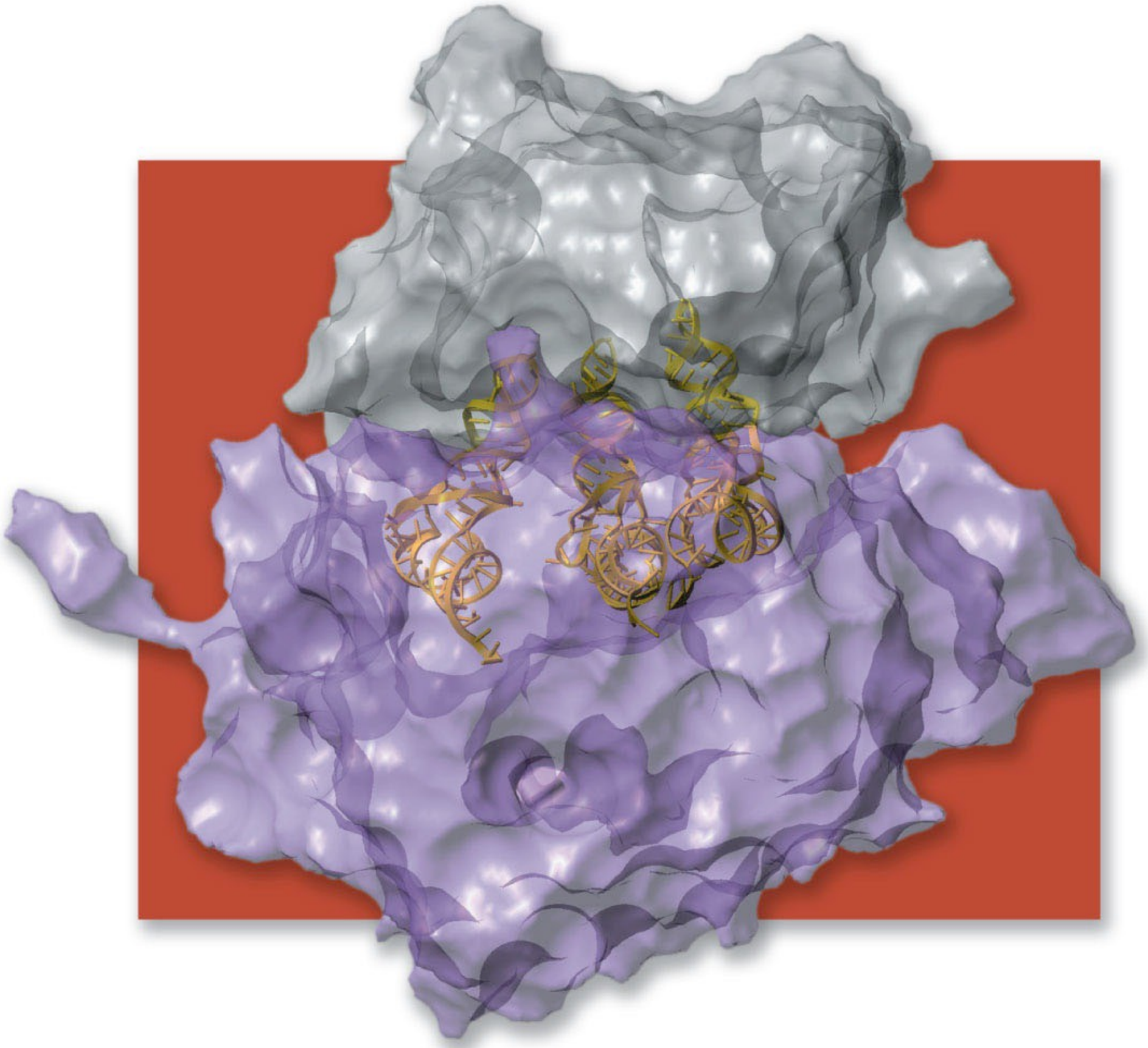
- ???
- Made up of cells
- Use of DNA as a genetic material
- Use of ACTG in DNA
- Use of ACUG in RNA
- Three letter genetic code
- Central dogma (DNA » RNA » protein)
- Use water as a solvent
- Lipoprotein cell envelope
- 20 core amino acids in proteins
- Lives on Earth
- Ribosome for translation
- RNA polymerase proteins
- Acquires energy from environment
- Store energy in chemicals

Can we use these to infer phylogenetic relationships?

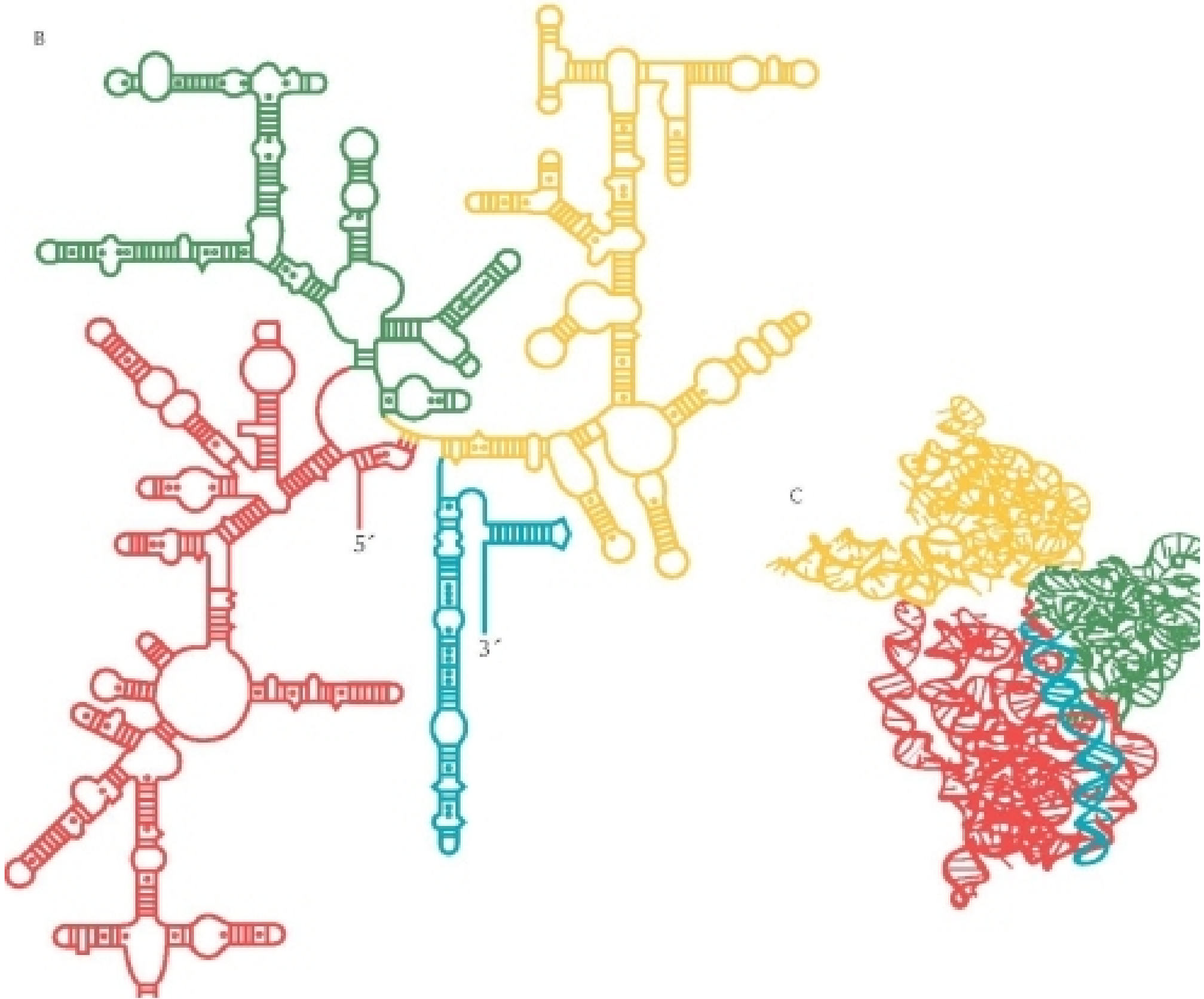
Not directly - because all organisms have these, presence is not informative

But … within some of these there is variation in structure

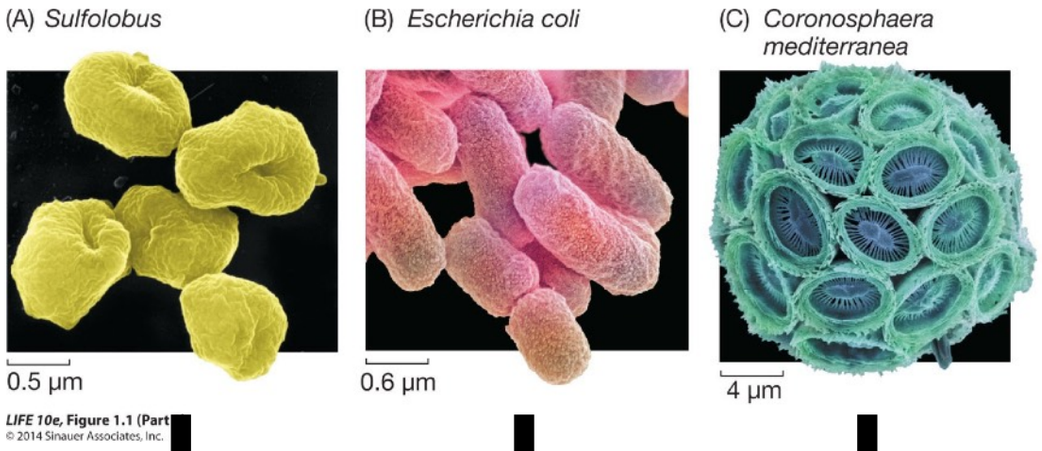
The Ribosome

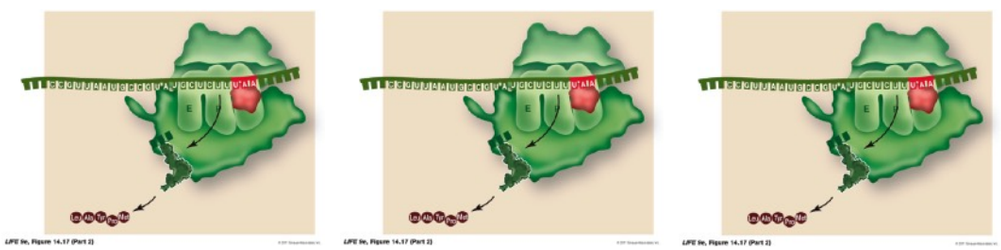


Ribosomal RNA structure



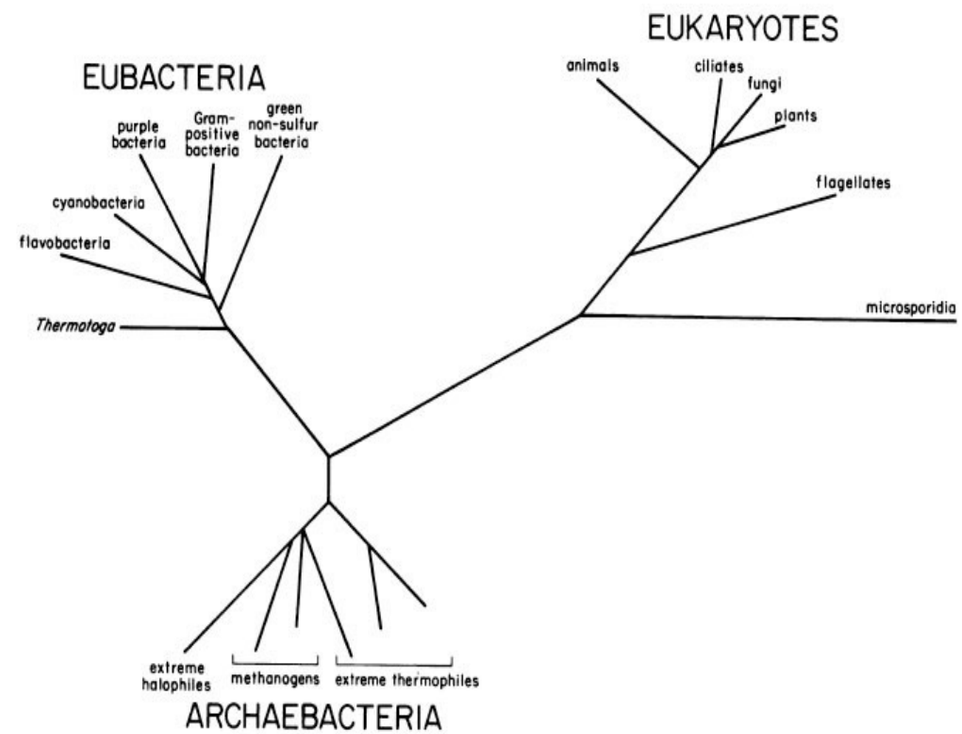
Woese Tree of Life



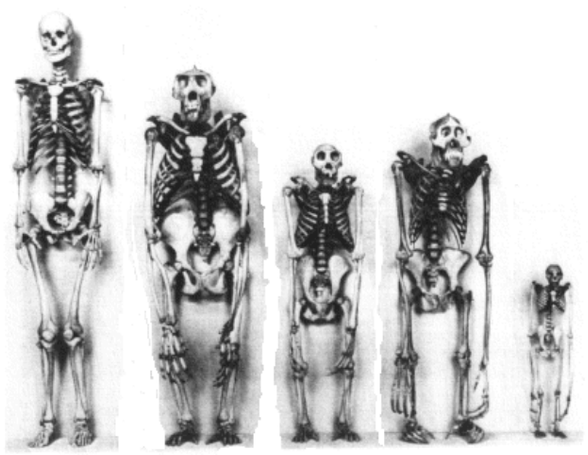


rRNA rRNA rRNA

ACUGC ACUCC ACCCC
ACCUAU AGCUAU AGCUCU
CGUUCG CGAUCG CGCUCG



Taxa	Characters	
S		G
R		G
E		G
F		G
C		G
W		G



Front Page of the New York Times, 11/3/1977

Scientists Discover a Form of Life That Predates Higher Organisms

By RICHARD D. LYONS

Special to The New York Times

URBANA, Ill., Nov. 2—Scientists studying the evolution of primitive organisms reported today the existence of a separate form of life that is hard to find in nature. They described it as a "third kingdom" of living material, composed of ancestral cells that abhor oxygen, digest carbon dioxide and produce methane.

The research group working here at the University of Illinois reported that this third form of life on earth was genetically distinct from the higher organisms that evolved from it—bacteria and, finally, the plant and animal world. Bacteria, with their own distinct form of cells, are more primitive than plant and animal life, which have vastly more complicated cellular structures.

Believed to have evolved 3.5 billion to 4 billion years ago, these organisms have yet to be named but are being referred to informally as either archaebacteria or methanogens. Before today's report, the oldest form of life, bacteria, was believed to have evolved about 3.4 billion years ago.

"We have shown that they are genetically distinct from the higher organisms," said Dr. Carl R. Woese, the leader of the group investigating the evolution of microorganisms.

The genetic tracking efforts of the scientific group, which spanned five years, were made public today by two of the Federal agencies that supported the research, the National Aeronautics and Space Administration and the National Science Foundation.

The work is described in detail in the October and November issues of the Proceedings of the National Academy of Sciences.

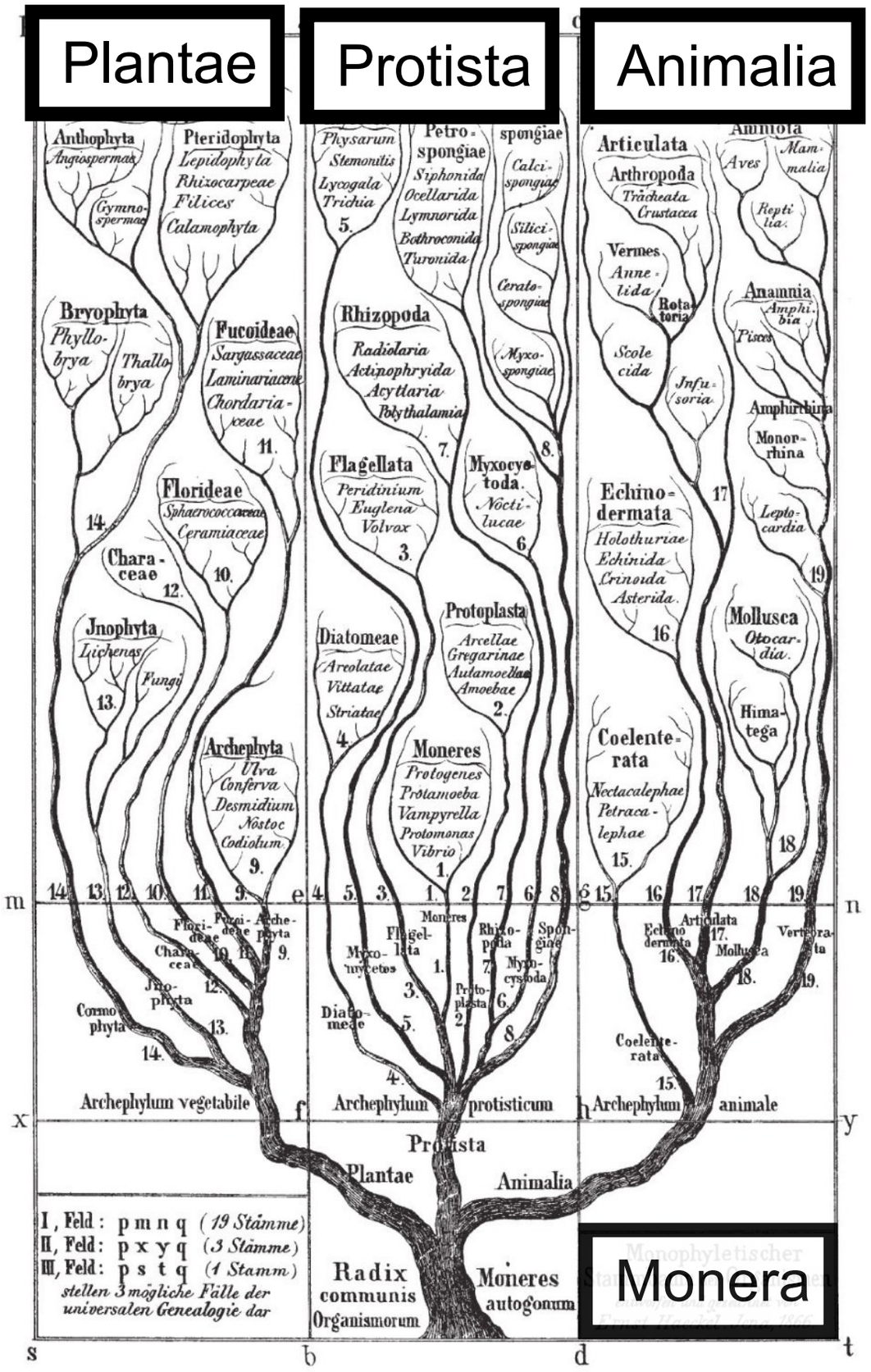
Asked for their evaluation of the results of the team at the University of Illinois, two other scientists familiar with the genetics of microbiology described the reports as "important" and "exciting," adding that it would further what is known of the basic processes of evolution.

Dr. Woese and his colleagues conclude that before the emergence on the earth of bacteria, usually regarded as the simplest form of life as we know it, at least one and perhaps several earlier forms of primitive organisms had evolved from the primordial ooze that developed after the

Continued on Page A20, Column 1

The New York Times
Published: November 3, 1977

Ernst Haeckel 1866



Early 1900s - Two Kingdoms

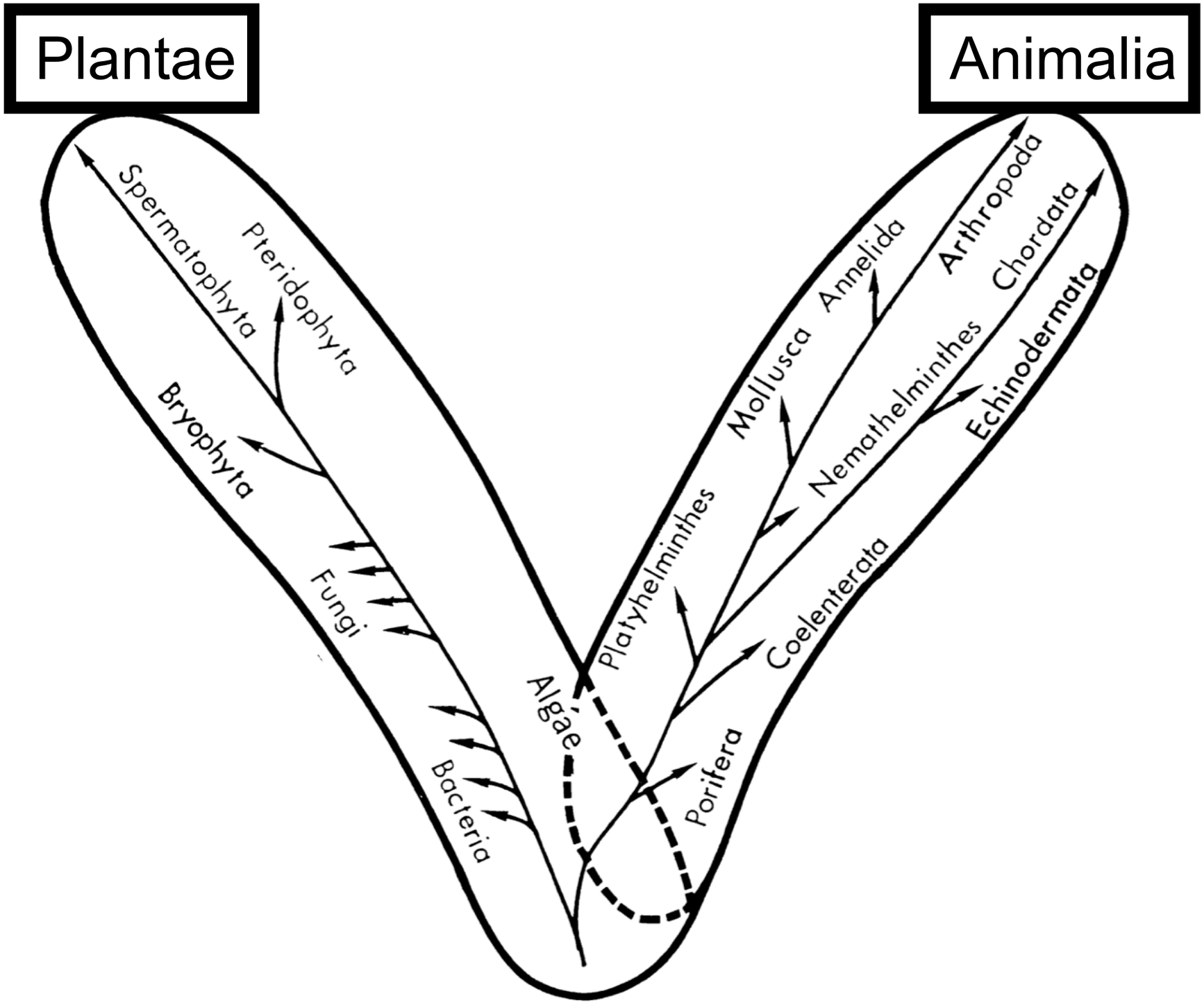
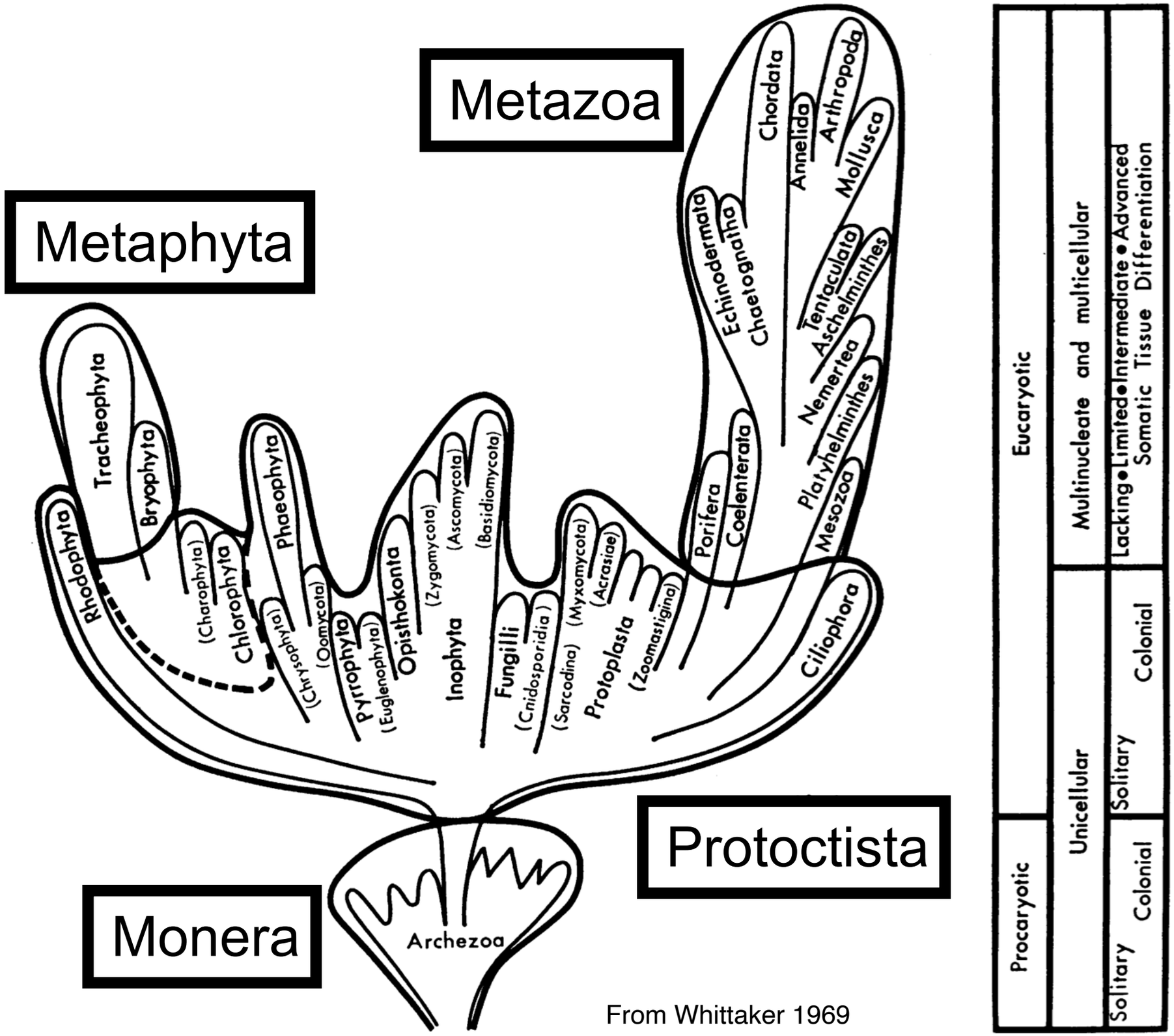


Fig. 1. A simplified evolutionary scheme of the two-kingdom system as it might have appeared early in the century. The plant kingdom comprised four divisions—Thallophyta (algae, bacteria, fungi), Bryophyta, Pteridophyta, and Spermatophyta. Only major animal phyla are indicated.

From Whittaker 1969

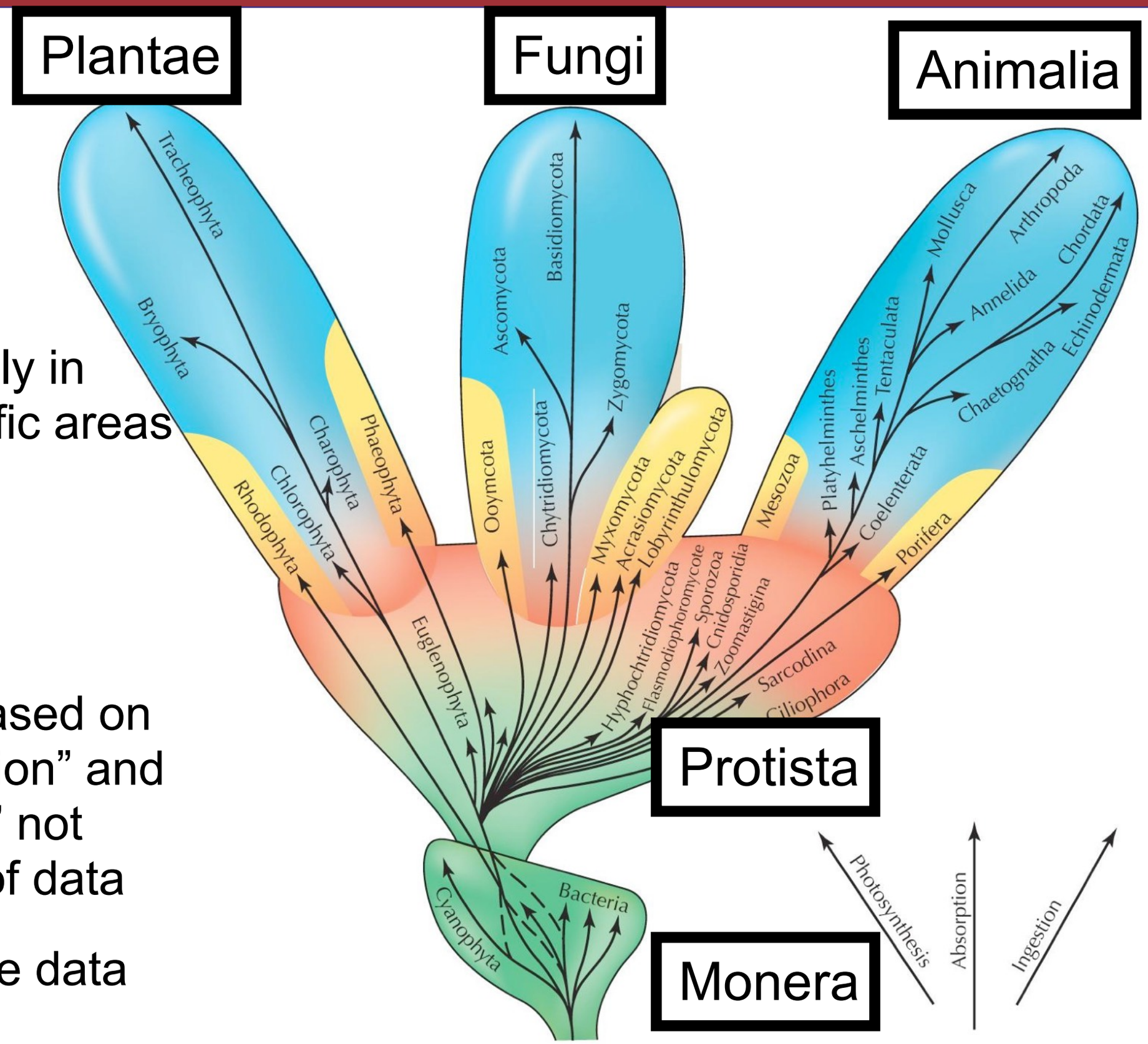
Copeland Four Kingdoms



From Whittaker 1969

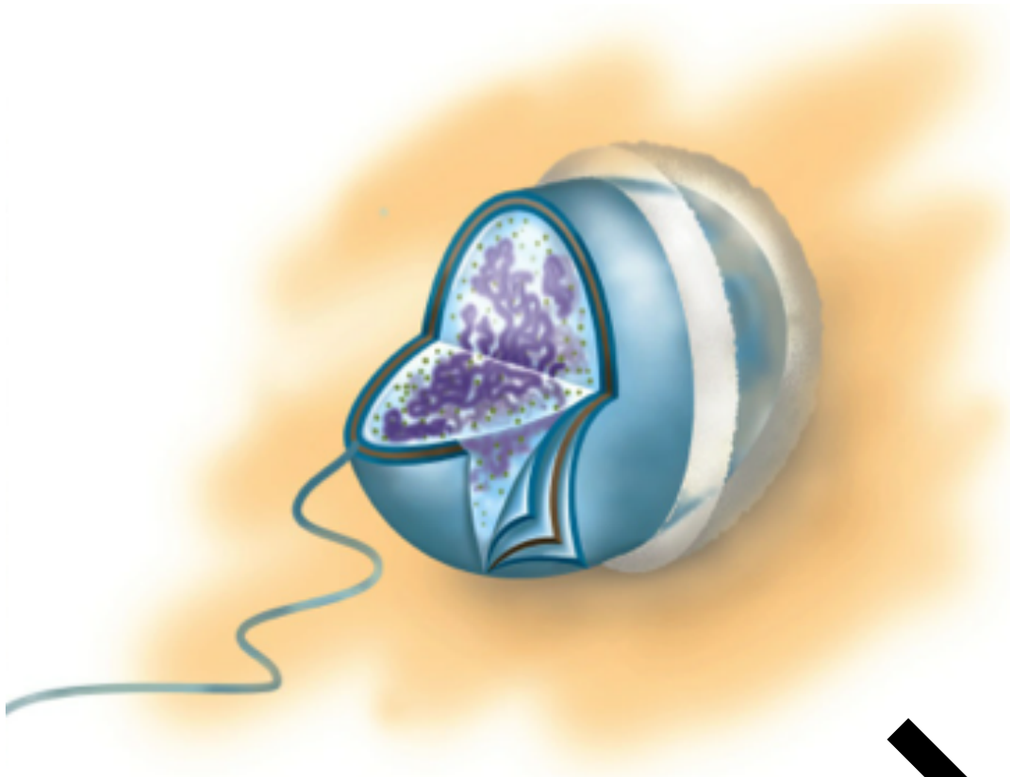
Whittaker – Five Kingdoms 1969

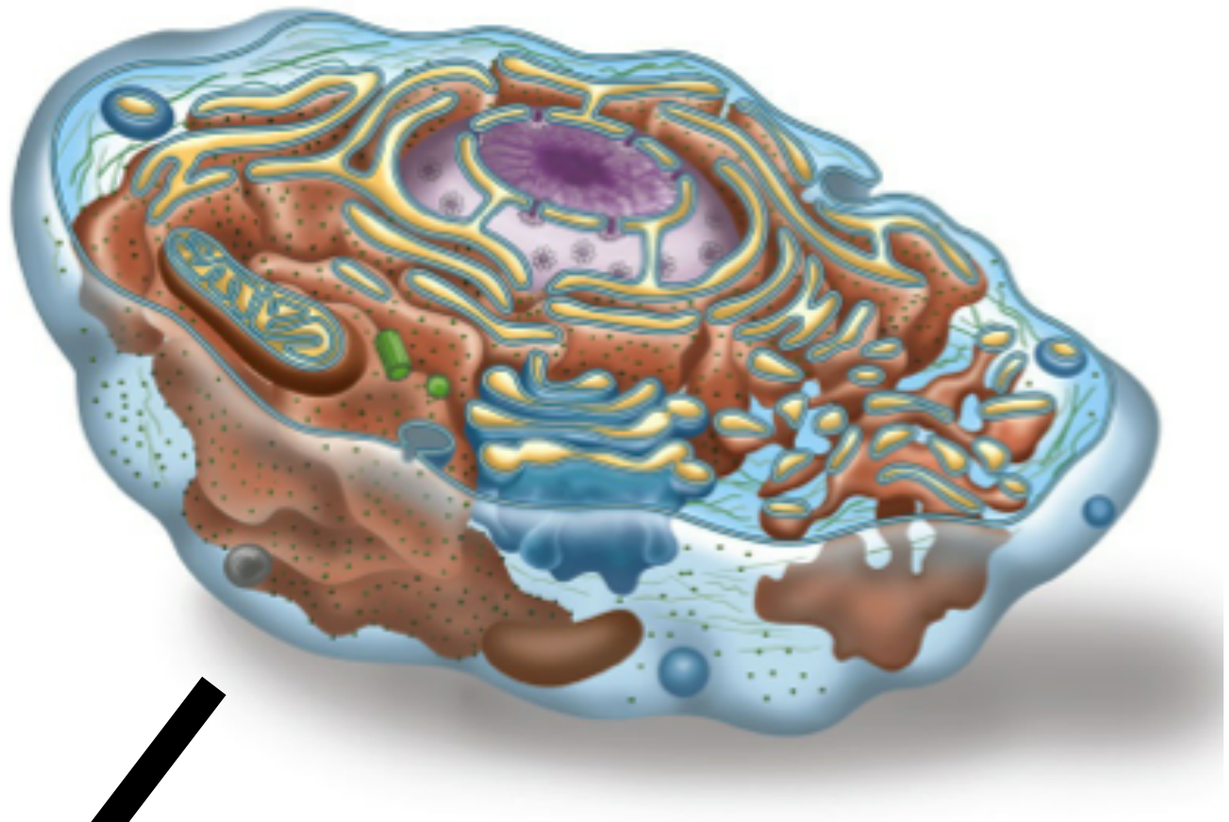
- Still used extensively in popular and scientific areas
- Improved upon two kingdoms and four kingdoms
- Overall structure based on "levels of organization" and "means of nutrition" not objective analysis of data
- What does objective data analysis tell us?

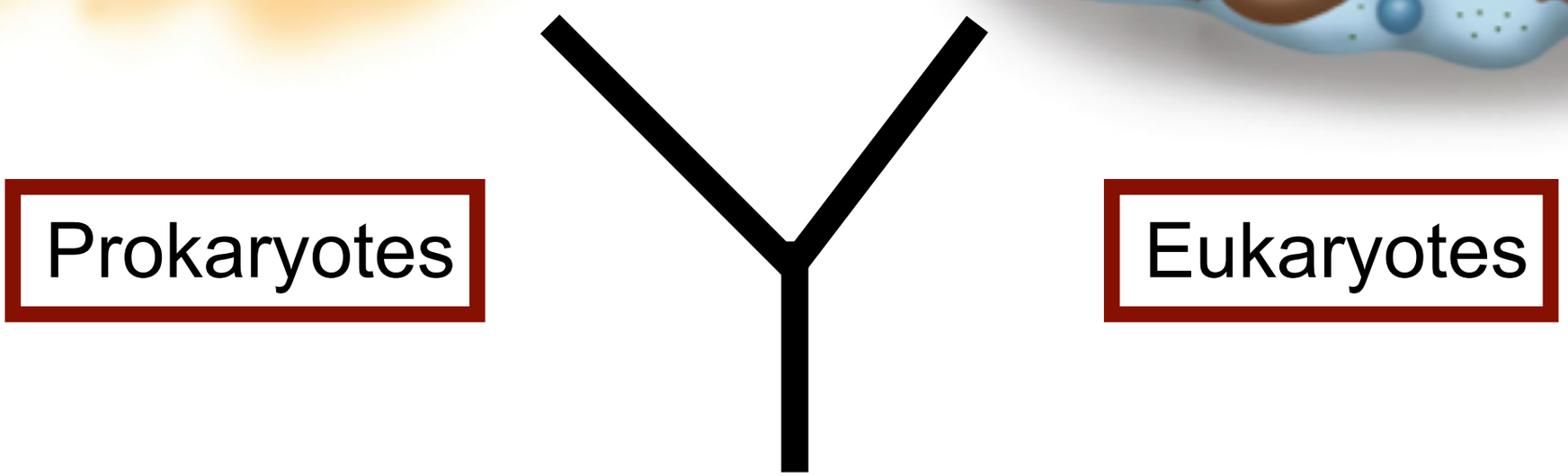


Prokaryotes vs. Eukaryotes (types of organisms)

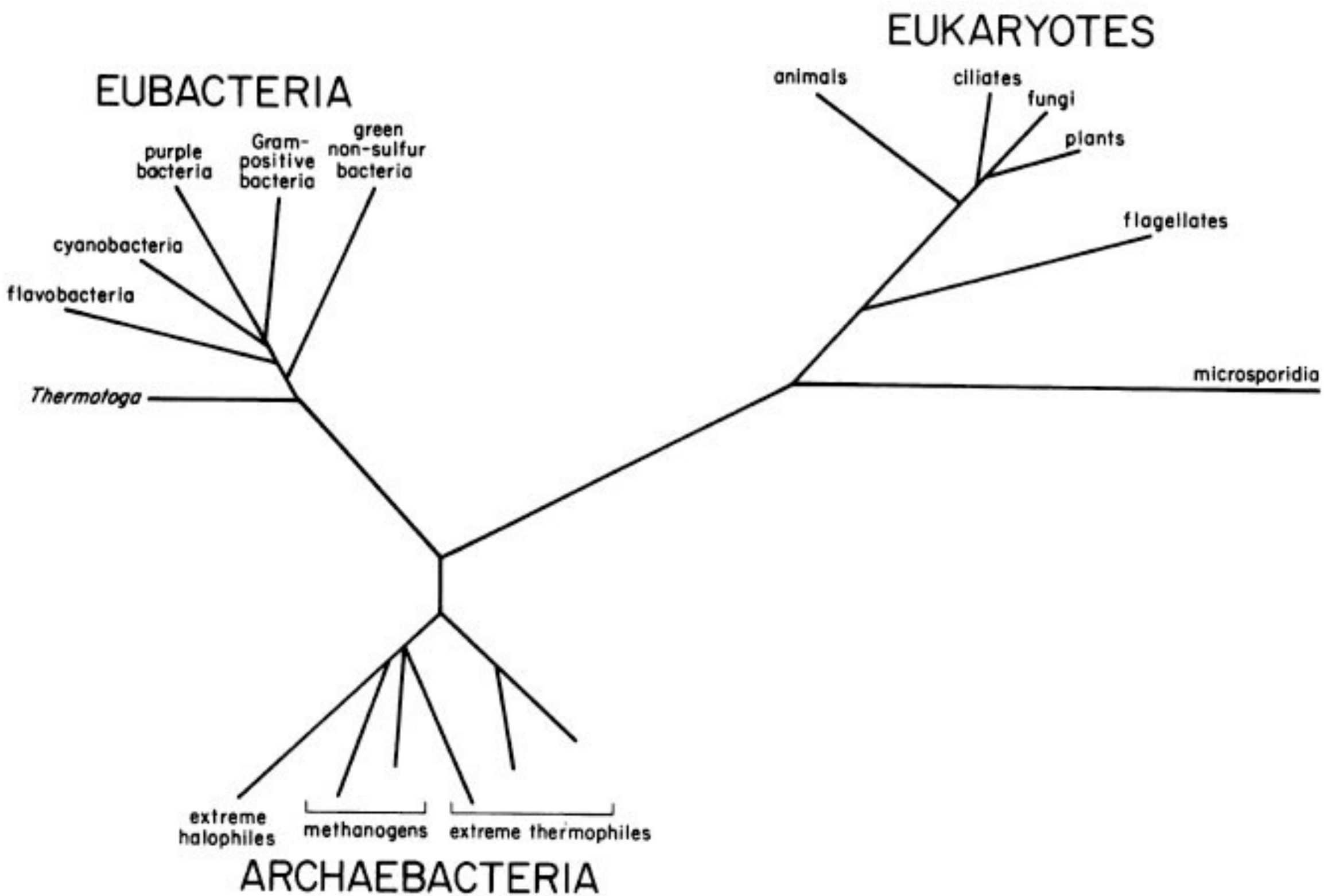
Interpreted As Implying This Tree (e.g. Chatton 1937)



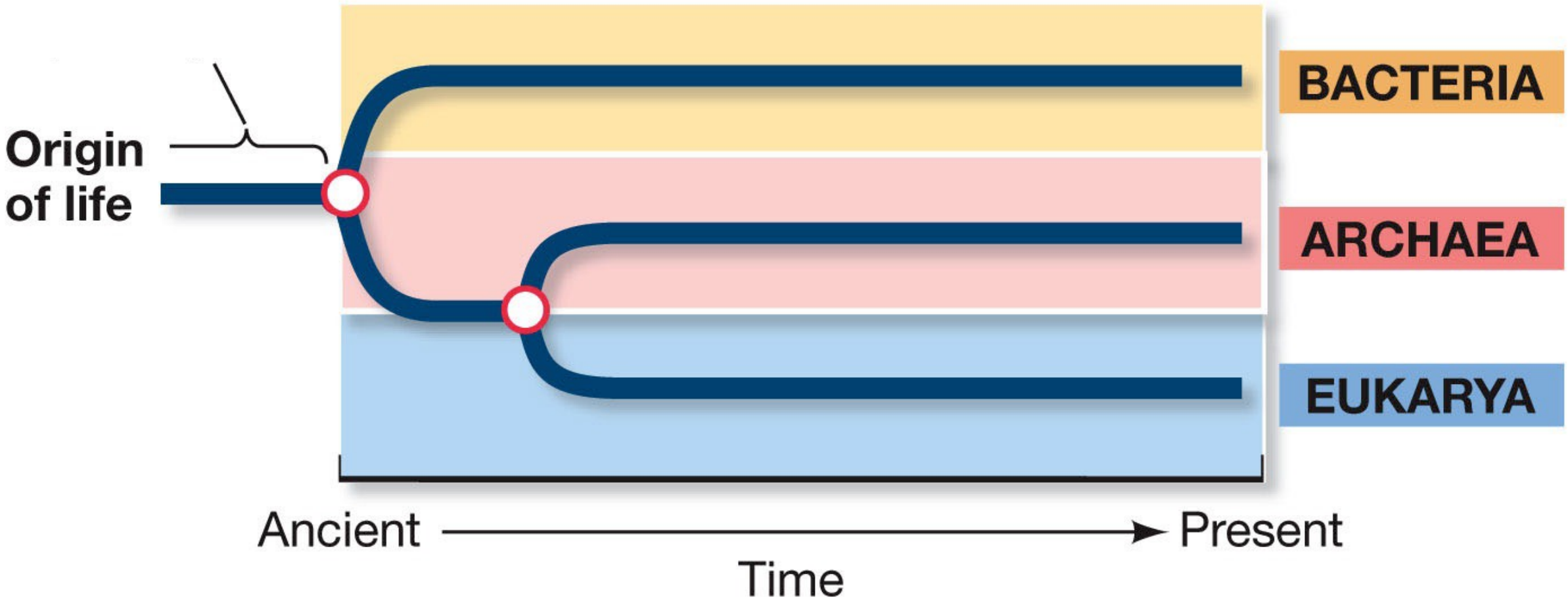




Woese 1987


EUKARYOTES
animals
ciliates
fungi
plants
flagellates
microsporidia
EUBACTERIA
purple bacteria
Gram-positive bacteria
green non-sulfur bacteria
cyanobacteria
flavobacteria
Thermotoga
extreme halophiles
methanogens
extreme thermophiles
ARCHAEBACTERIA

Simplified, Rooted Tree of Life



***LIFE 8e*, Figure 26.1**

Sequencing and Microbes

- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

Great Plate Count Anomaly

Great Plate Count Anomaly

Culturing

Microscopy

Great Plate Count Anomaly

Culturing

Count

Microscopy

Count

Great Plate Count Anomaly

Culturing

Microscopy

Count <<<< Count

Great Plate Count Anomaly

Problem because appearance not effective for “who is out there?” or “what are they doing?”

Culturing

Count

<<<<

Microscopy

Count

Great Plate Count Anomaly

Problem because appearance not effective for “who is out there?” or “what are they doing?”

Solution?

Culturing

Microscopy

Count

<<<<

Count

Great Plate Count Anomaly

Problem because appearance not effective for “who is out there?” or “what are they doing?”

Solution?

→ DNA

Culturing

↓

Count

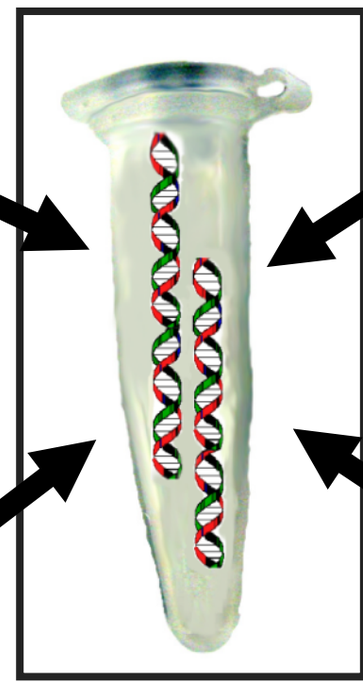
<<<<

Microscopy

↓

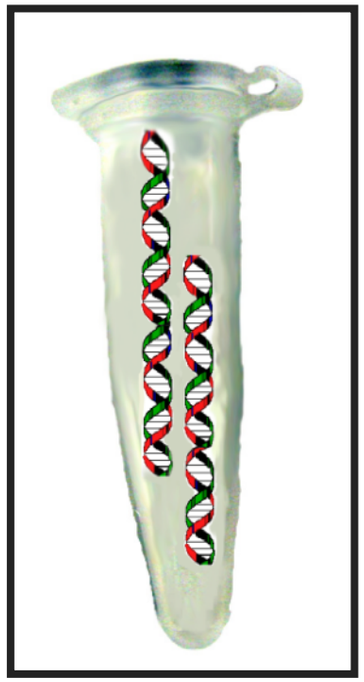
Count

Analysis of uncultured microbes



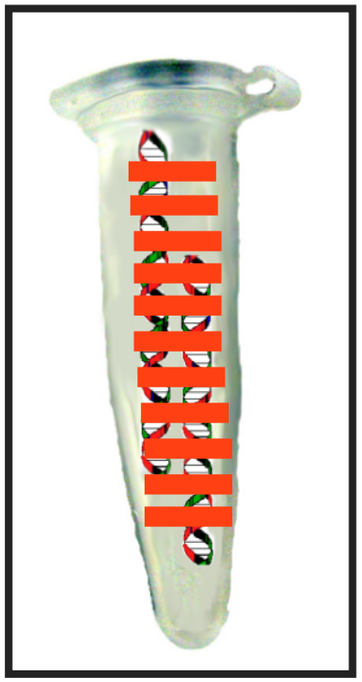
PCR and phylogenetic analysis of rRNA genes

DNA extraction



PCR

Makes lots of copies of the rRNA genes in sample



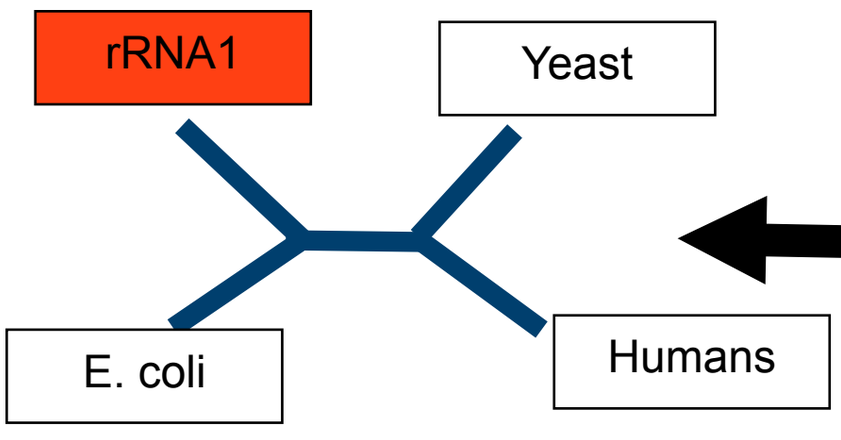
Sequence rRNA genes

rRNA1
5' ...TACAGTATAGGT
GGAGCTAGCGATCG
ATCGA... 3'

Sequence alignment = Data matrix

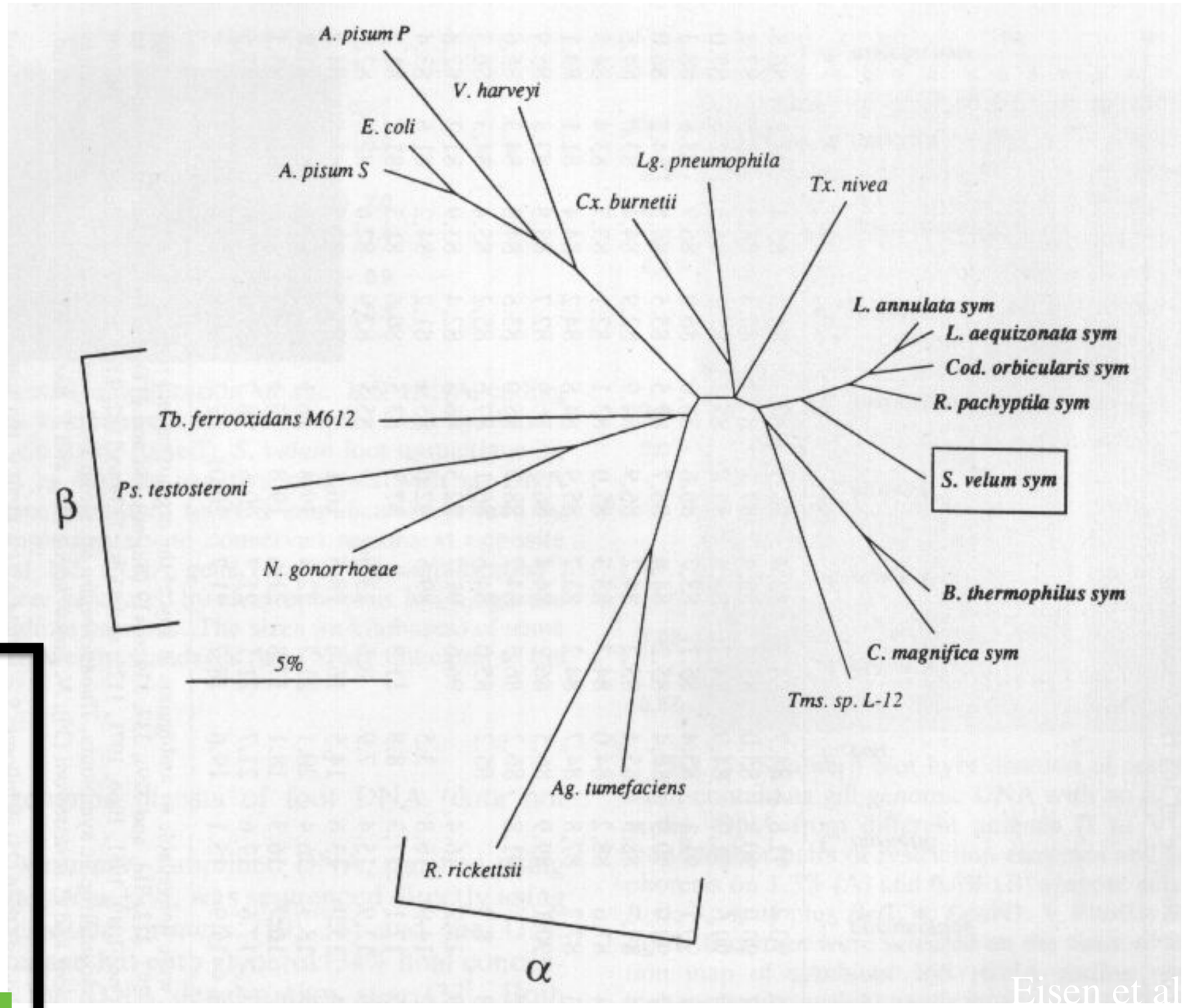
rRNA1	A	C	A	C	A	C
Yeast	T	A	C	A	G	T
E. coli	A	G	A	C	A	G
Humans	T	A	T	A	G	T

Phylogenetic tree



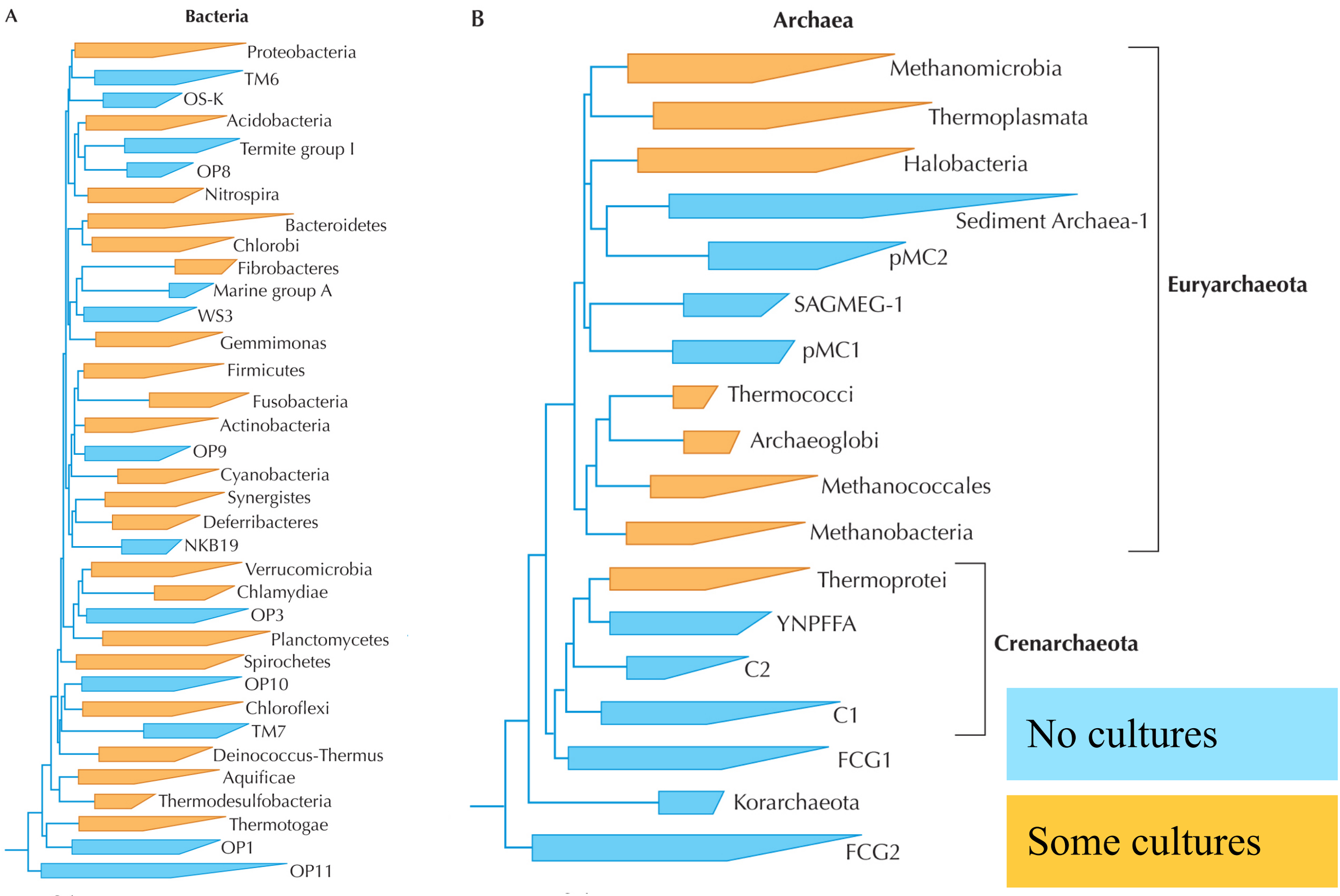
rRNA Phylotyping: One Taxon

Colleen Cavanaugh

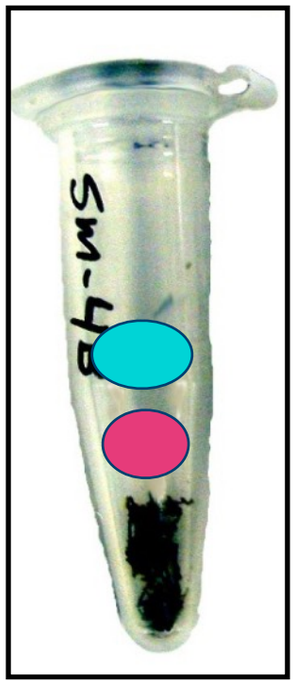


Eisen et al. 1992. J. Bact.174: 3416

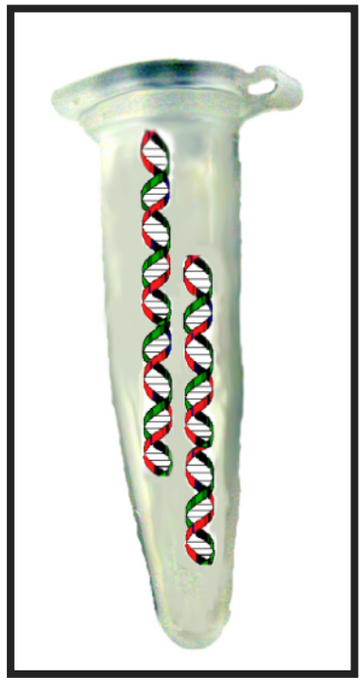
Major phyla of bacteria & archaea (as of 2002)



PCR and phylogenetic analysis of rRNA genes

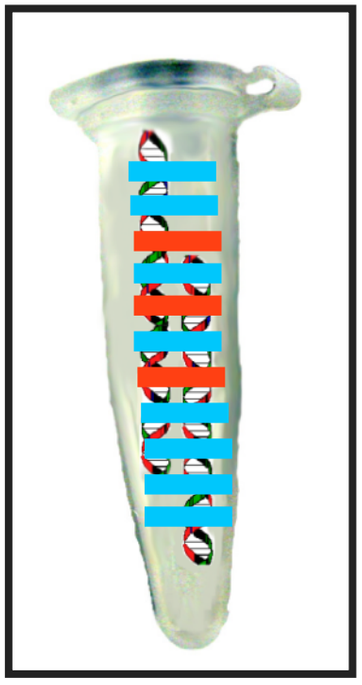


DNA extraction



PCR

Makes lots of copies of the rRNA genes in sample



Sequence rRNA genes

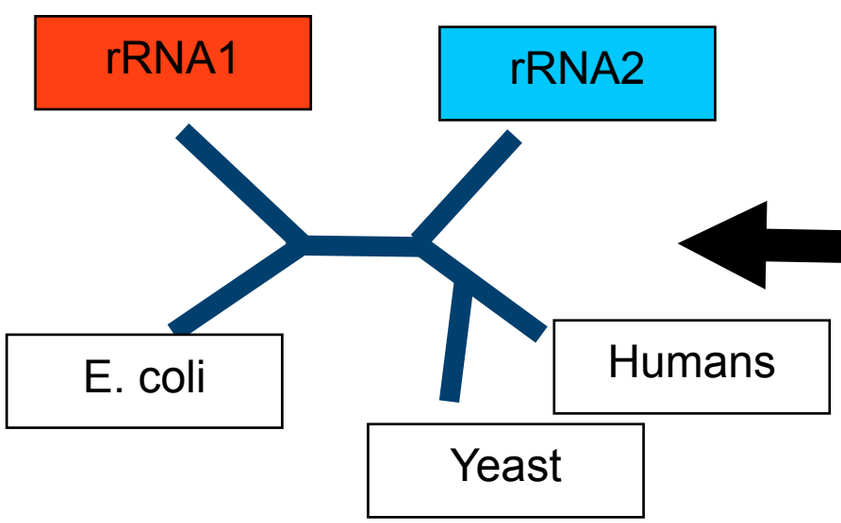
rRNA1
5' ...ACACACATAGGT
GGAGCTAGCGATCG
ATCGA... 3'

rRNA2
5' ...TACAGTATAGGT
GGAGCTAGCGATCG
ATCGA... 3'

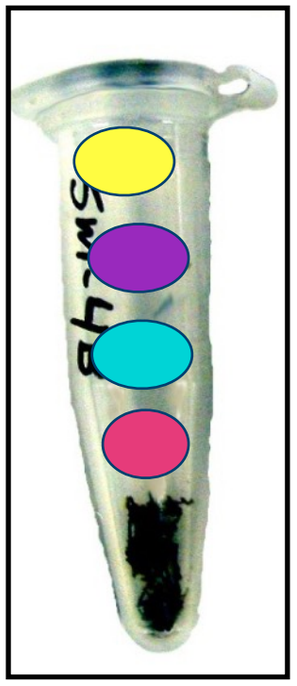
Sequence alignment = Data matrix

rRNA1	A	C	A	C	A	C
rRNA2	T	A	C	A	G	T
E. coli	A	G	A	C	A	G
Humans	T	A	T	A	G	T
Yeast	T	A	C	A	G	T

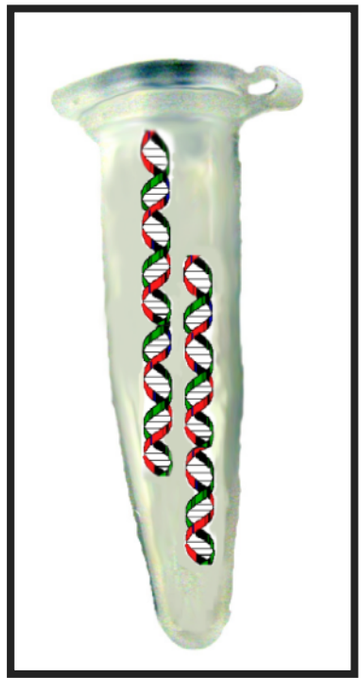
Phylogenetic tree



PCR and phylogenetic analysis of rRNA genes

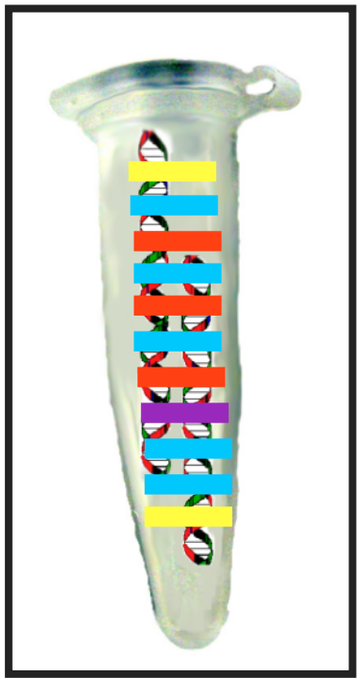


DNA extraction



PCR

Makes lots of copies of the rRNA genes in sample



Sequence rRNA genes

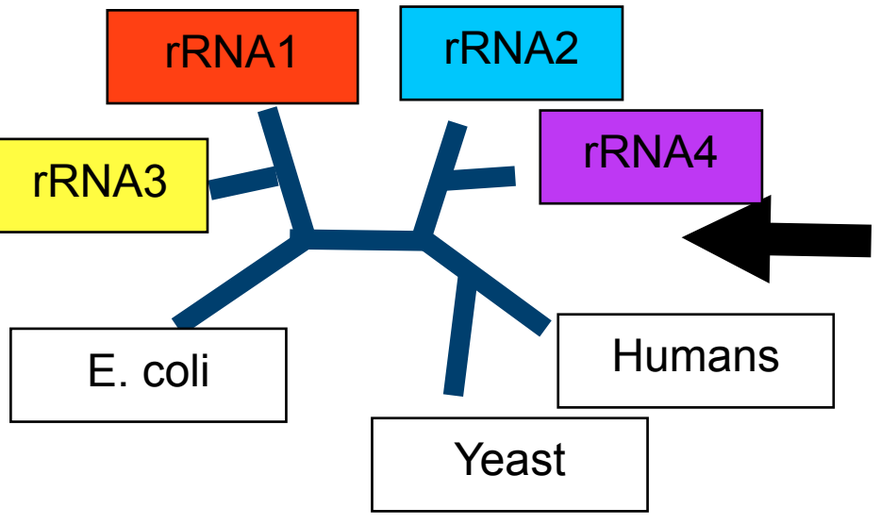
rRNA1
5'...ACACACATAGGTGGAGCTA
GCGATCGATCGA... 3'

rRNA2
5'..TACAGTATAGGTGGAGCTAG
CGACGATCGA... 3'

rRNA3
5'...ACGGCAAAATAGGTGGATT
CTAGCGATATAGA... 3'

rRNA4
5'...ACGGCCCGATAGGTGGATT
CTAGCGCCATAGA... 3'

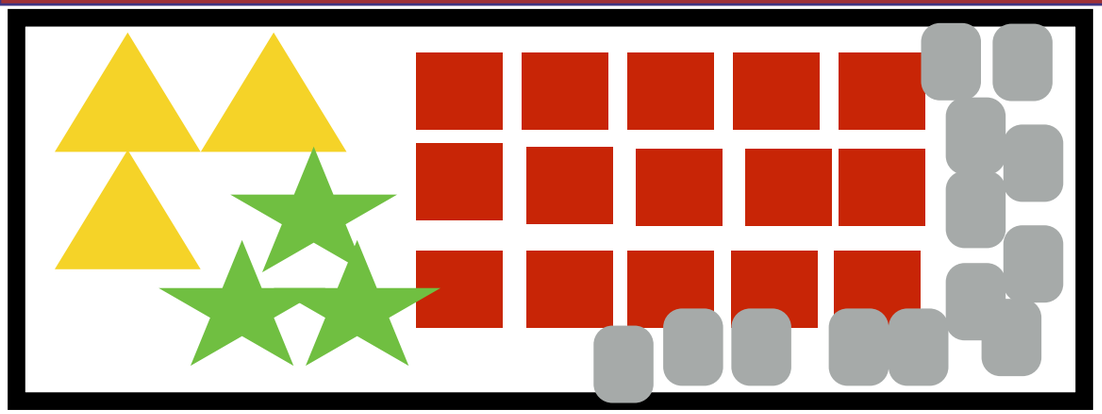
Phylogenetic tree

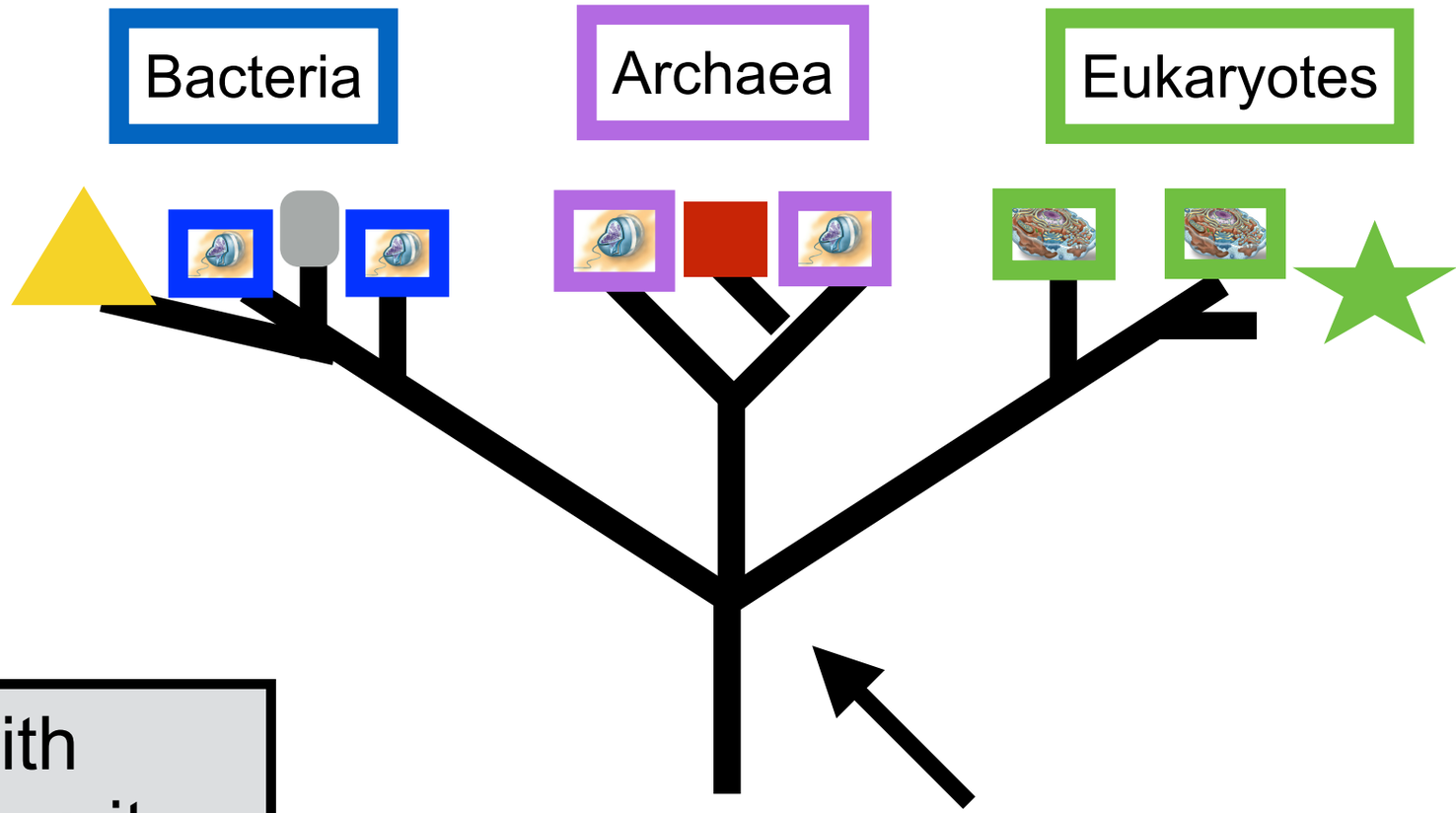


Sequence alignment = Data matrix

rRNA1	A	C	A	C	A	C
rRNA2	T	A	C	A	G	T
rRNA3	C	A	C	T	G	T
rRNA4	C	A	C	A	G	T
E. coli	A	G	A	C	A	G
Humans	T	A	T	A	G	T
Yeast	T	A	C	A	G	T

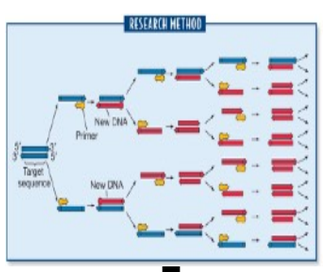
rRNA Phylotyping: Relative Abundance





DNA

Even with more taxa it still works



ACTGC
ACCTAT
CGTTCG

AGGGG
AGCTCT
CGCTCG

ACTGC
ACCTAT
CGTTCG

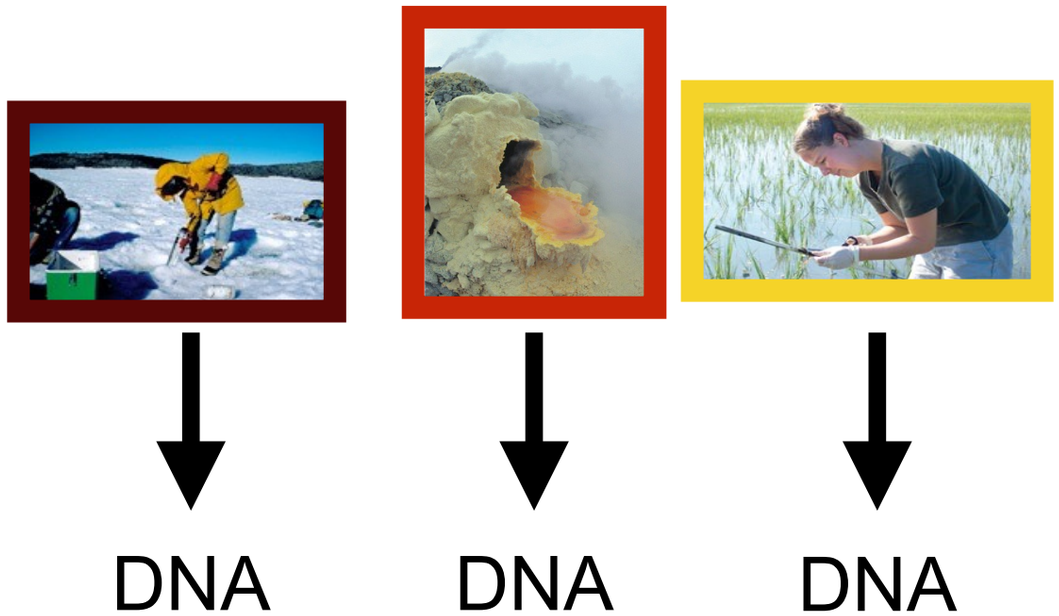
ACTCC
AGCTAT
CGATCG

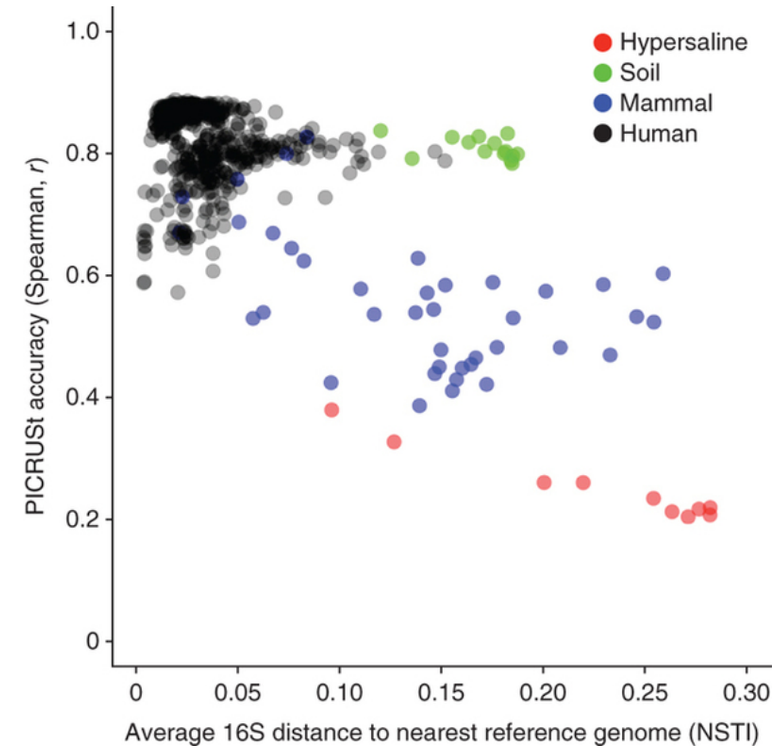
ACCCC
AGCTCT
CGCTCG

AGGGG
AGCTCT
CGCTCG

Taxa	Characters
B1	ACTGCACCTATCGTTCG
B2	ACTCCACCTATCGTTCG
E1	ACTCCAGCTATCGATCG
E2	ACTCCAGGTATCGATCG
A1	ACCCCAGCTCTCGCTCG
A2	ACCCCAGCTCTGGCTCG
New1	ACCCCAGCTCTGCCTCG
New2	AGGGGAGCTCTGCCTCG
New3	ACTCCAGCTATCGATCG
New4	ACTGCACCTATCGTTCG

rRNA PCR: Community Comparisons





Taxa	Characters
B1	ACTGCACCTATCGTTCG
B2	ACTCCACCTATCGTTCG
E1	ACTCCAGCTATCGATCG
E2	ACTCCAGGTATCGATCG
A1	ACCCCAGCTCTCGCTCG
A2	ACCCCAGCTCTGGCTCG
New1	ACCCCAGCTCTGCCTCG
New2	ACGGCAGCTCTGCCTCG

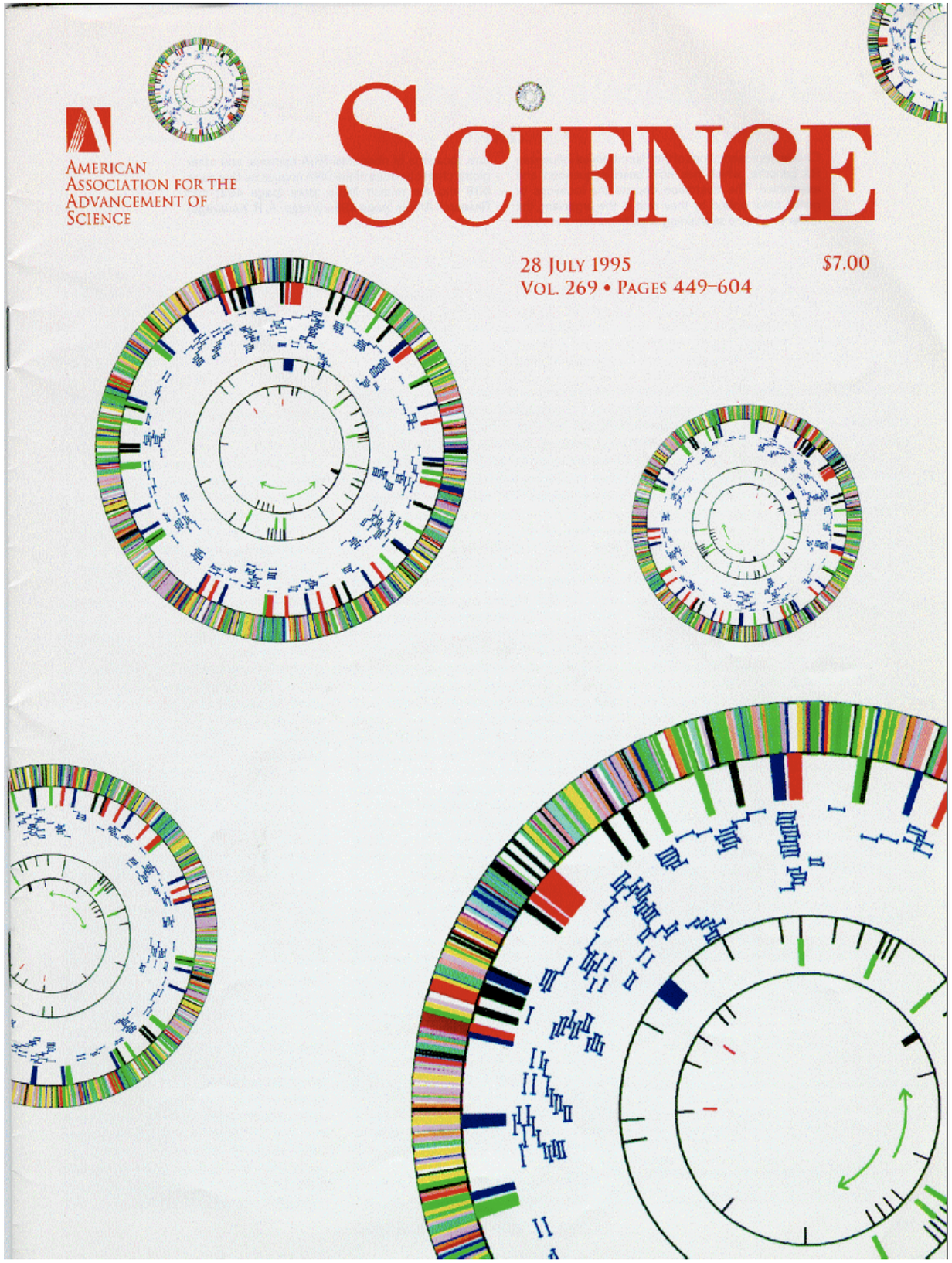
Limitations of rRNA PCR Surveys

- PCR primer bias
- Taxa without rRNA
- Copy number issues
- rRNA phylogeny imperfect
- rRNA evolves too slowly
- Relative abundances usually

Sequencing and Microbes

- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

1st Genome Sequence

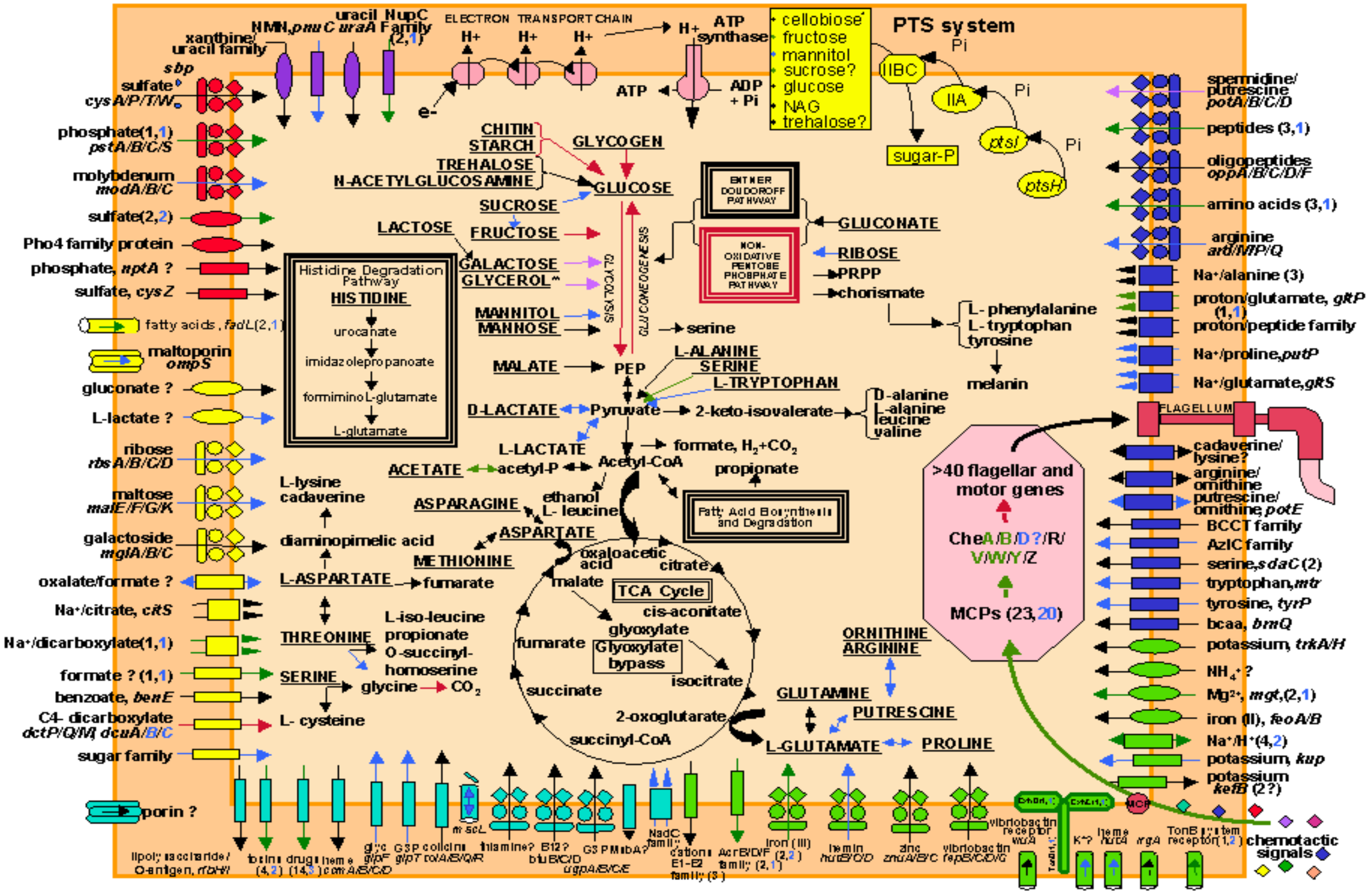


Fleischmann et al. 1995

Methods

- Random (shotgun) sequencing
- Assembly into contigs and scaffolds
- Finishing gaps (not done as much these days)
- Annotation I: Finding genes
- Annotation II: Predicting gene function
- Comparative genomics
- Phylogenomics

Metabolic Predictions



Sequencing and Microbes

- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

rRNA PCR: Community Comparisons

DNA DNA DNA

Genomes of Uncultured Taxa

AKA Metagenomics

Delong Lab - Sequencing Large Inserts

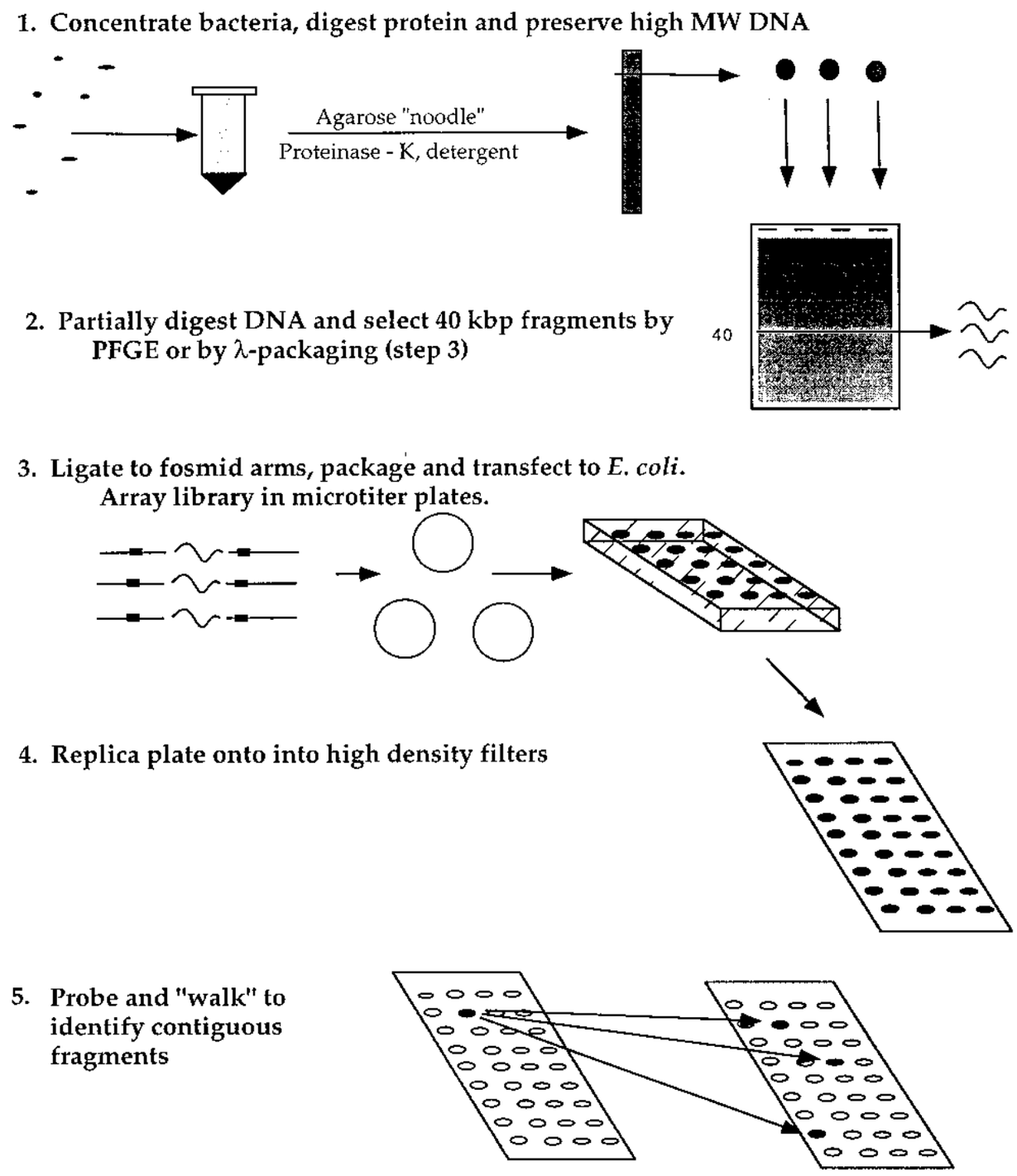


FIG. 1. Flowchart depicting the construction and screening of an environmental library from a mixed picoplankton sample. MW, molecular weight; PFGE, pulsed-field gel electrophoresis.

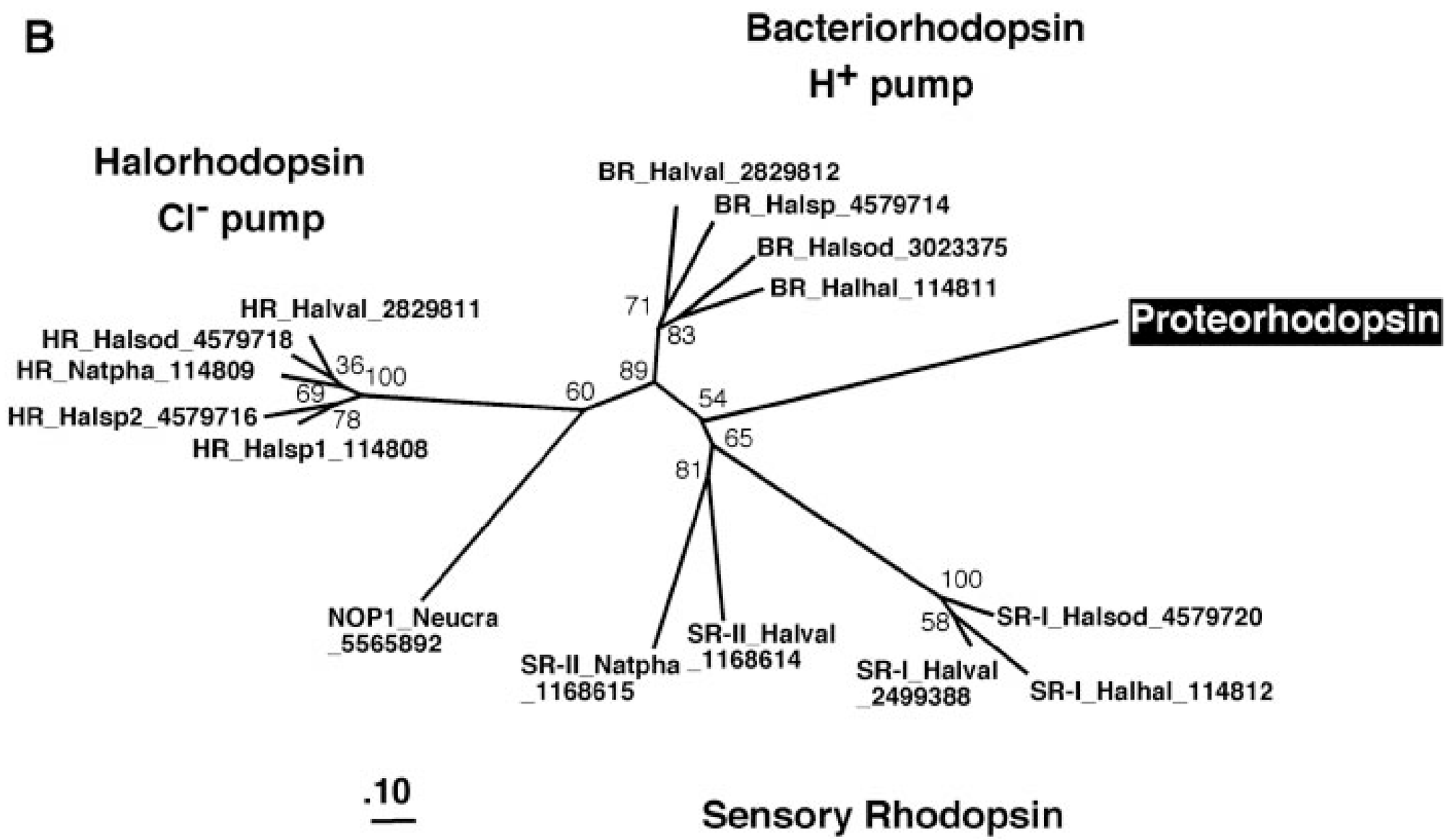
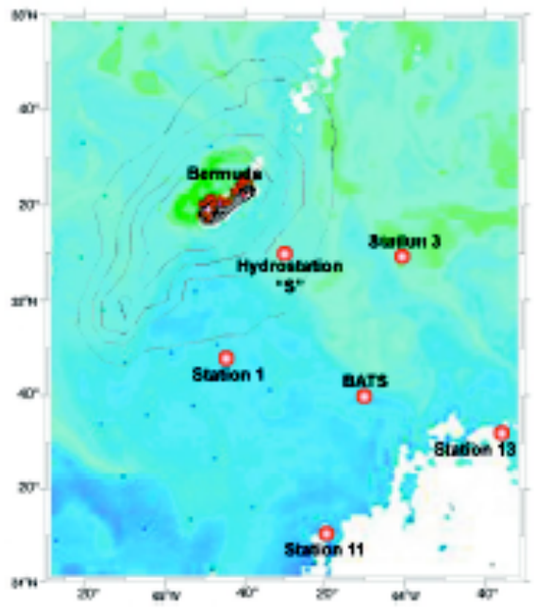
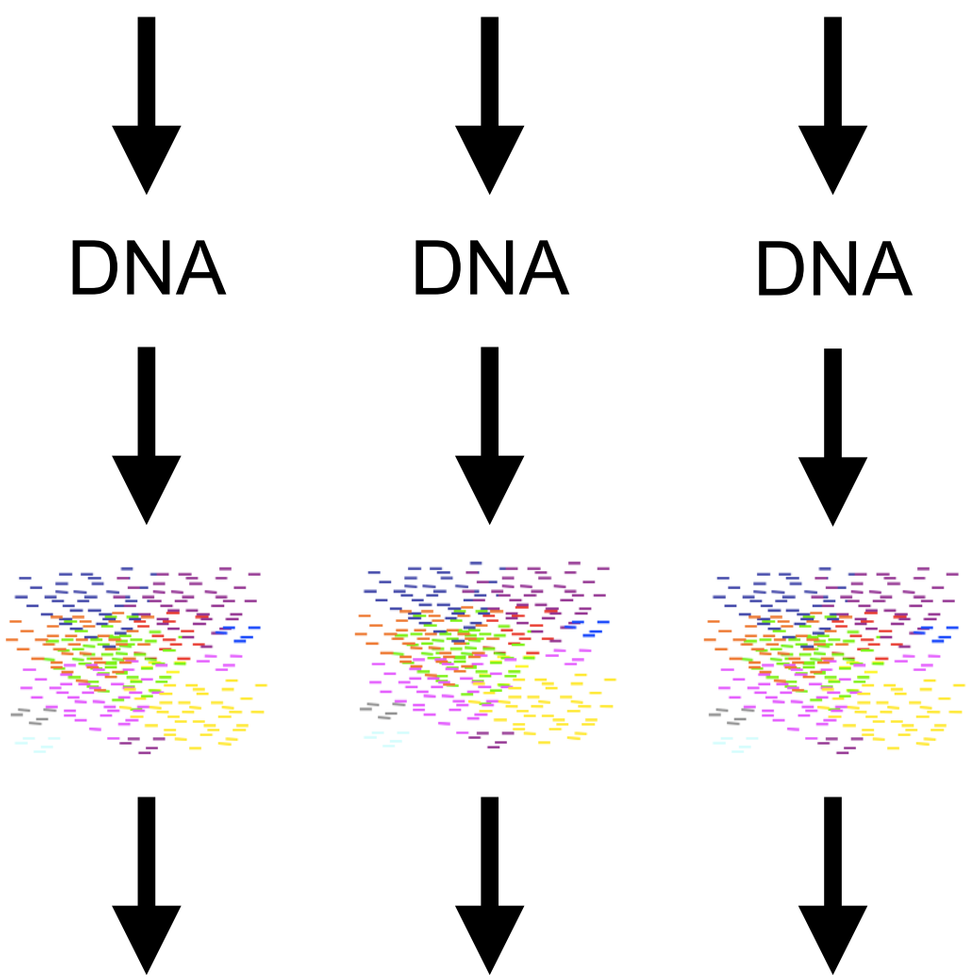


Fig. 1. (**A**) Phylogenetic tree of bacterial 16*S* rRNA gene sequences, including that encoded on the 130-kb bacterioplankton BAC clone (EBAC31A08) (*16*). (**B**) Phylogenetic analysis of proteorhodopsin with archaeal (BR, HR, and SR prefixes) and *Neurospora crassa* (NOP1 prefix) rhodopsins (*16*). Nomenclature: Name_Species.abbreviation_Genbank.gi (HR, halorhodopsin; SR, sensory rhodopsin; BR, bacteriorhodopsin). Halsod, *Halorubrum sodomense*; Halhal, *Halobacterium salinarum* (*halobium*); Halval, *Haloarcula vallismortis*; Natpha, *Natronomonas pharaonis*; Halsp, *Halobacterium* sp; Neucra, *Neurospora crassa*.

Shotgun “Metagenomics” - 2004





RESEARCH ARTICLE

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1*] Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

Community structure and metabolism through reconstruction of microbial genomes from the environment

Gene W. Tyson[1], Jarrod Chapman[3,4], Philip Hugenholtz[1], Eric E. Allen[1], Rachna J. Ram[1], Paul M. Richardson[4], Victor V. Solovyev[4], Edward M. Rubin[4], Daniel S. Rokhsar[3,4] & Jillian F. Banfield[1,2]

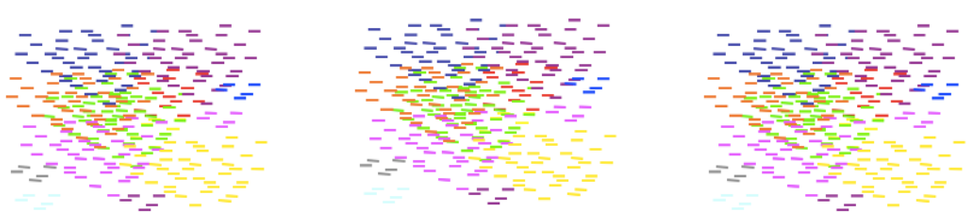
[1]*Department of Environmental Science, Policy and Management,* [2]*Department of Earth and Planetary Sciences, and* [3]*Department of Physics, University of California, Berkeley, California 94720, USA*
[4]*Joint Genome Institute, Walnut Creek, California 94598, USA*

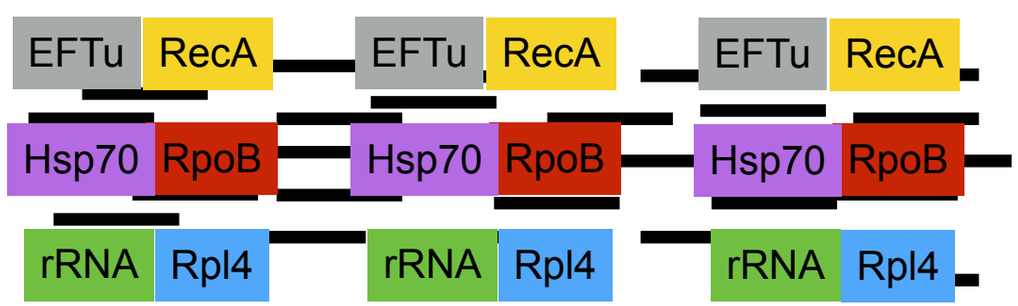
Side Lesson:

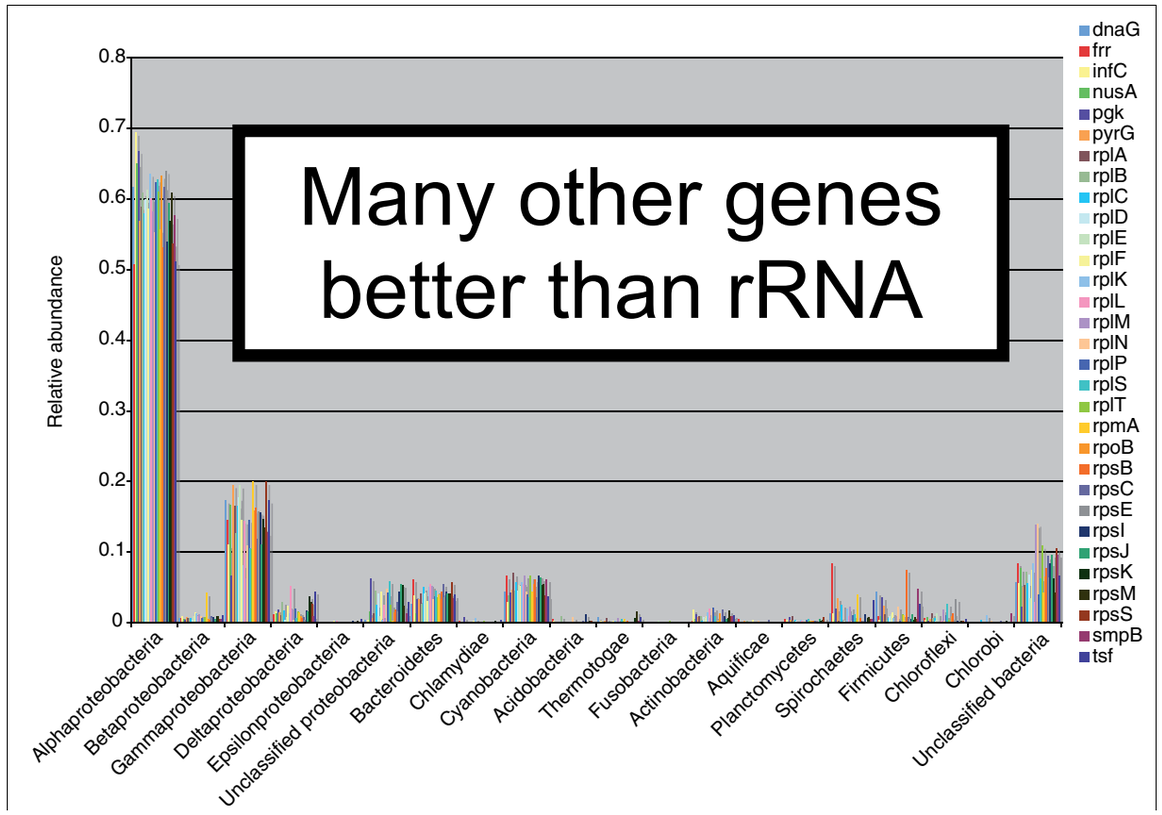
Other genes can be better for phylotyping than rRNA

Shotgun "Metagenomics"

DNA DNA DNA

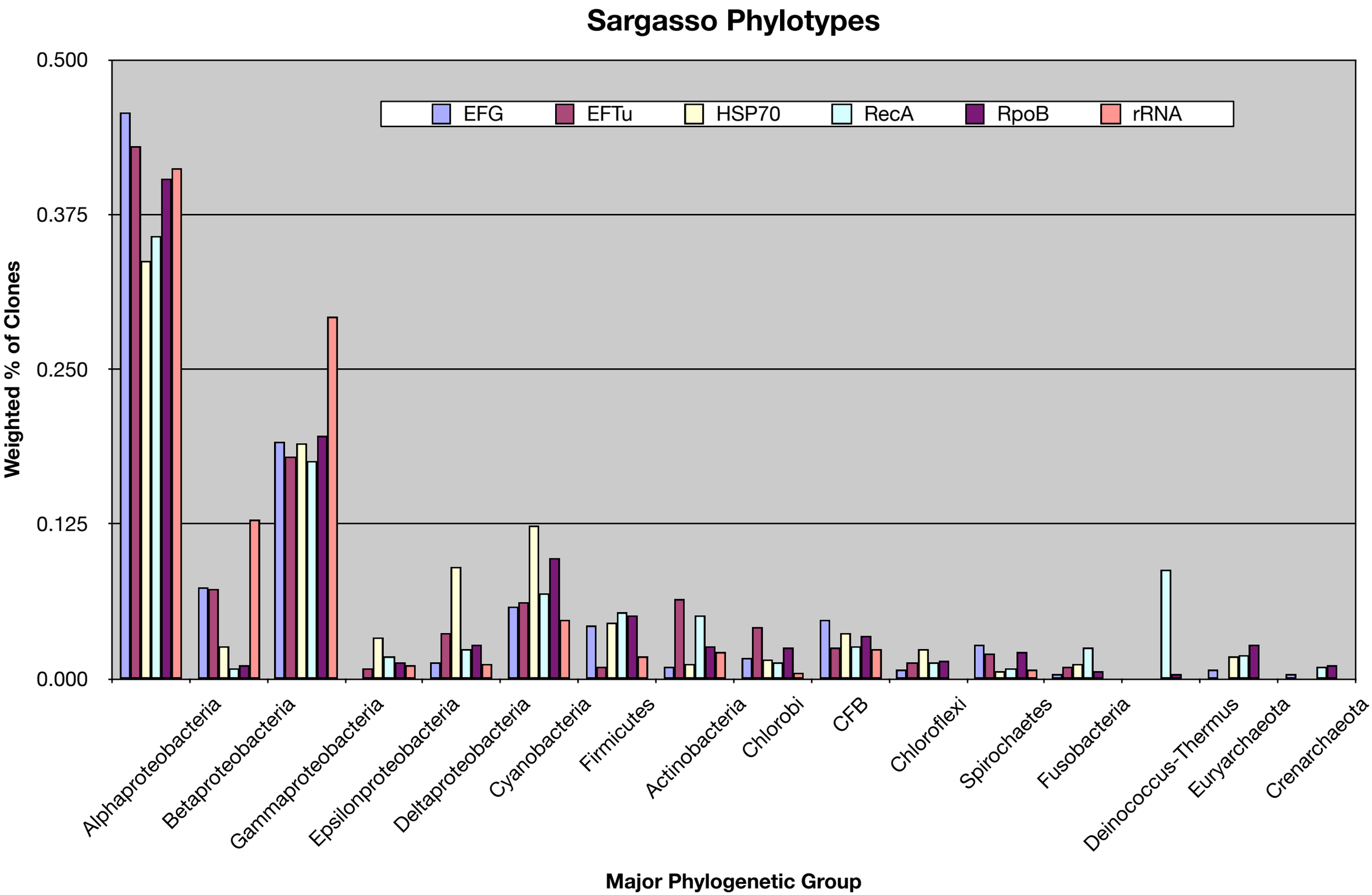






Taxa	Characters
B1	ACTGCACCTATCGTTCG
B2	ACTCCACCTATCGTTCG
E1	ACTCCAGCTATCGATCG
E2	ACTCCAGGTATCGATCG
A1	ACCCCAGCTCTCGCTCG
A2	ACCCCAGCTCTGGCTCG
New1	ACCCCAGCTCTGCCTCG
New2	AGGGGAGCTCTGCCTCG
New3	ACTCCAGCTATCGATCG
New4	ACTGCACCTATCGTTCG

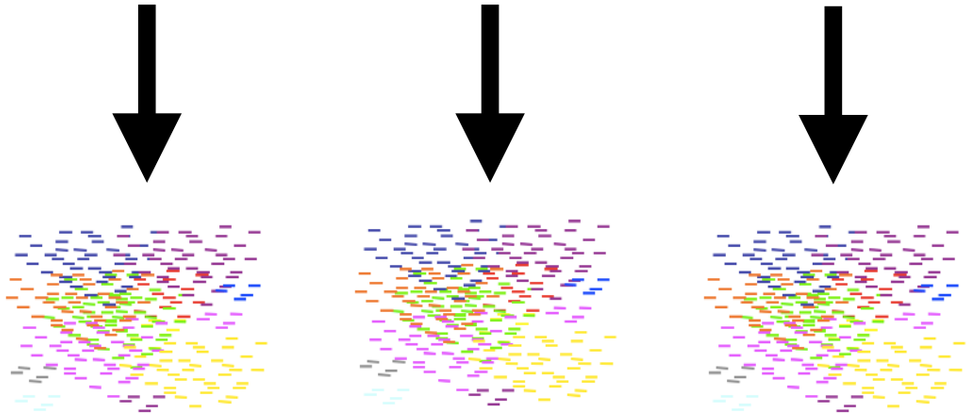
Sargasso Sea Five Other Markers

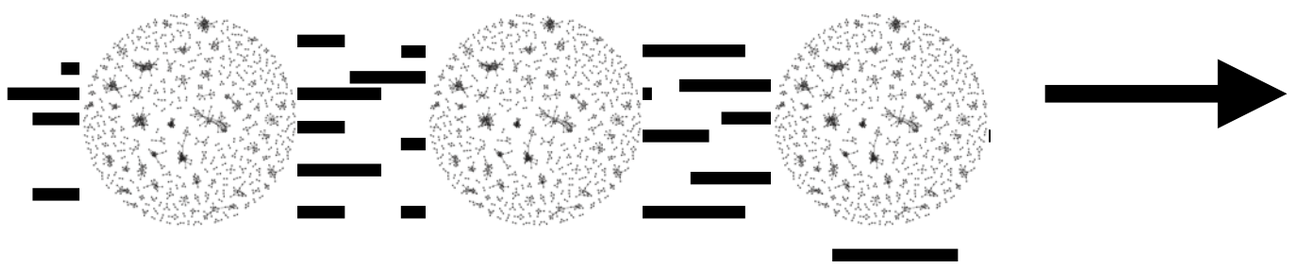


Venter et al., Science 304: 66. 2004

Shotgun Metagenomics - Functional Predictions

DNA DNA DNA





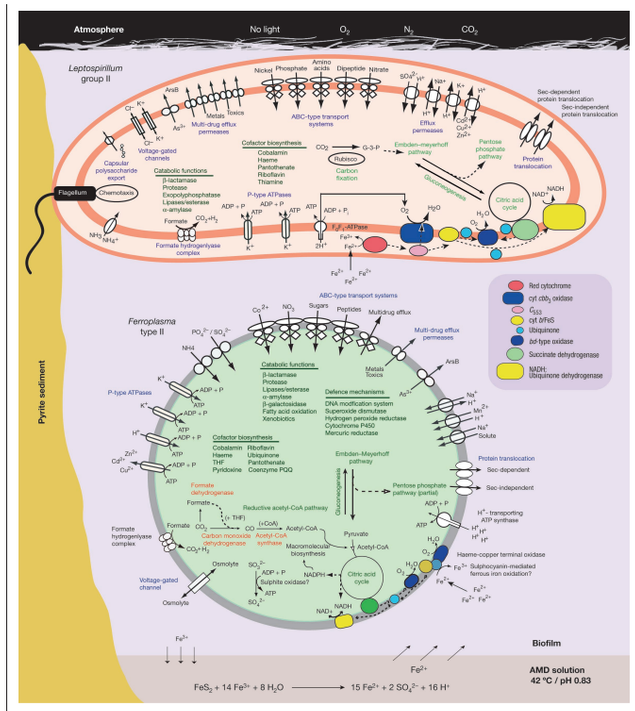
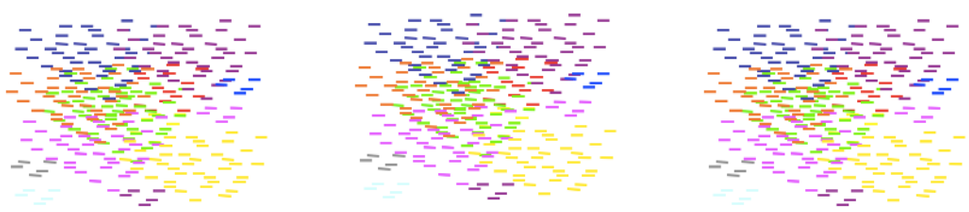


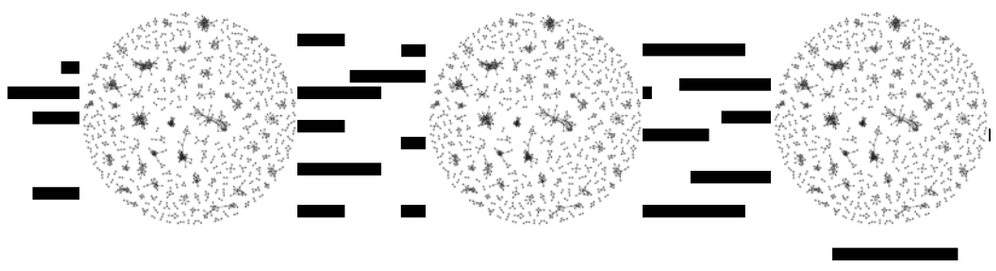
Figure 4 Cell metabolic cartoons constructed from the annotation of 2,180 ORFs identified in the *Leptospirillum* group II genome (63% with putative assigned function) and 1,931 ORFs in the *Ferroplasma* type II genome (58% with assigned function). The cell cartoons are shown within a biofilm that is attached to the surface of an acid mine drainage stream (viewed in cross-section). Tight coupling between ferrous iron oxidation, pyrite dissolution and acid generation is indicated. Rubisco, ribulose 1,5-bisphosphate carboxylase–oxygenase. THF, tetrahydrofolate.

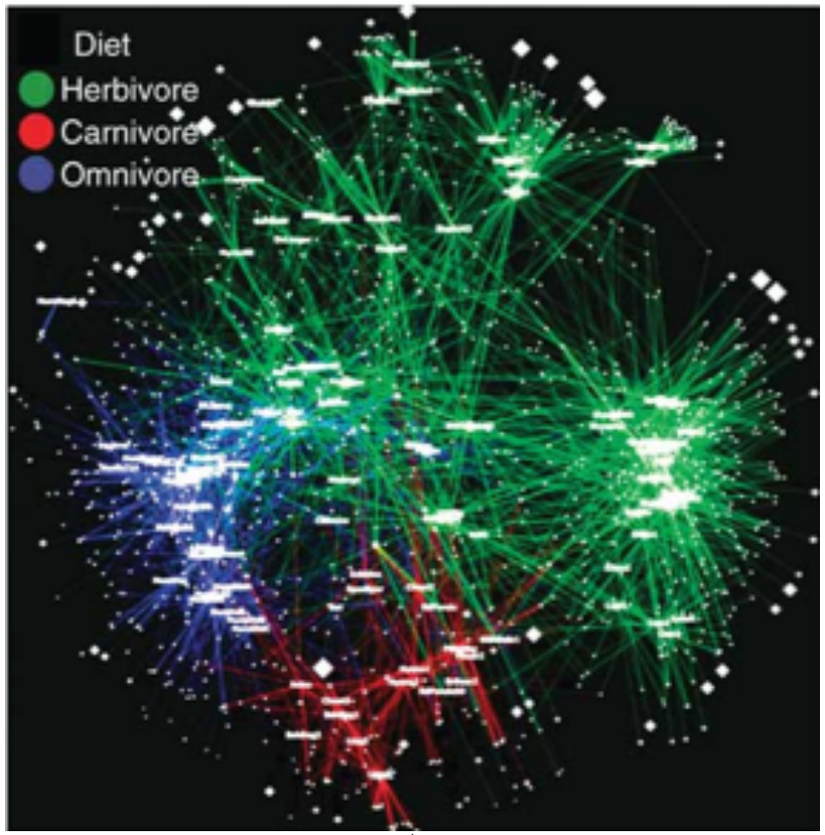
Taxa	Characters
B1	ACTGCACCTATCGTTCG
B2	ACTCCACCTATCGTTCG
E1	ACTCCAGCTATCGATCG
E2	ACTCCAGGTATCGATCG
A1	ACCCCAGCTCTCGCTCG
A2	ACCCCAGCTCTGGCTCG
New1	ACCCCAGCTCTGCCTCG
New2	AGGGGAGCTCTGCCTCG
New3	ACTCCAGCTATCGATCG
New4	ACTGCACCTATCGTTCG

Shotgun Metagenomics - Community Comparisons

DNA DNA DNA



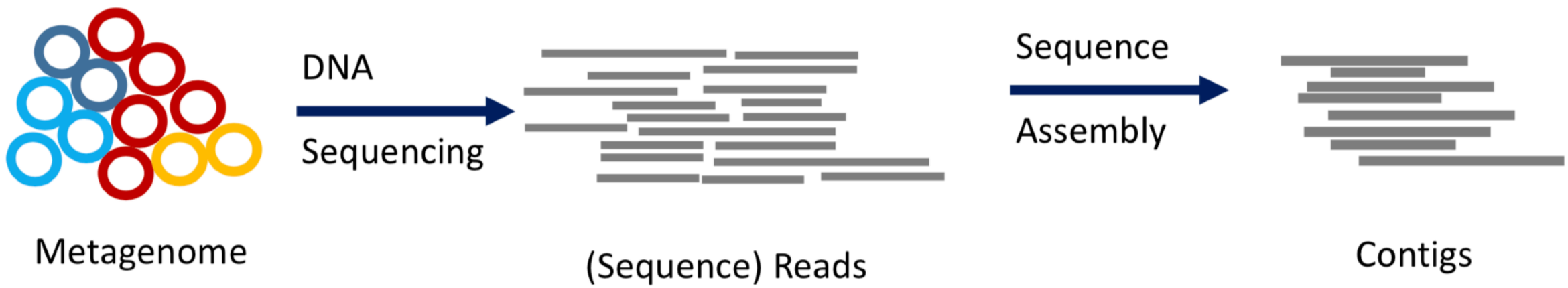




Taxa	Characters
B1	ACTGCACCTATCGTTCG
B2	ACTCCACCTATCGTTCG
E1	ACTCCAGCTATCGATCG
E2	ACTCCAGGTATCGATCG
A1	ACCCCAGCTCTCGCTCG
A2	ACCCCAGCTCTGGCTCG
New1	ACCCCAGCTCTGCCTCG
New2	AGGGGAGCTCTGCCTCG
New3	ACTCCAGCTATCGATCG
New4	ACTGCACCTATCGTTCG

MAGs

Binning : Grouping nucleotide sequences belonging to individual/similar organism/s



Sequencing and Microbes

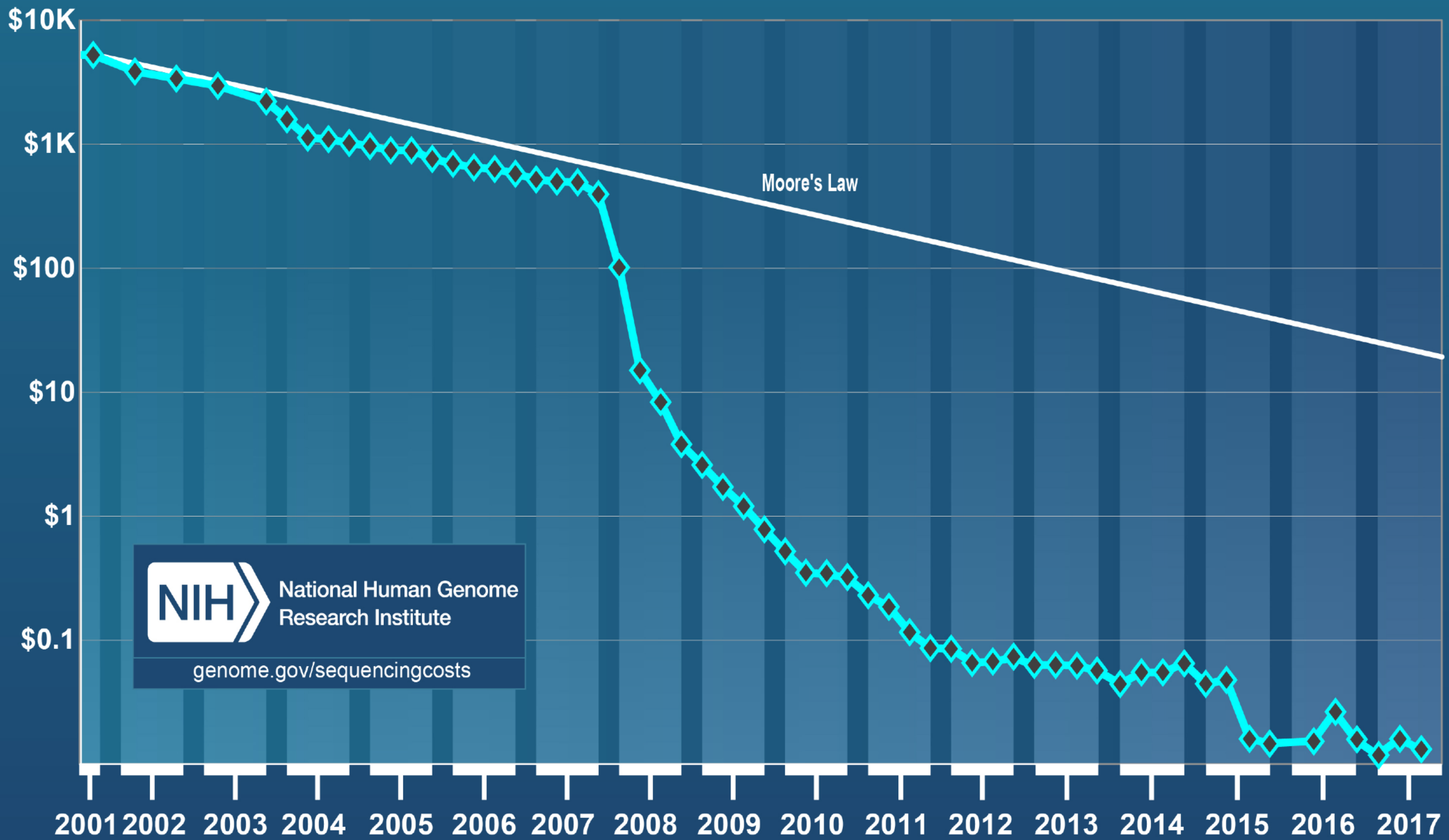
- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Sanger / Maxam-Gilbert
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

Sequencing and Microbes

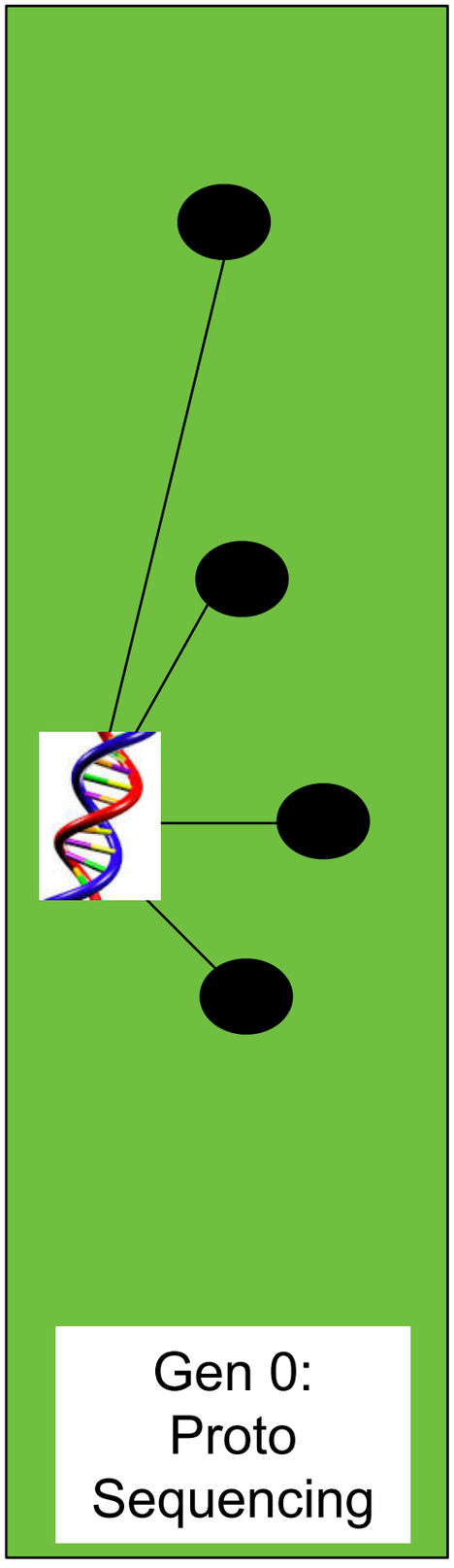
- Part 1: Four Eras of Sequencing and Microbes
  - Era 1: rRNA and the Tree of Life
  - Era 2: rRNA from environmental samples
  - Era 3: Genome sequencing
  - Era 4: Genomes from environmental samples
- Part 2: Evolution of Sequencing
  - Generation 0: Protosequencing
  - Generation 1: Manual Sequencing
  - Generation 2: Automation of Sanger
  - Generation 3: Clusters not clones
  - Generation 4: Single molecule sequencing

NOTE - New Eras Add On to Past Ones, Past Ones Do Not End

Cost per Raw Megabase of DNA Sequence



Evolution of Sequencing



Gen 0: Proto-Sequencing

Proc. Nat. Acad. Sci. USA
Vol. 70, No. 12, Part I, pp. 3581–3584, December 1973

The Nucleotide Sequence of the *lac* Operator

(regulation/protein–nucleic acid interaction/DNA–RNA sequencing/oligonucleotide priming)

WALTER GILBERT AND ALLAN MAXAM

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Communicated by J. D. Watson, August 9, 1973

TABLE 1. *Pyrimidine tracts from the lac operator*

Tract	Moles	Yields
pCp	4–5	(4.6)
pTp	7	(7.0)
pTpTp	4	(3.8)
pTpCpCp	1	(2.0)
pCpTpCp	1	
pTpTpCpCp	1	(0.7)

Pyrimidine tracts were isolated and fingerprinted. The sequences were determined by partial digestion of phosphatase-treated material by spleen and by venom phosphodiesterase. The relative molar yields are the averages of three experiments, taking the TCC and CTC isostichs together as 2 mol/mol of operator.

Proc. Nat. Acad. Sci. USA
Vol. 72, No. 2, pp. 737–741, February 1975

Nucleotide Sequence of an RNA Polymerase Binding Site from the DNA of Bacteriophage fd

(promoters/DNA sequencing/protein–DNA interaction)

HEINZ SCHALLER, CHRISTOPHER GRAY, AND KARIN HERRMANN

Max-Planck-Institut für Virusforschung, Tübingen and Lehrstuhl für Mikrobiologie der Universität Heidelberg, 6900 Heidelberg, Im Neuenheimer Feld 280, West Germany

Communicated by H. Gobind Khorana, December 6, 1974

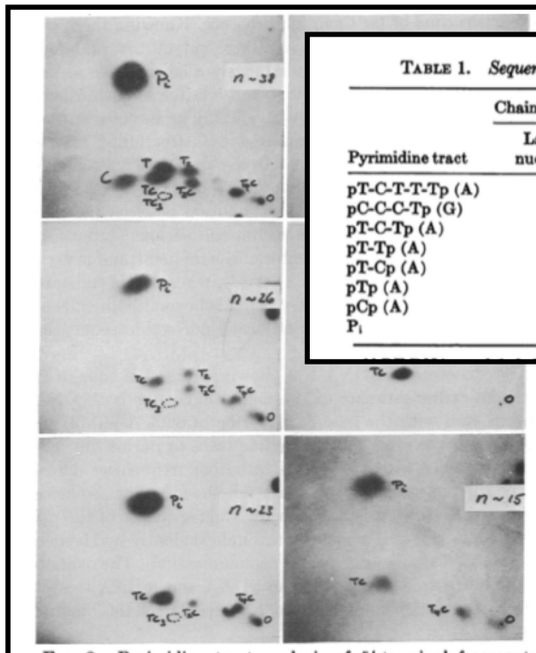
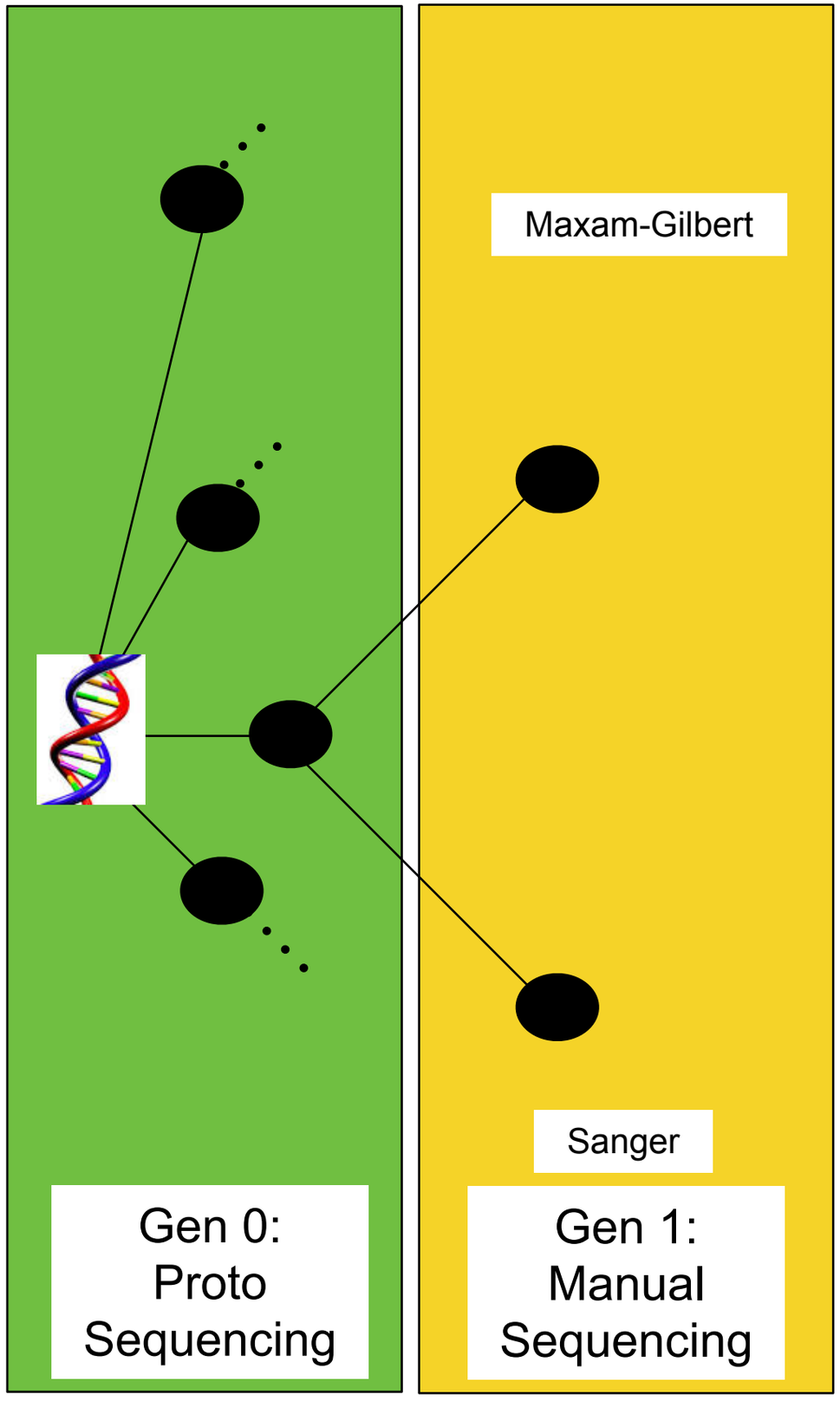


FIG. 2. Pyrimidine tract analysis of 5′-terminal fragments from pDNA-I minus strand. DNA fragments were prepared and separated as described in Fig. 1. Pyrimidine tracts from the individual spots were identified by two-dimensional thin-layer chromatography. Base composition of pyrimidine tracts and chainlength of the DNA fragments are indicated. Plate $n \sim 38$ was derived from full-length pDNA-I.

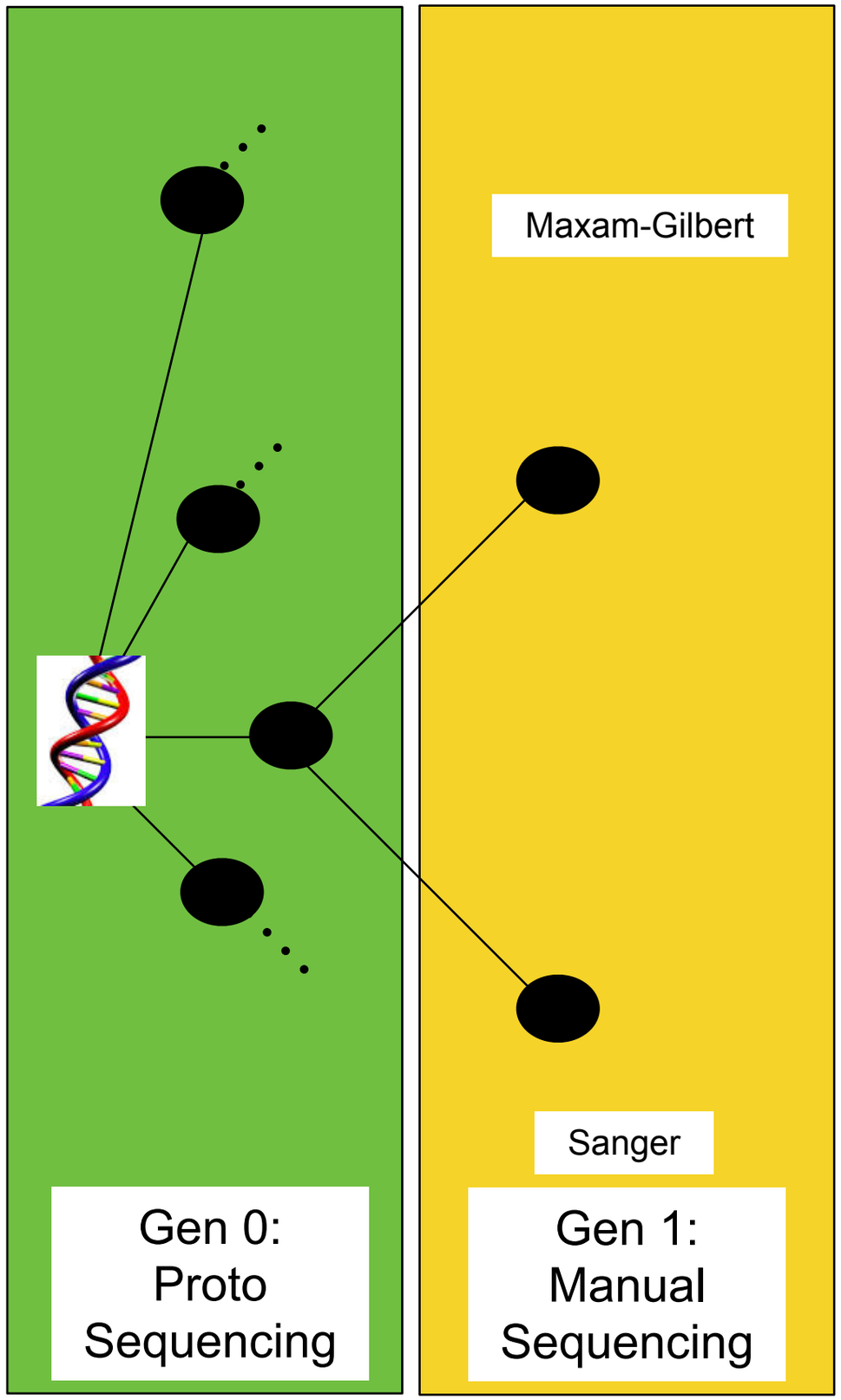
TABLE 1. *Sequences and yields of pyrimidine tracts from pDNA-I template (minus) strand and 5′-terminal fragments*

Pyrimidine tract	Chainlength (n)	38	38	38	38	26	23	21	19	15	9	4
	Labeled nucleotide	T(C)	T, C	G	A(C)				A (C)			
pT-C-T-T-Tp (A)		1.1	1.0	—	1.0	1.0	1.0	1.0	1.0	1.0	—	—
pC-C-C-Tp (G)		1.1	1.0	1.0	(0.8)	(0.8)	(0.7)	(0.8)	—	—	—	—
pT-C-Tp (A)		1.0	0.9	—	0.9	0.7	0.4	—	—	—	—	—
pT-Tp (A)		1.0	0.9	—	0.9	0.7	0.1	—	—	—	—	—
pT-Cp (A)		2.1	1.8	—	2.2	1.1	1.2	1.1	1.4	1.0	1.0	—
pTp (A)		1.0	0.9	—	1.1	0.2	—	—	—	—	—	—
pCp (A)		(0.7)	0.8	—	(0.7)	(—)	(—)	(—)	(—)	(—)	(—)	(—)
P_i		—	—	5.1	4.6	4.2	5.1	4.5	4.2	4.1	4.9	1

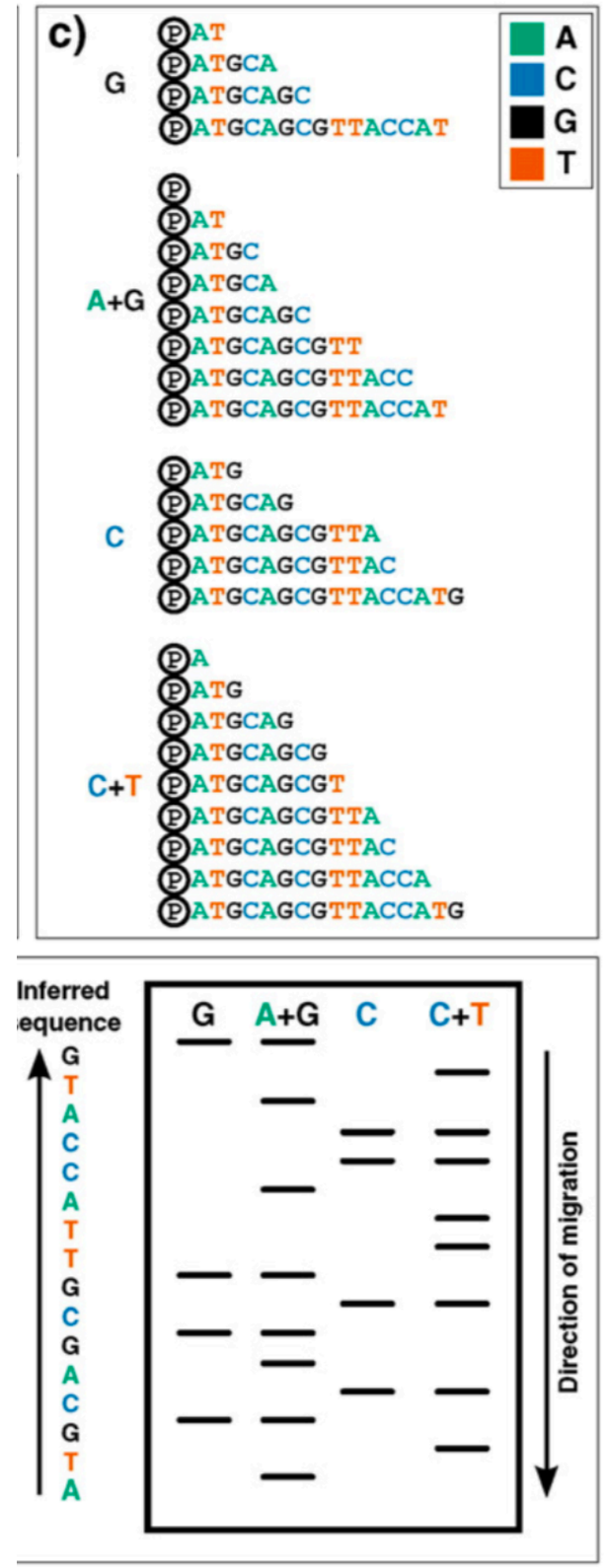
Generation 1: Manual Sequencing



Generation 1: Maxam-Gilbert



Maxam-Gilbert



Heather and Chain 2016

Generation 1: Maxam-Gilbert

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 2, pp. 560–564, February 1977
Biochemistry

A new method for sequencing DNA

(DNA chemistry/dimethyl sulfate cleavage/hydrazine/piperidine)

ALLAN M. MAXAM AND WALTER GILBERT

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Proc. Nat. Acad. Sci. USA
Vol. 70, No. 12, Part I, pp. 3581–3584, December 1973

The Nucleotide Sequence of the *lac* Operator

(regulation/protein–nucleic acid interaction/DNA–RNA sequencing/oligonucleotide priming)

WALTER GILBERT AND ALLAN MAXAM

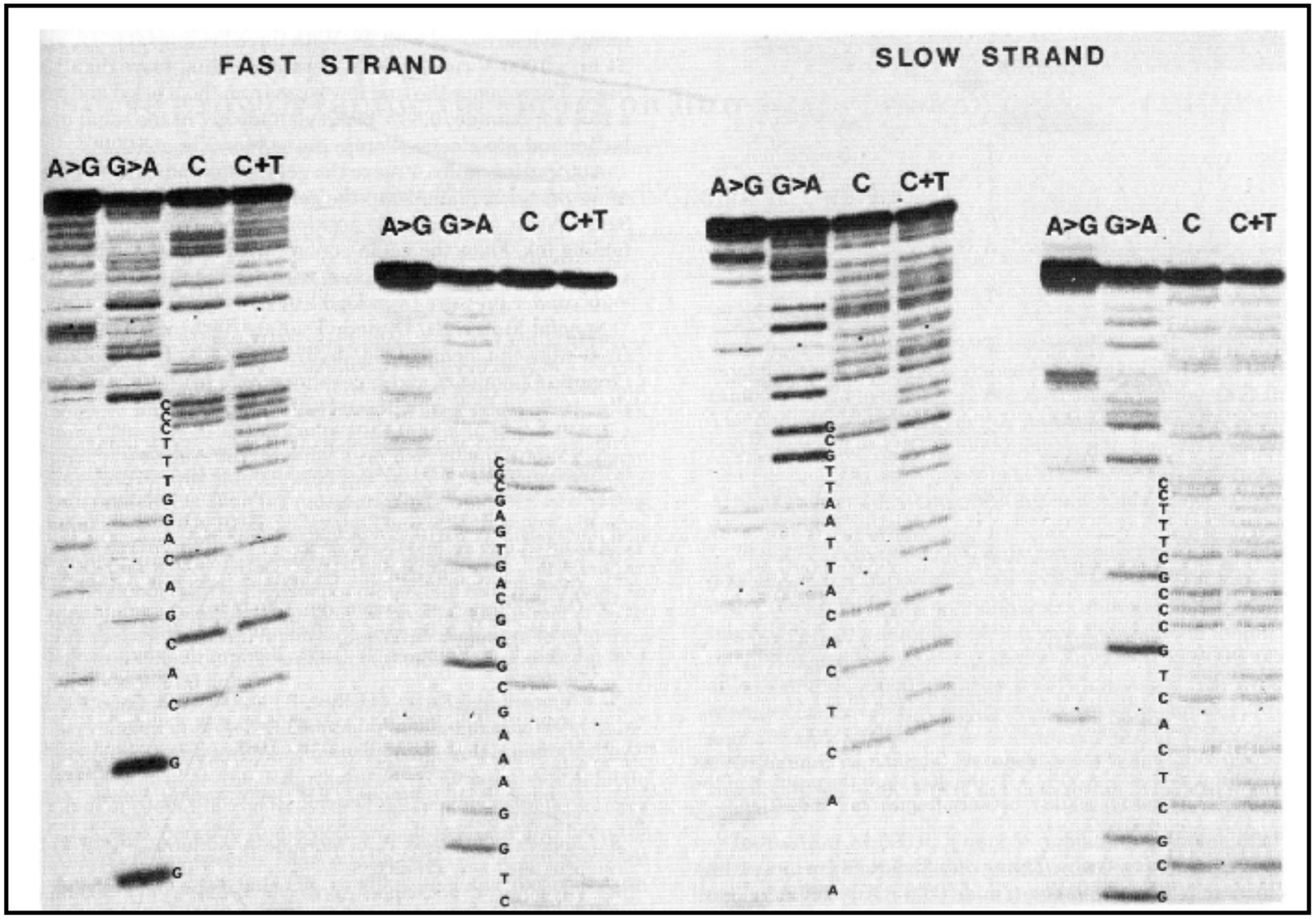
Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Communicated by J. D. Watson, August 9, 1973

ABSTRACT The *lac* repressor protects the *lac* operator against digestion with deoxyribonuclease. The protected fragment is double-stranded and about 27 base-pairs long. We determined the sequence of RNA transcription copies of this fragment and present a sequence for 24 base pairs. It is:

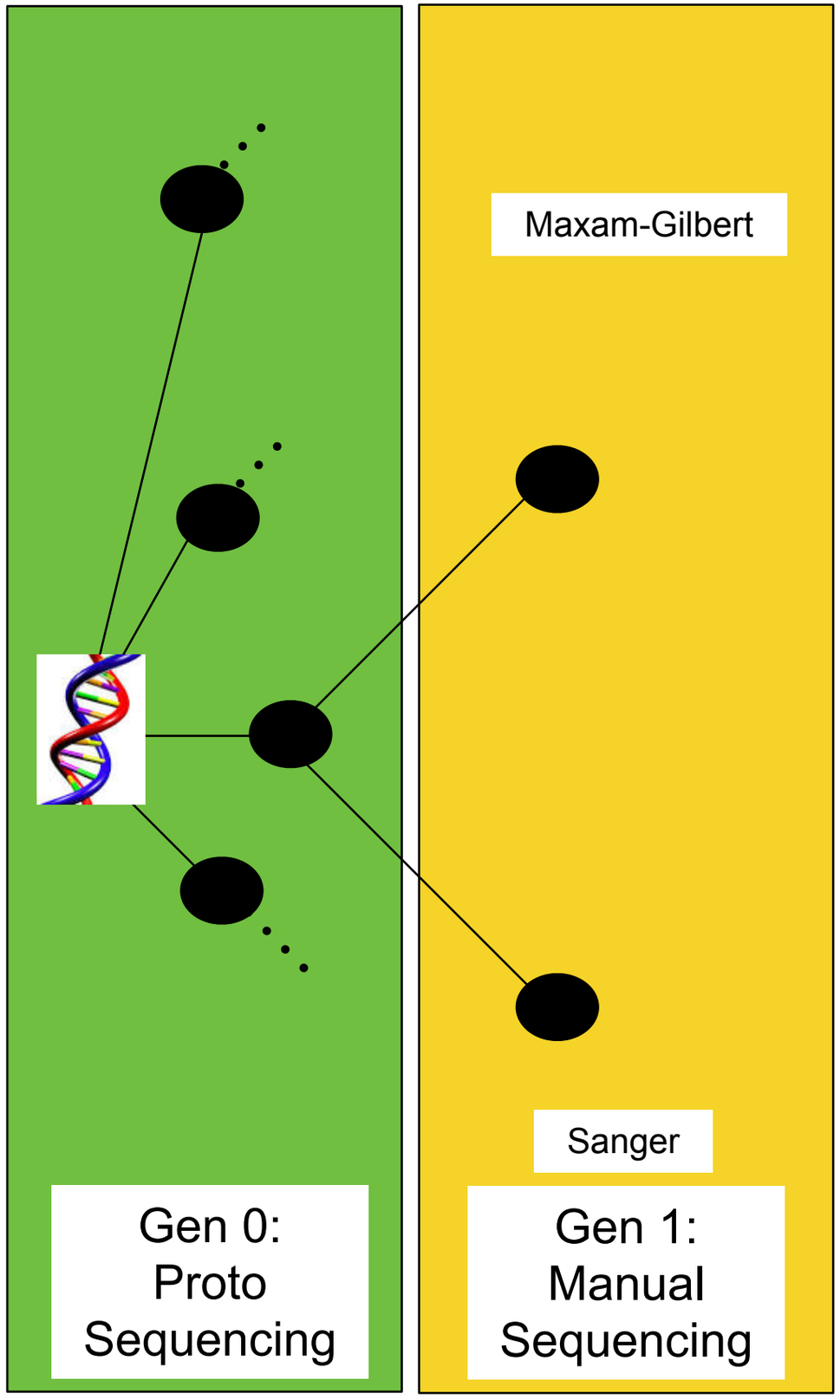
5′--T G G A A T T G T G A G C G G A T A A C A A T T 3′
3′--A C C T T A A C A C T C G C C T A T T G T T A A 5′

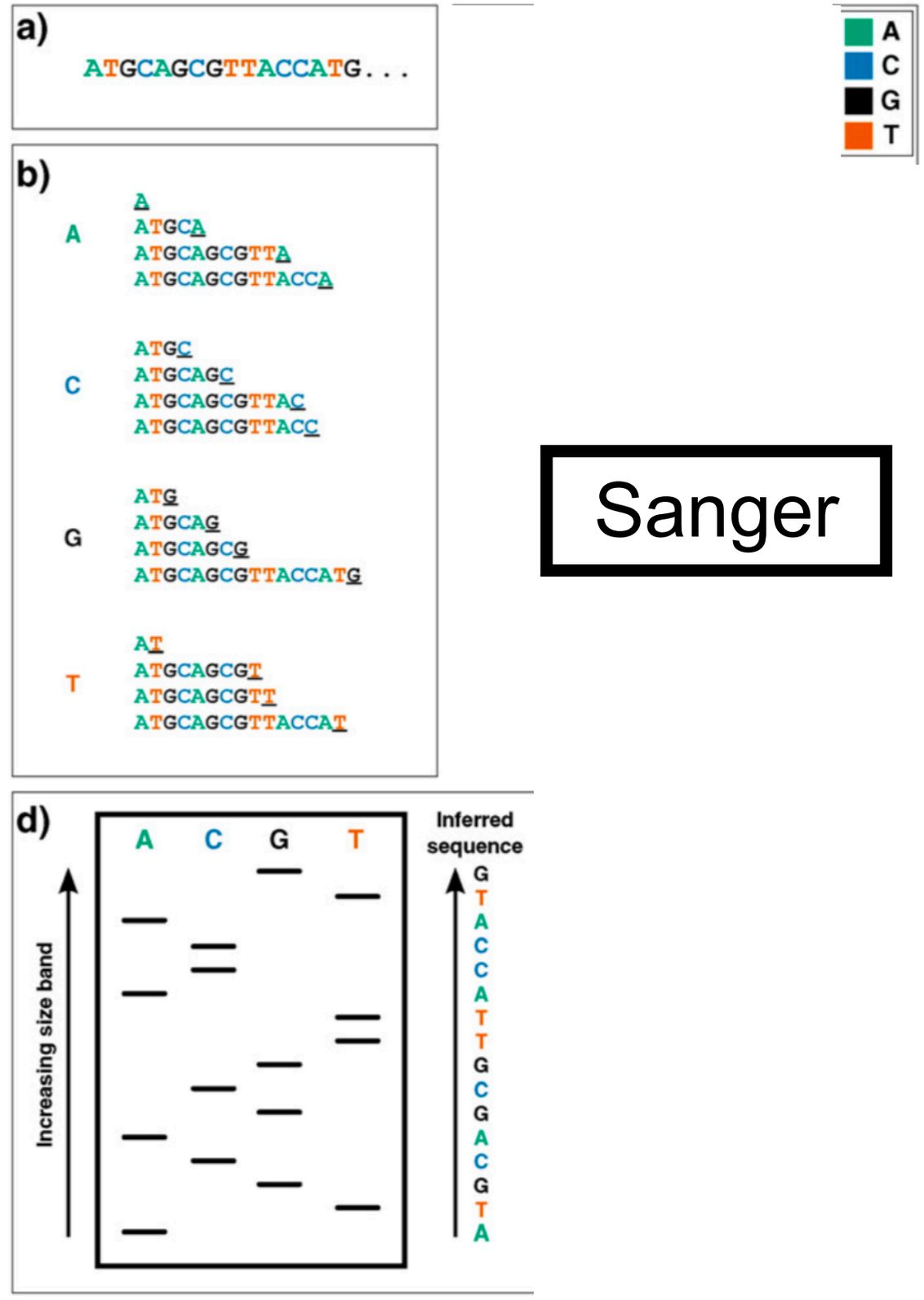
The sequence has 2-fold symmetry regions; the two longest are separated by one turn of the DNA double helix.



From http://www.pnas.org/content/74/2/560.full.pdf

Generation 1: Sanger Sequencing





Sanger

Heather and Chain 2016

Generation 1: Sanger Sequencing

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

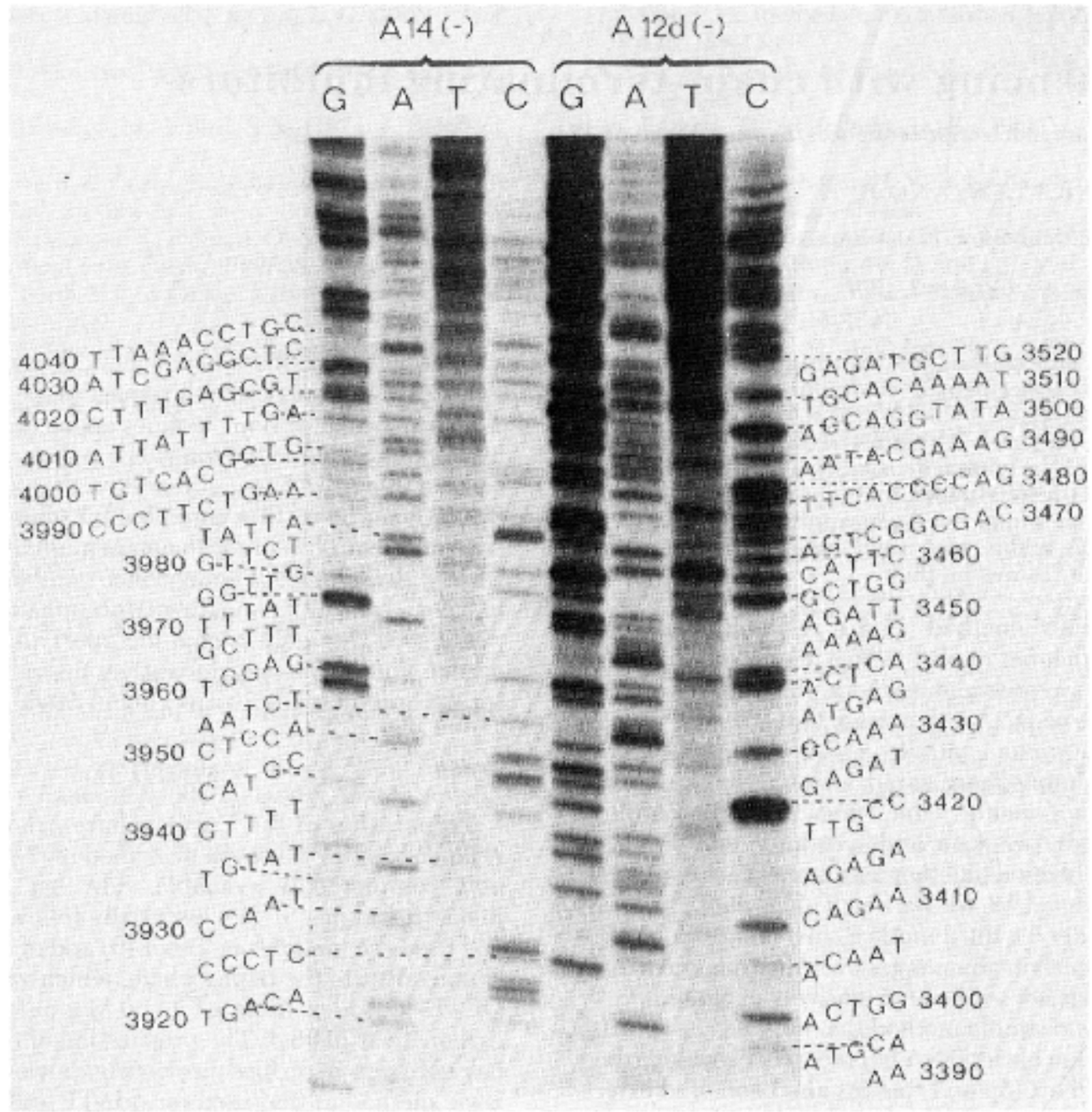
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

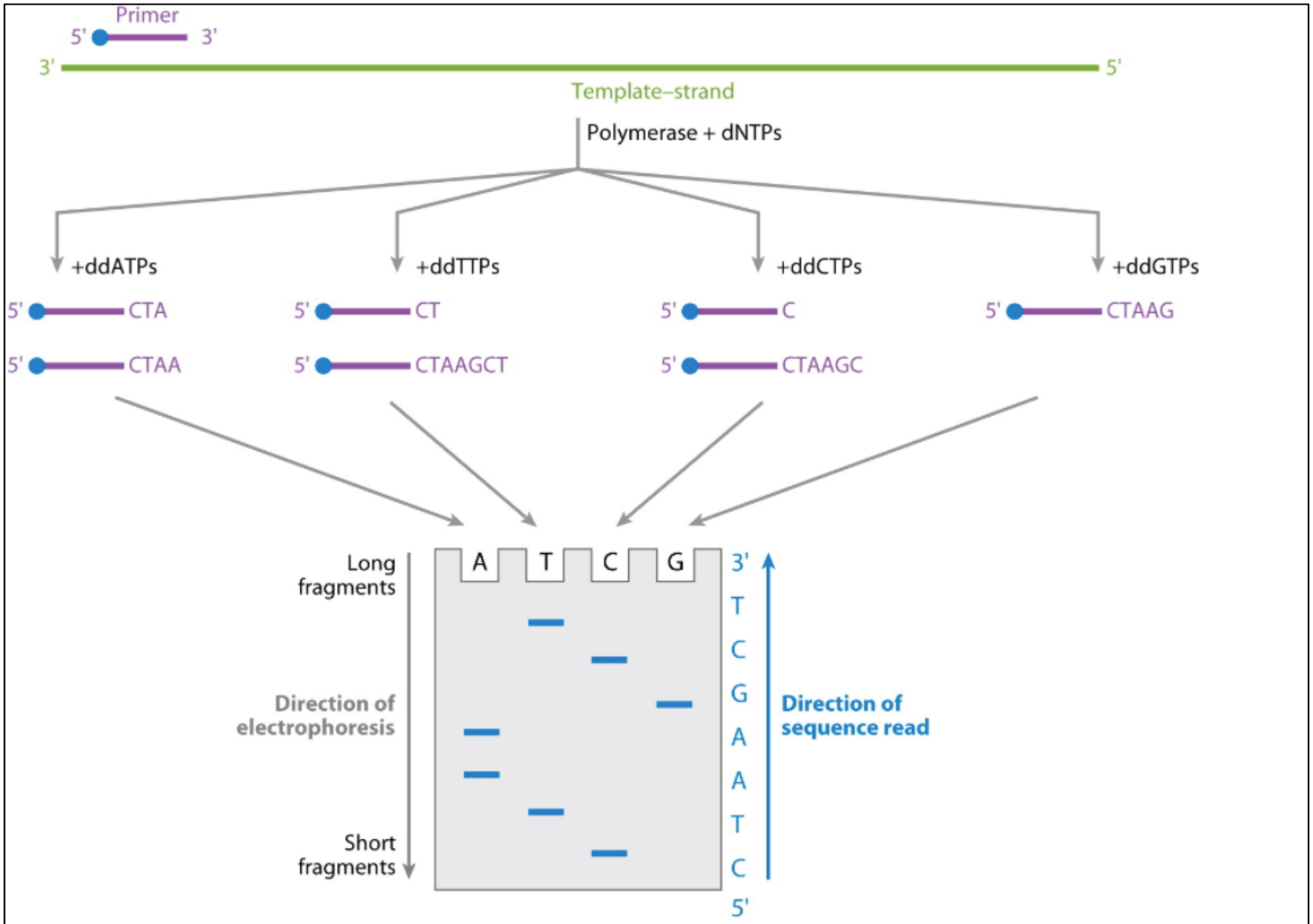
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/

Generation 1: Sanger Sequencing



Mardis ER. 2013.
Annu. Rev. Anal. Chem. 6:287–303

- http://www.annualreviews.org/na101/home/literatum/publisher/ar/journals/content/anchem/2013/anchem.2013.6.issue-1/annurev-anchem-062012-092628/20130605/images/medium/ac60287.f1.gif

Nobel Prize 1980: Berg, Gilbert, Sanger

The Nobel Prize in Chemistry 1980
Paul Berg, Walter Gilbert, Frederick Sanger

The Nobel Prize in Chemistry 1980


Paul Berg


Walter Gilbert


Frederick Sanger

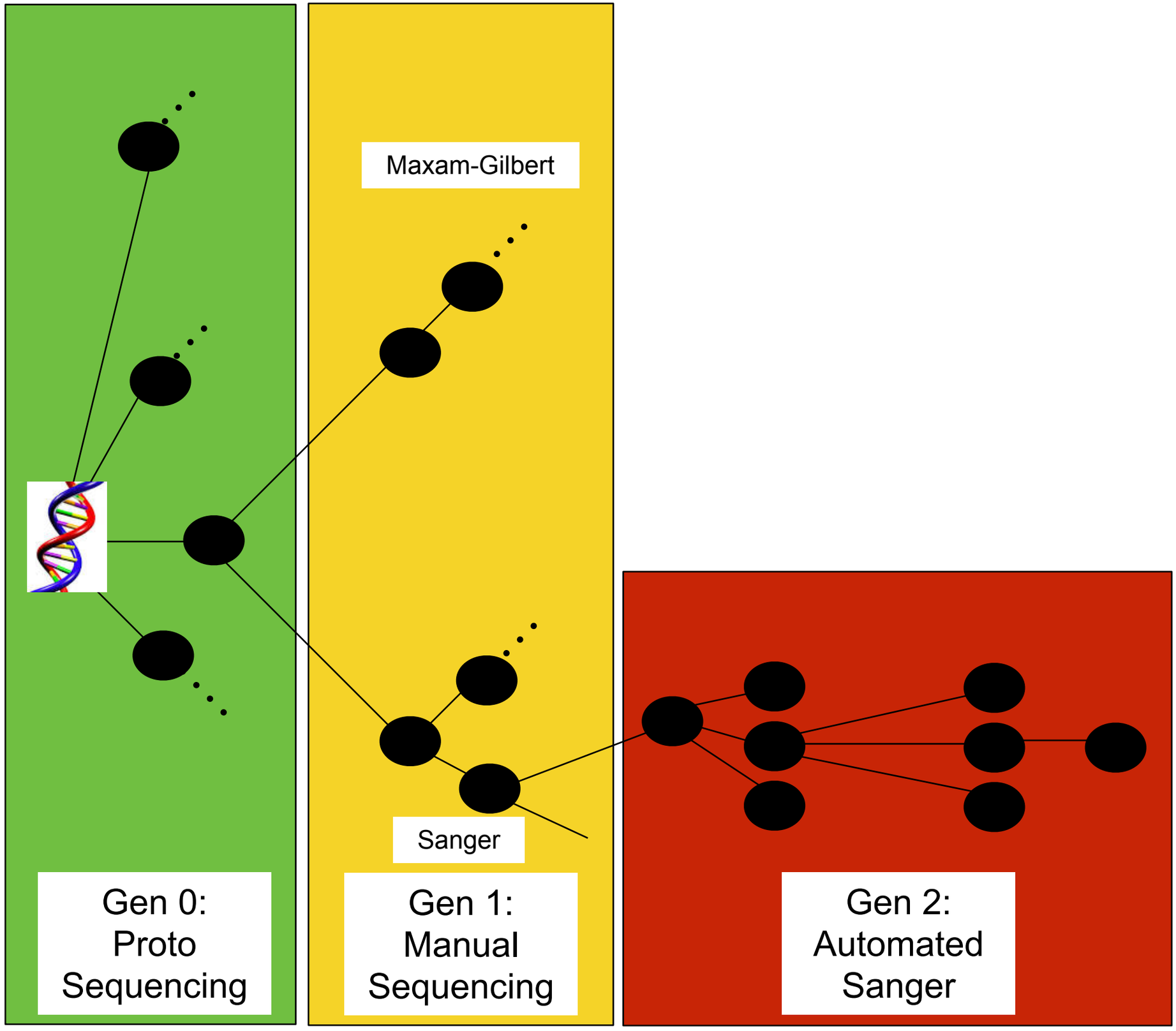
The Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg *"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"*, the other half jointly to Walter Gilbert and Frederick Sanger *"for their contributions concerning the determination of base sequences in nucleic acids"*.

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/

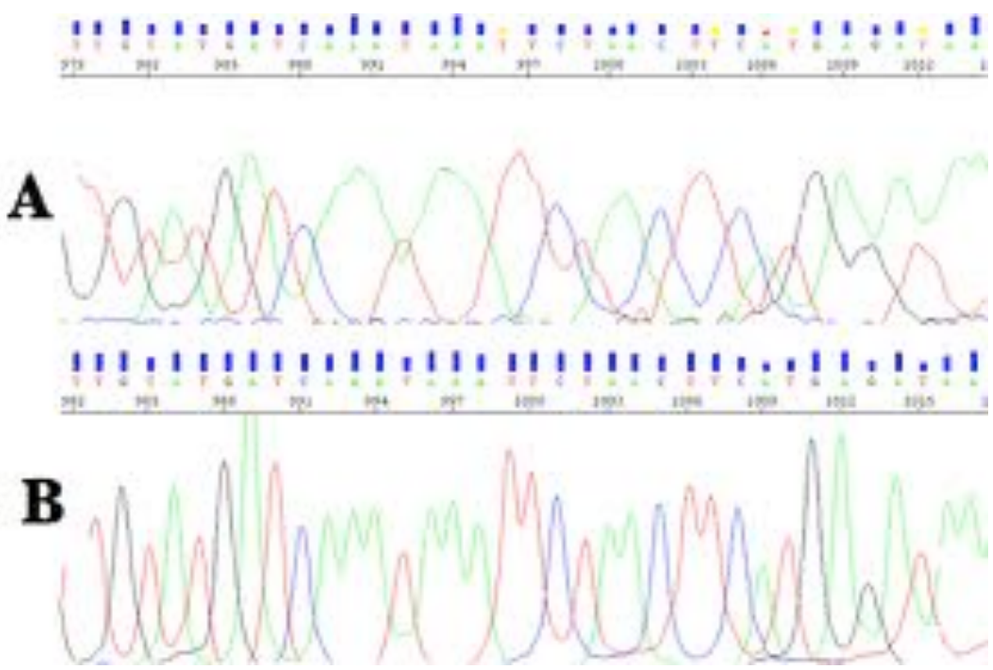
Some Key Innovations for Generation 1

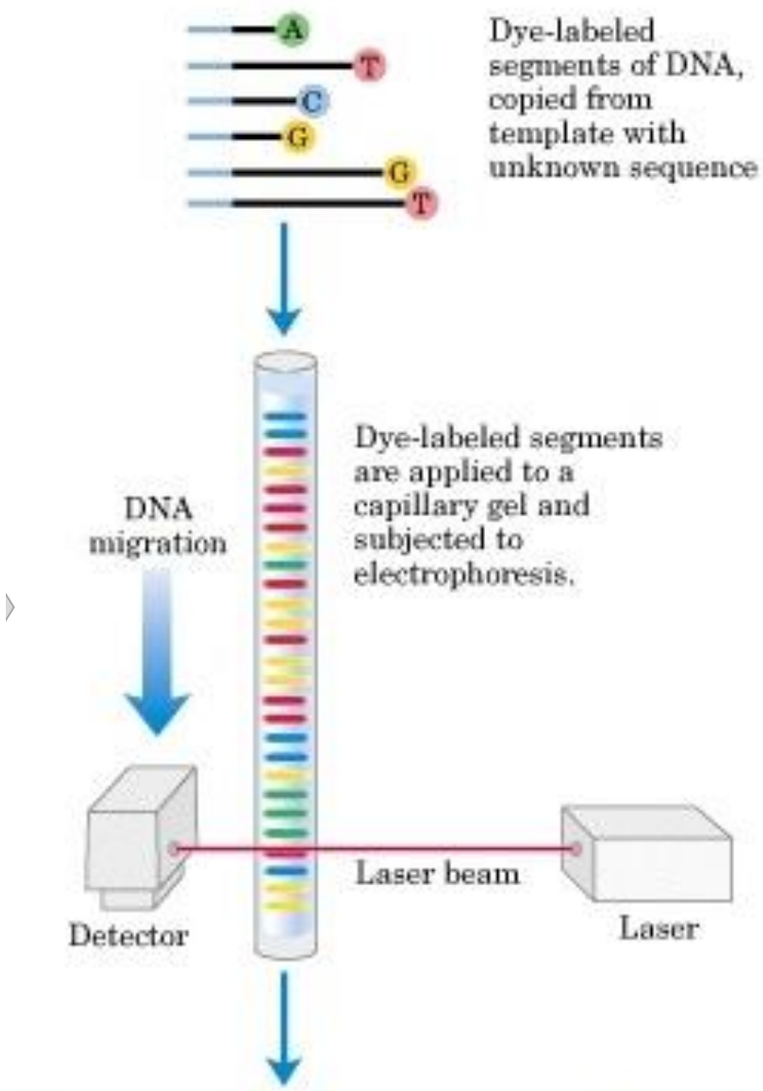
- Polyacrylamide gels
- Nucleotide chemistry
- Synthesis of primers
- Chain termination by ddNTPs

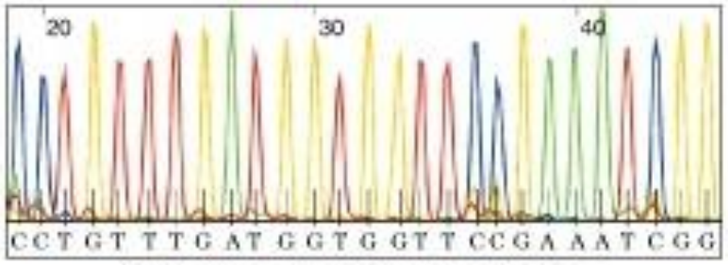
Generation 2: Automation of Sanger Sequencing



Automation of Sanger







Computer-generated result after bands migrate past detector

Many Systems for Sanger Automation

ABI 3700

ABI 3730

ABI 3730xl

Megabase 1000

Megabase 4000

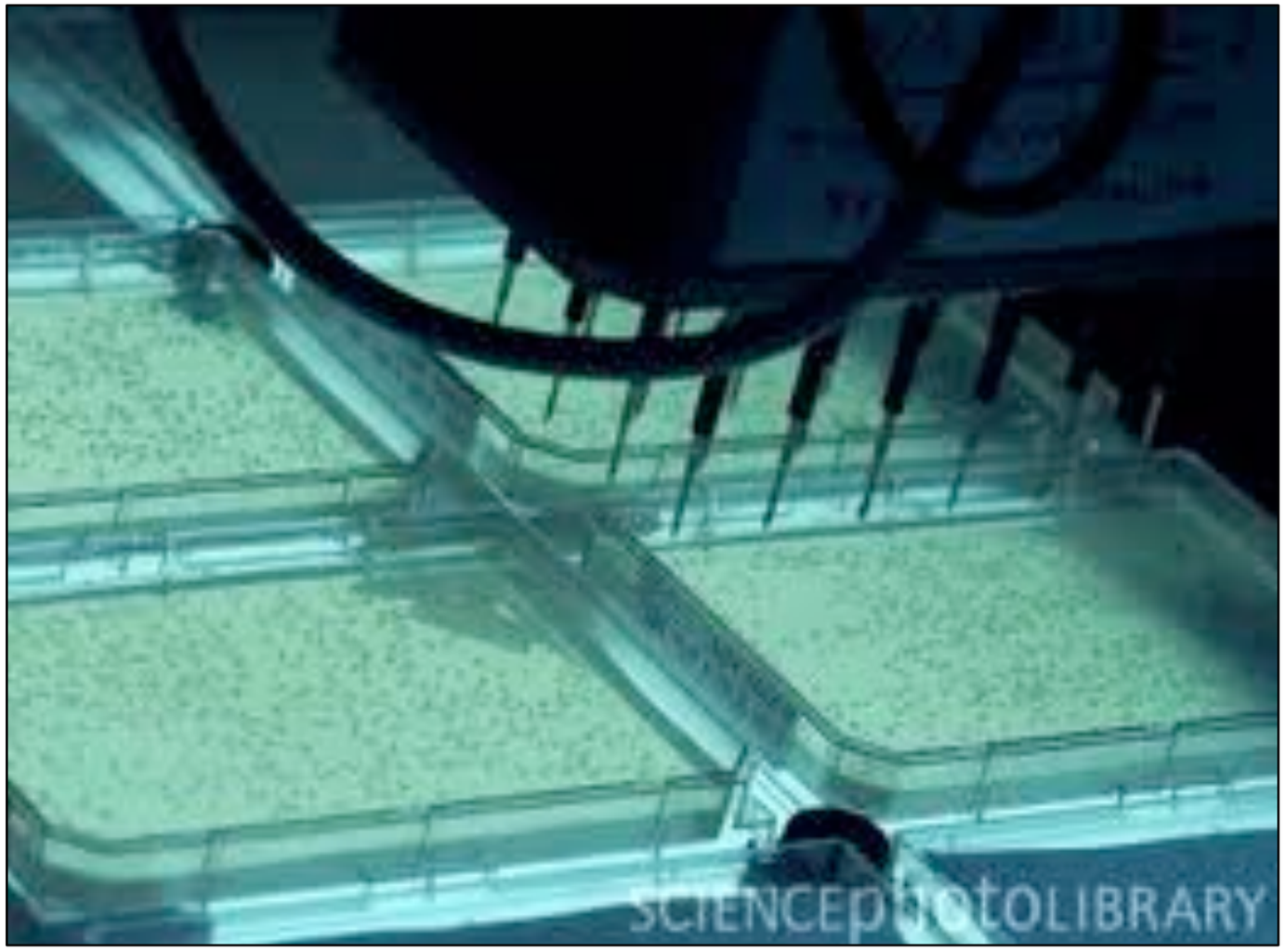
LiCor

Thanks to Robin Coope and Dale Yazuki for comments on 2014 talk

Automation of Sanger Innovations

- Fluorescence not radioactivity
- Capillaries not gels

Other Innovations for Autokation of Sanger

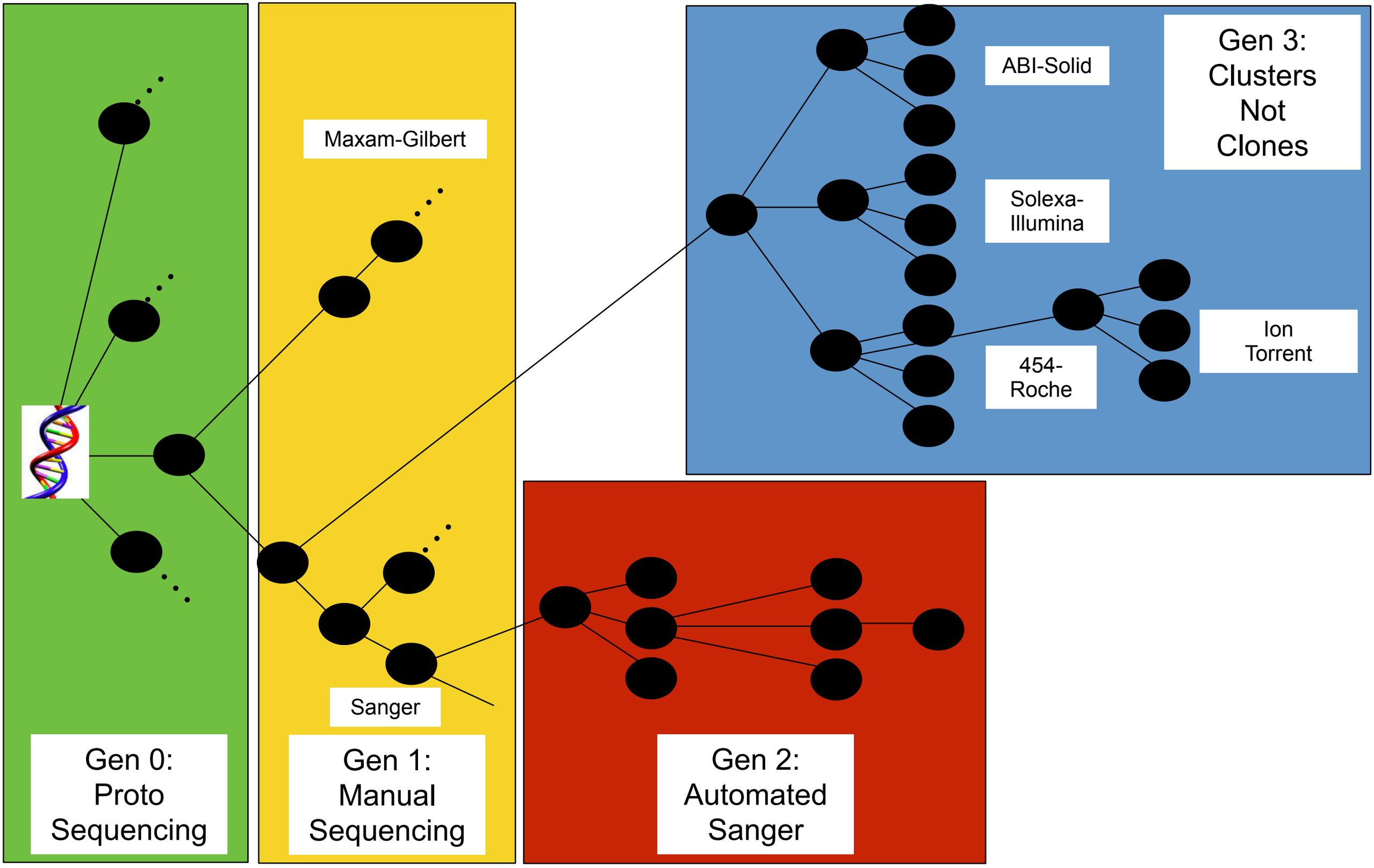


Some Automated Sanger Highlights

- 1991: ESTs by Venter
- 1995: *H. influenzae* shotgun genome
- 1996: Yeast, archaeal genomes
- 1998: 1st animal genome - *C. elegans*
- 1999: *Drosophila* shotgun genome
- 2000: *Arabidopsis* genome
- 2000: Human genome
- 2004: Shotgun metagenomics

Thanks to Keith Bradnam for C. elegans suggestion

Generation 3: Clusters Not Clones



Generation III = “NextGen”

Solexa / Illumina

454 / Roche

Ion Torrent

ABI Solid

Generation III: Clusters not Clones

Solexa / Illumina

454 / Roche

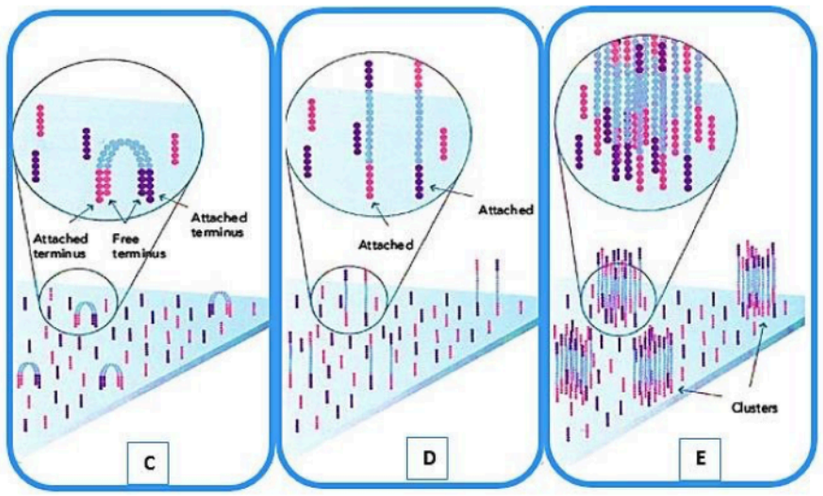
Ion Torrent

ABI Solid

Generation III: Clusters not Clones

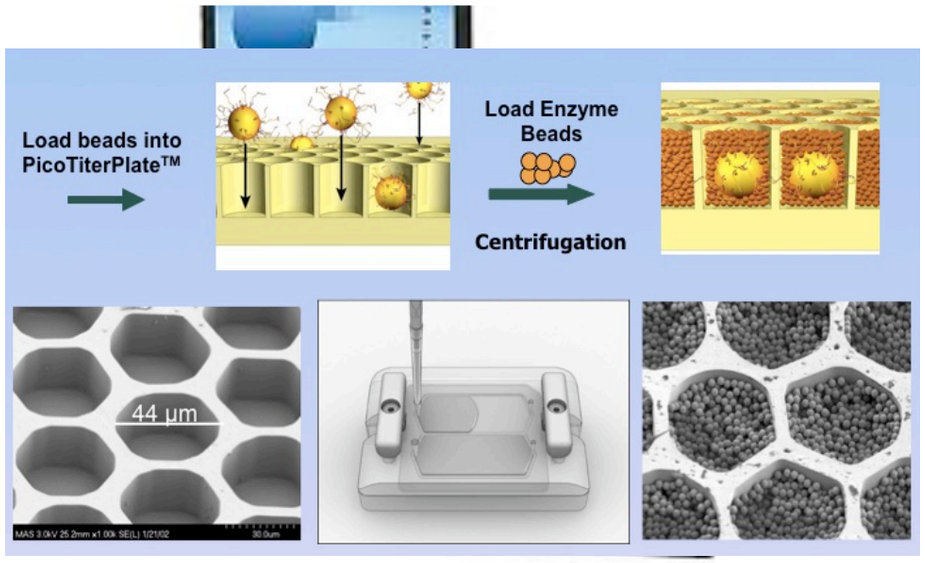
Solexa / Illumina

Clusters on Flowcell



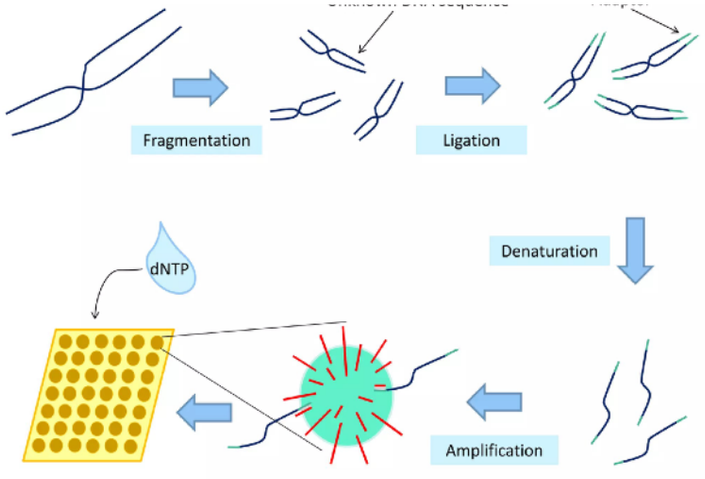
454 / Roche

Clusters in Beads



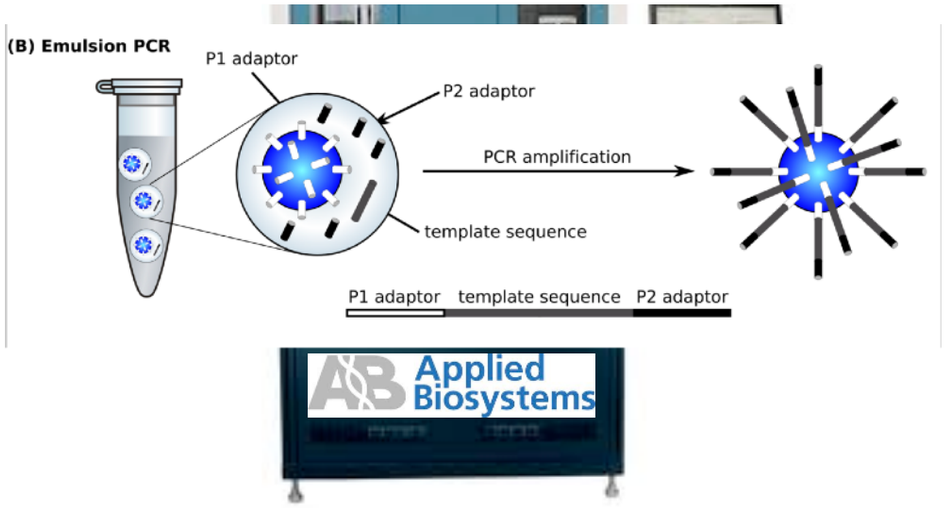
Ion Torrent

Clusters in Beads

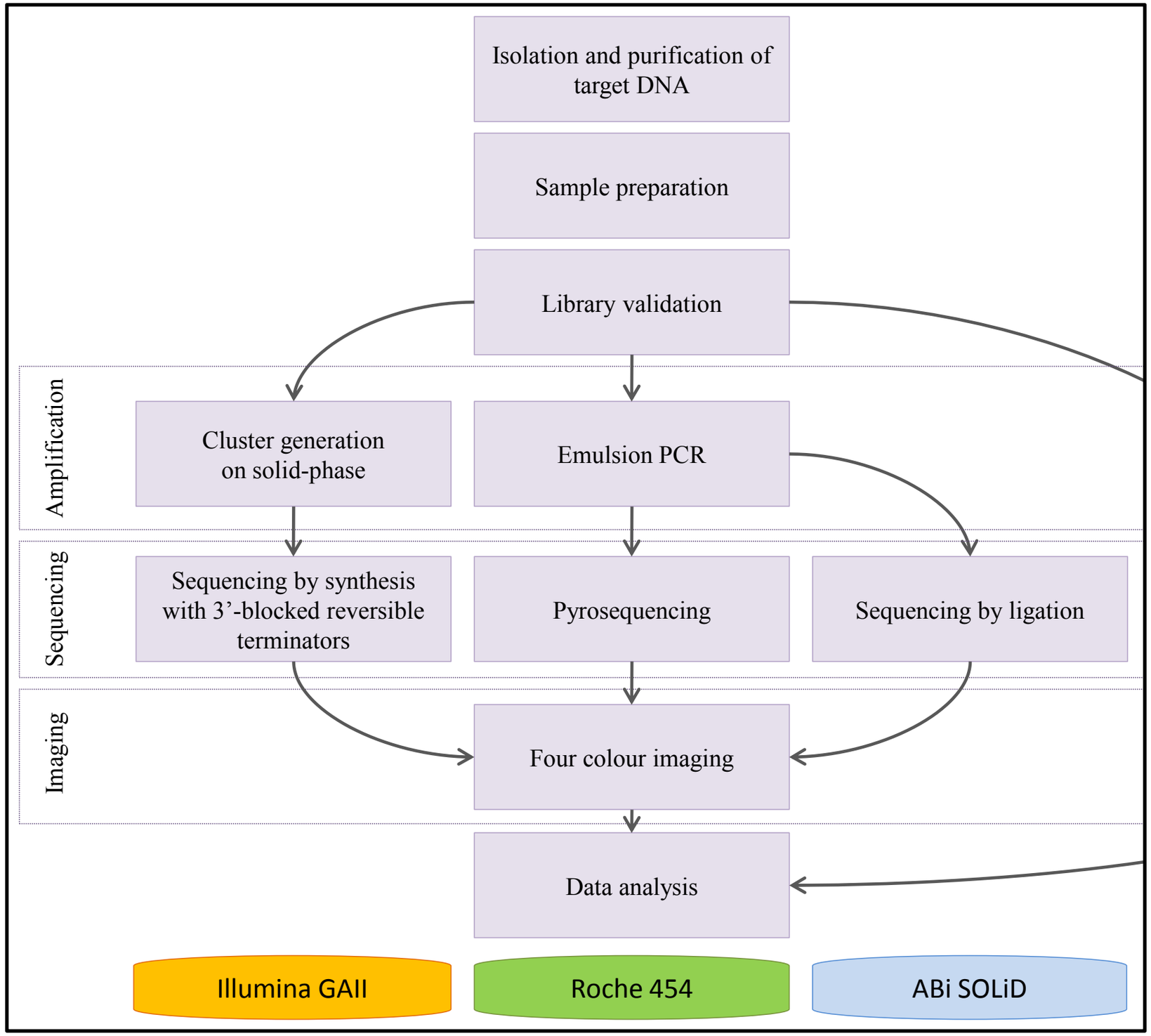


ABI Solid

Clusters in Beads



NextGen Sequencing Outline



From Slideshare presentation of Cosentino Cristian
http://www.slideshare.net/cosentia/high-throughput-equencing

Generation III: Dominant Player

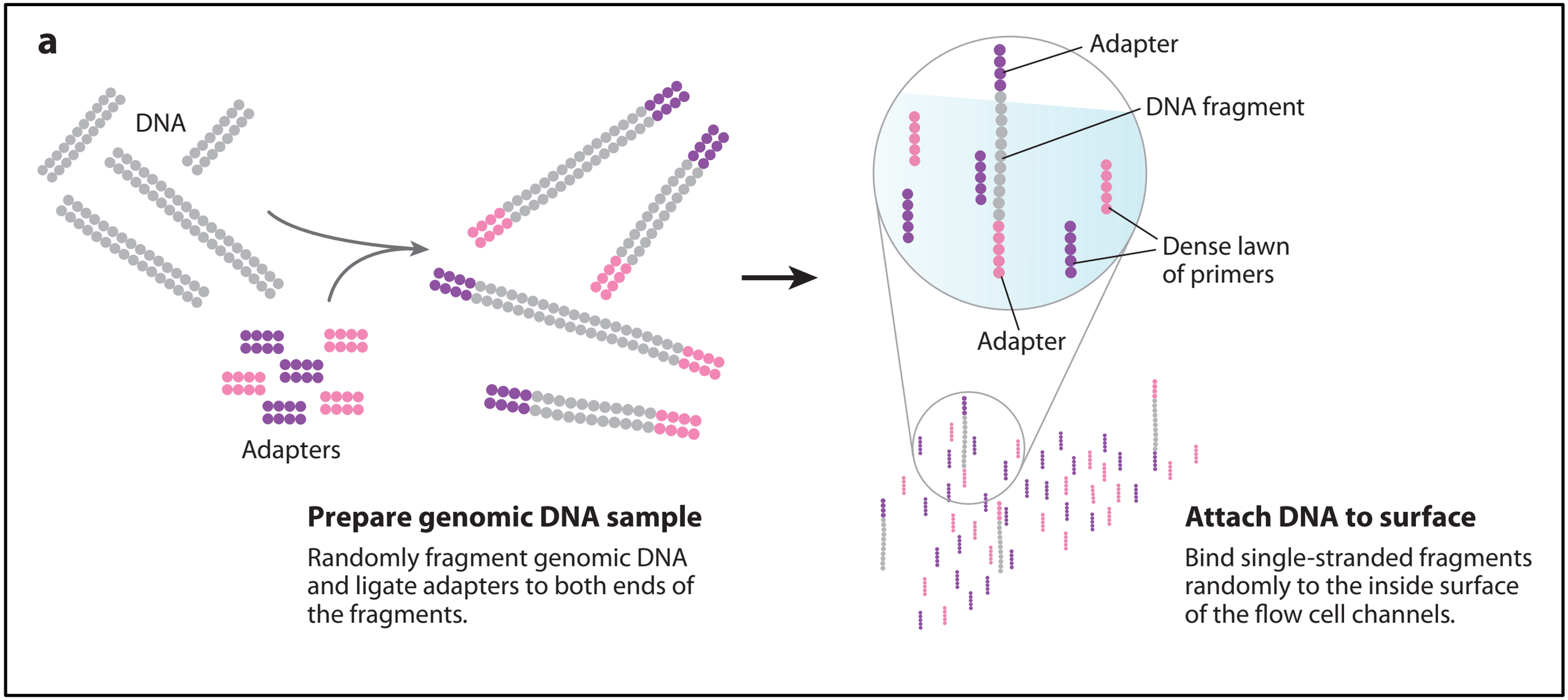
Solexa / Illumina

454 / Roche

Ion Torrent

ABI Solid

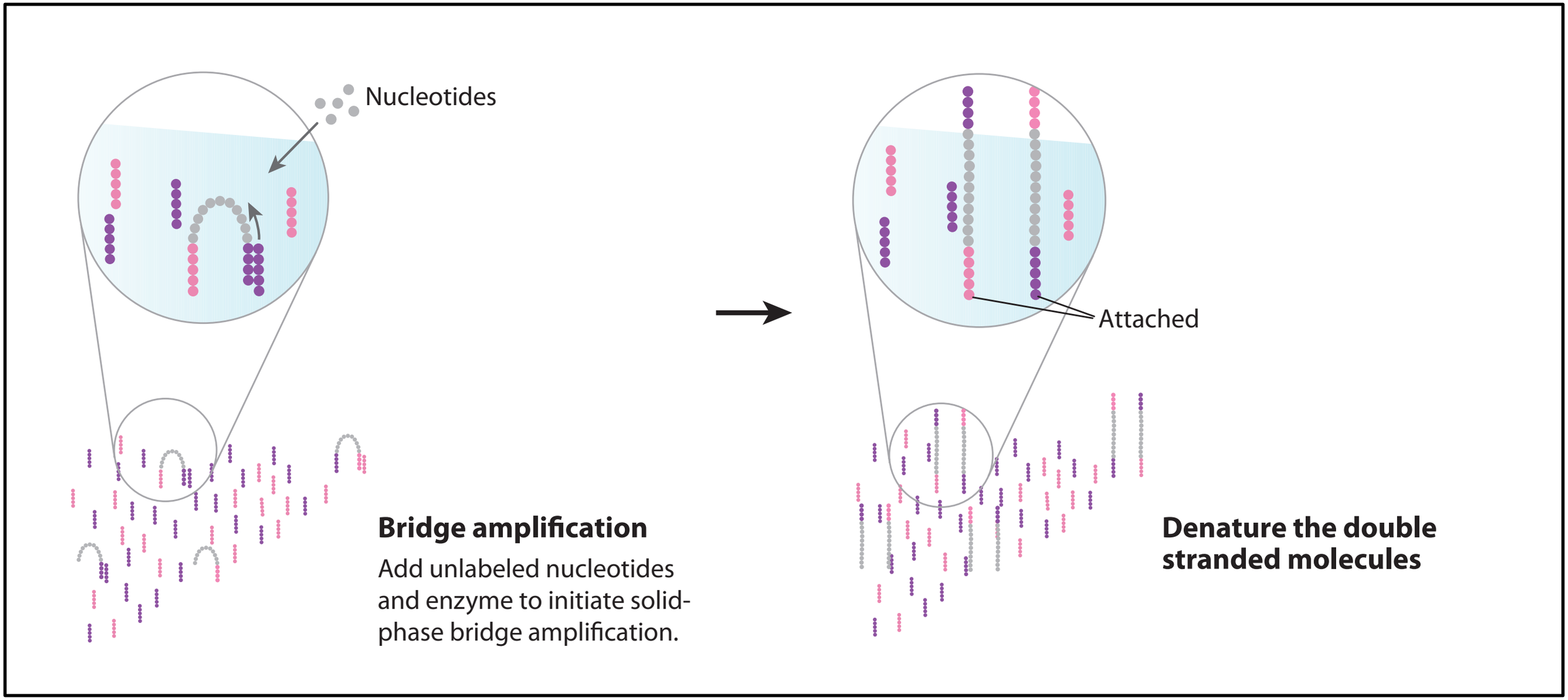
Illumina Step 1: Prep & Attach DNA



Step 1: Sample Preparation The DNA sample of interest is sheared to appropriate size (average ~800bp) using a compressed air device known as a nebulizer. The ends of the DNA are polished, and two unique adapters are ligated to the fragments. Ligated fragments of the size range of 150-200bp are isolated via gel extraction and amplified using limited cycles of PCR

From Mardis 2008. Annual Rev. Genetics 9: 387.

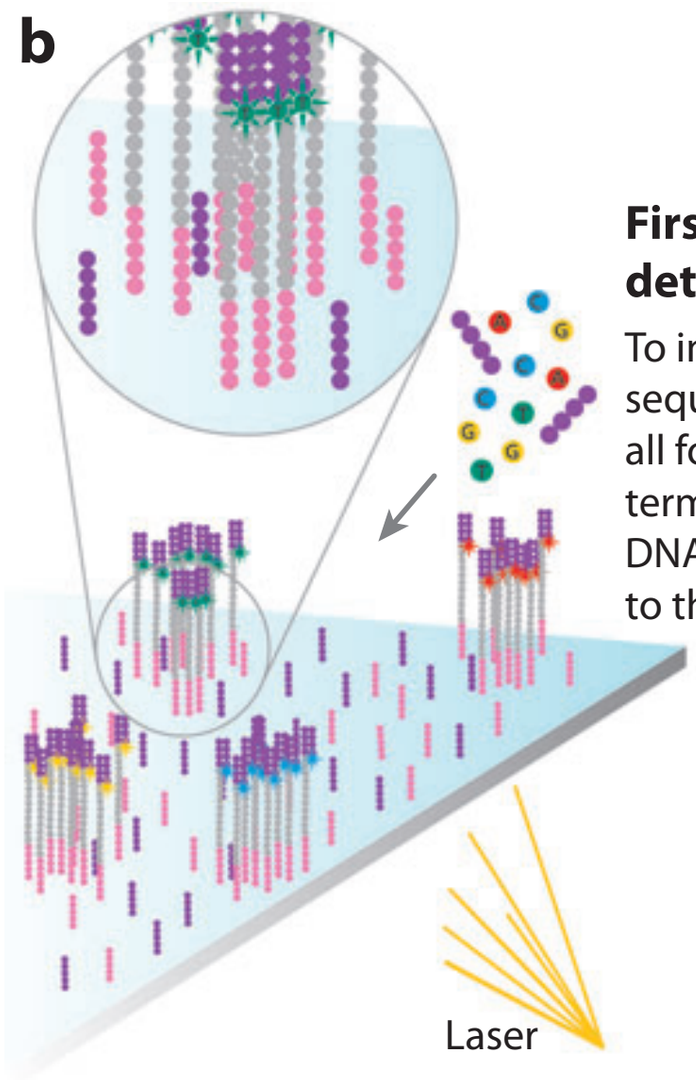
Illumina Step 2: Clusters by Bridge PCR



- From : http://seqanswers.com/forums/showthread.php?t=21. Steps 2-6: Cluster Generation by Bridge Amplification. In contrast to the 454 and ABI methods which use a bead-based emulsion PCR to generate "polonies", Illumina utilizes a unique "bridged" amplification reaction that occurs on the surface of the flow cell. The flow cell surface is coated with single stranded oligonucleotides that correspond to the sequences of the adapters ligated during the sample preparation stage. Single-stranded, adapter-ligated fragments are bound to the surface of the flow cell exposed to reagents for polyermase-based extension. Priming occurs as the free/distal end of a ligated fragment "bridges" to a complementary oligo on the surface. Repeated denaturation and extension results in localized amplification of single molecules in millions of unique locations across the flow cell surface. This process occurs in what is referred to as Illumina's "cluster station", an automated flow cell processor.

From Mardis 2008. Annual Rev. Genetics 9: 387.

Illumina Step 3: Sequencing



First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

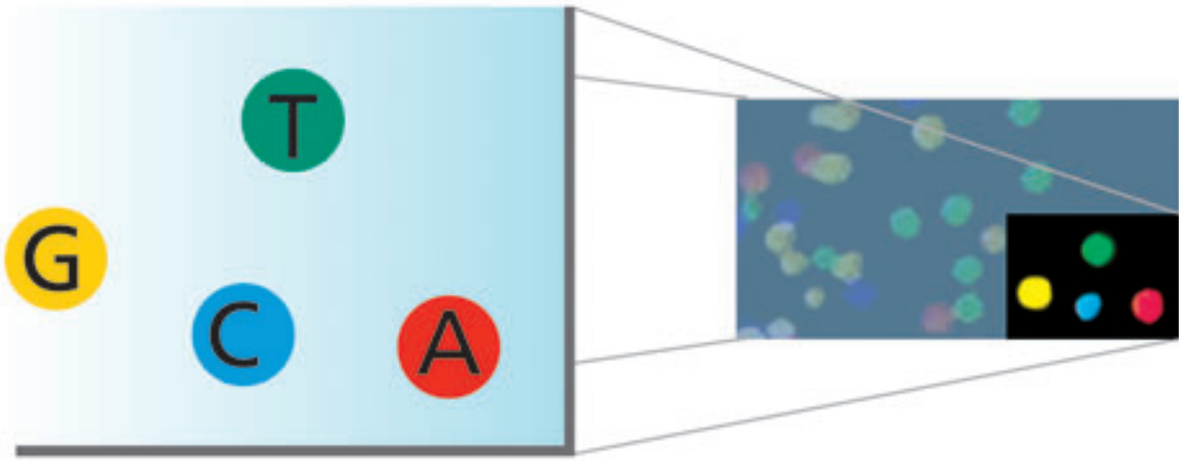


Image of first chemistry cycle

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

The blocked 3' terminus and the fluorophore from each incorporated base are removed.

From Mardis 2008. Annual Rev. Genetics 9: 387.

From : http://seqanswers.com/forums/showthread.php?t=21. Steps 7-12: Sequencing by Synthesis. A flow cell containing millions of unique clusters is now loaded into the 1G sequencer for automated cycles of extension and imaging. The first cycle of sequencing consists first of the incorporation of a single fluorescent nucleotide, followed by high resolution imaging of the entire flow cell. These images represent the data collected for the first base. Any signal above background identifies the physical location of a cluster (or polony), and the fluorescent emission identifies which of the four bases was incorporated at that position. This cycle is repeated, one base at a time, generating a series of images each representing a single base extension at a specific cluster. Base calls are derived with an algorithm that identifies the emission color over time. At this time reports of useful Illumina reads range from 26-50 bases.

Illumina Today …

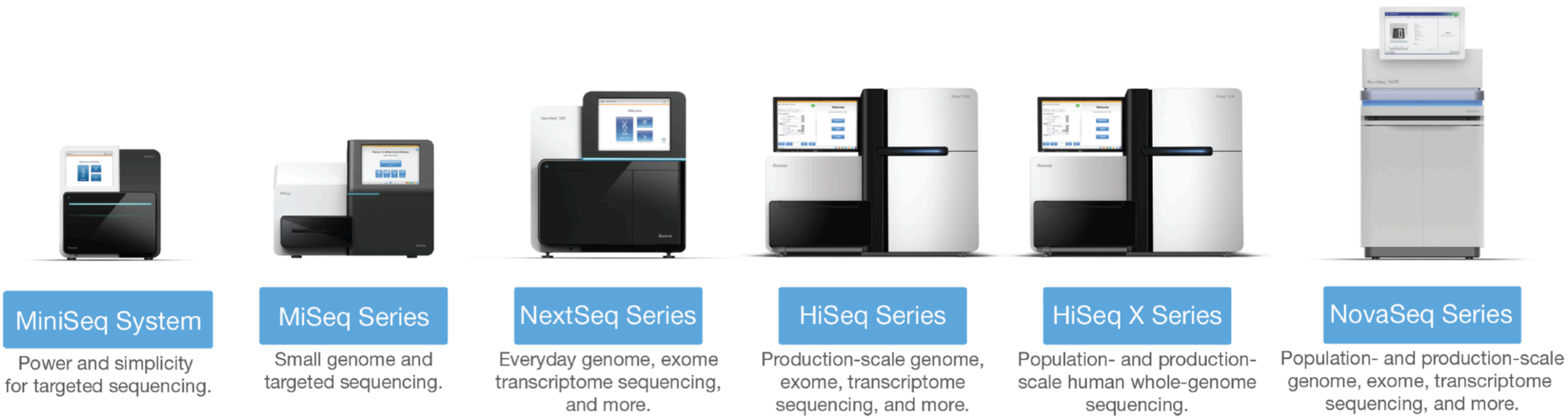


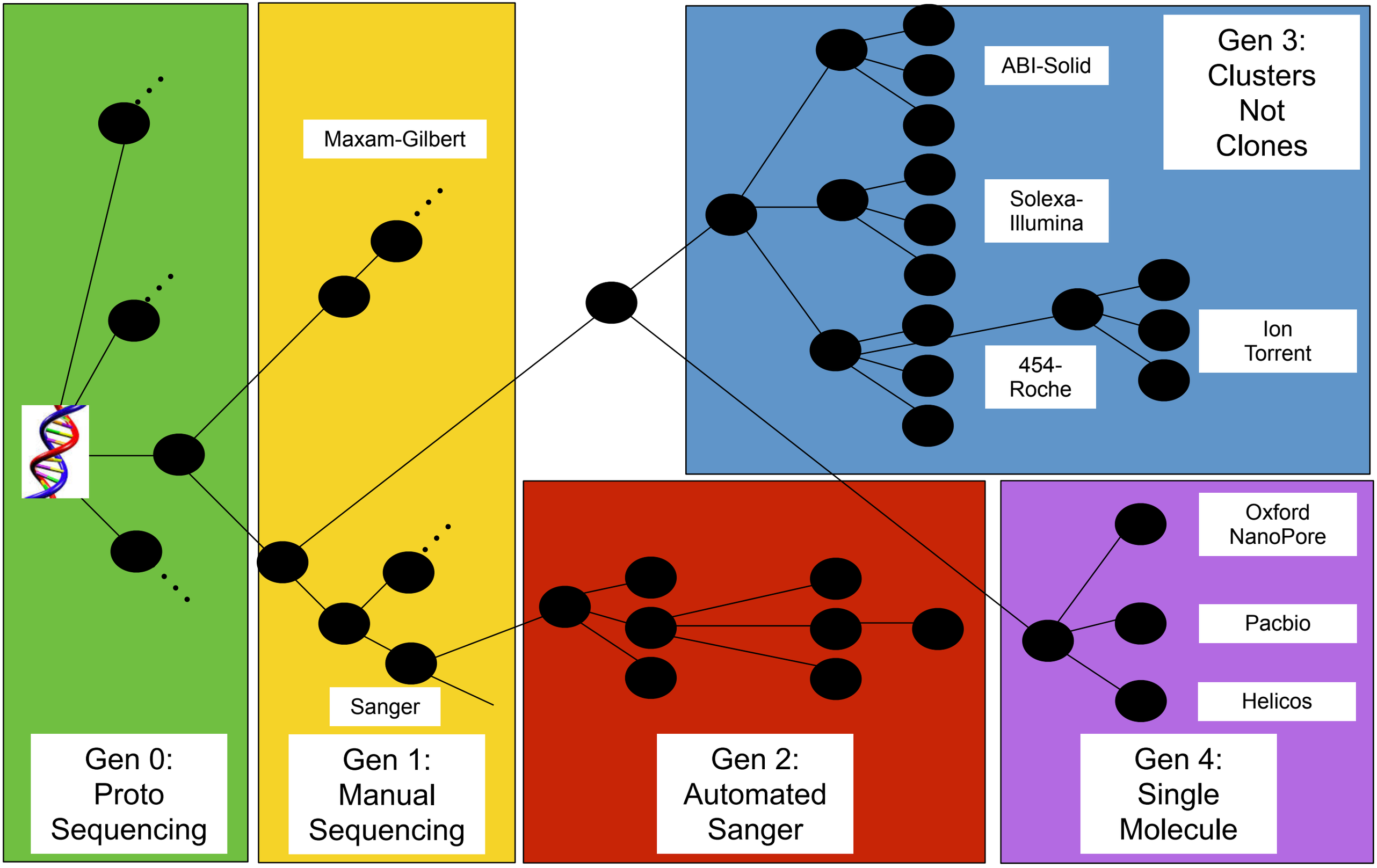
Figure 6: Sequencing Systems for Virtually Every Scale—Illumina offers innovative NGS platforms that deliver exceptional data quality and accuracy over a wide scale, from small benchtop sequencers to production-scale sequencing systems.

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

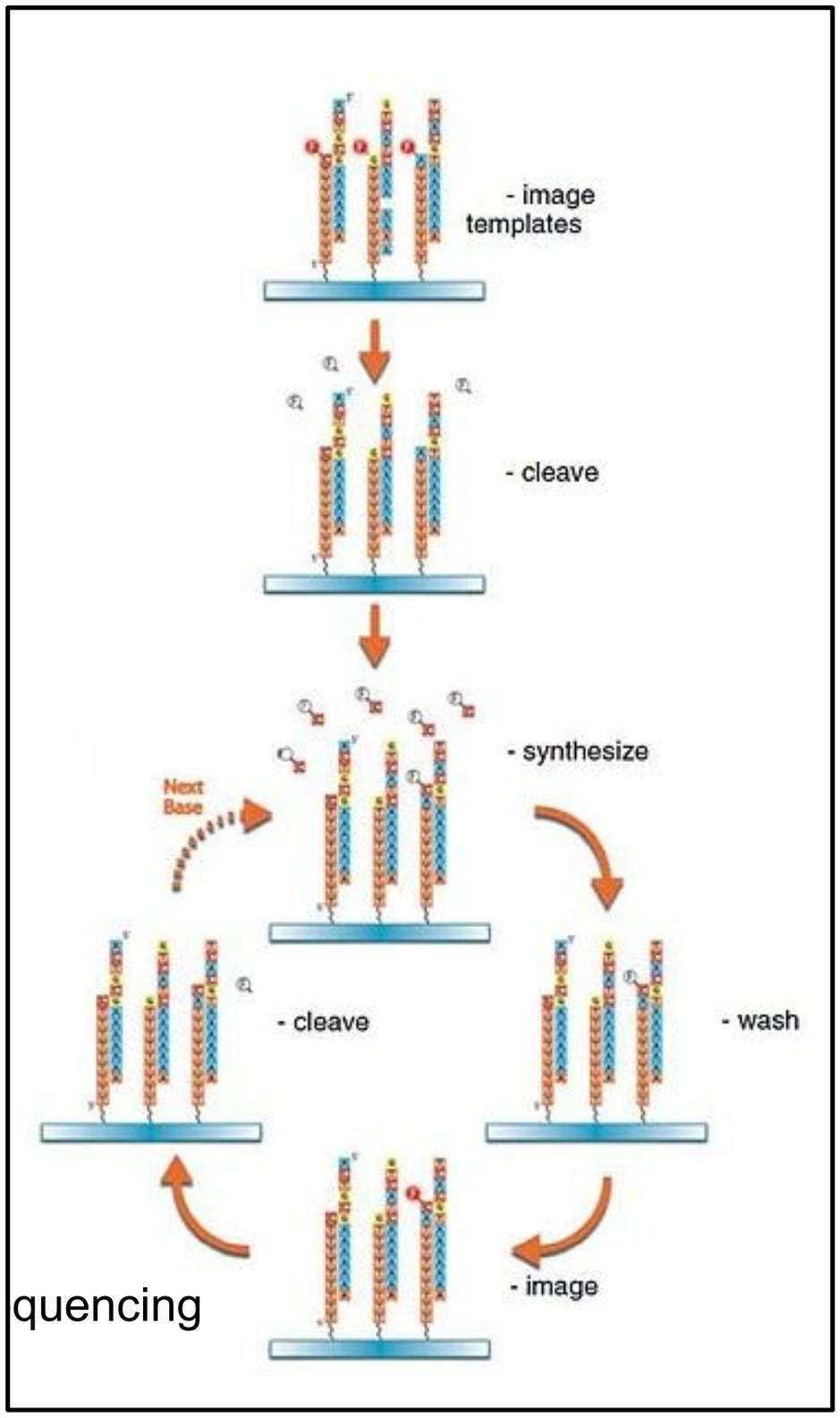
Some Key Innovations for Generation 3

- Diverse cluster creation methods
- Better microscopes, computers to process data
- Barcoding

Generation 4: Single Molecule



Single Molecule I: Helicos



Single Molecule II: Pacific Biosciences

Single Molecule II: Pacific Biosciences

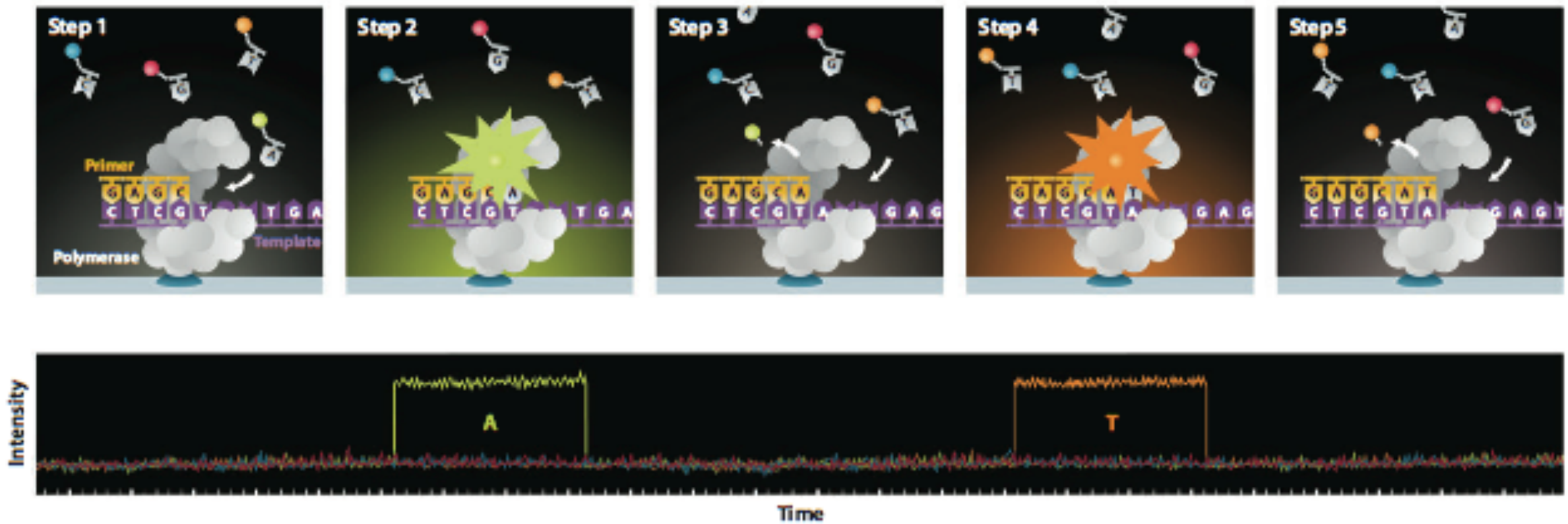


Figure 5

Single-molecule sequencing using Pacific Biosciences' zero-mode waveguides.

Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013;6:287-303.

Single Molecule II: Pacific Biosciences

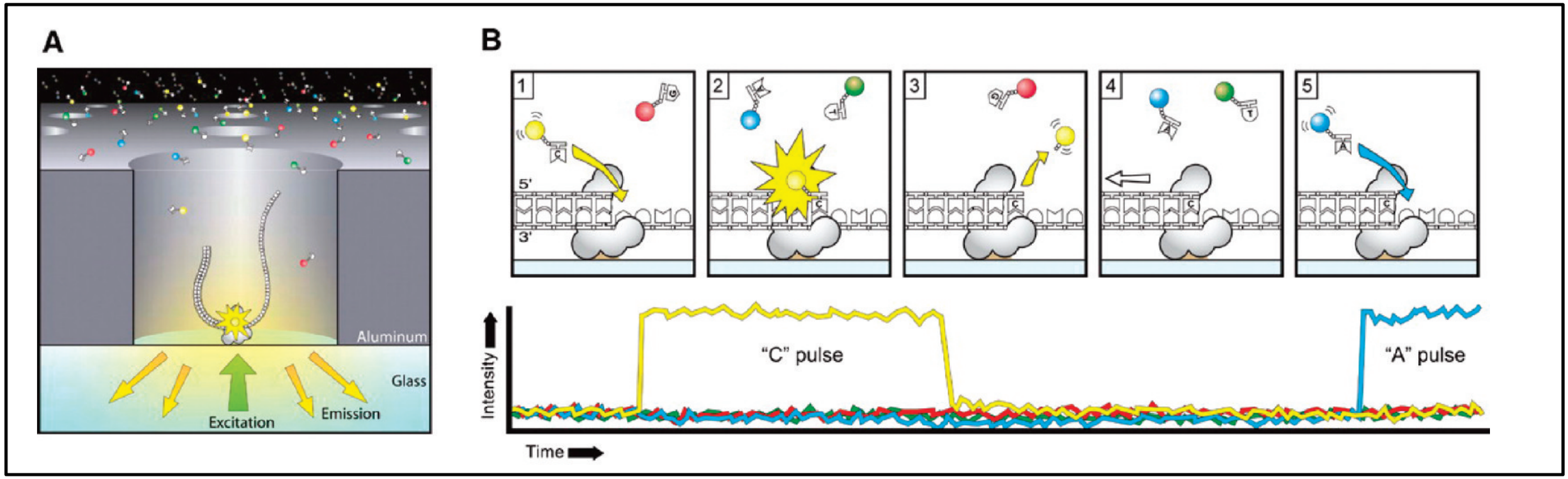
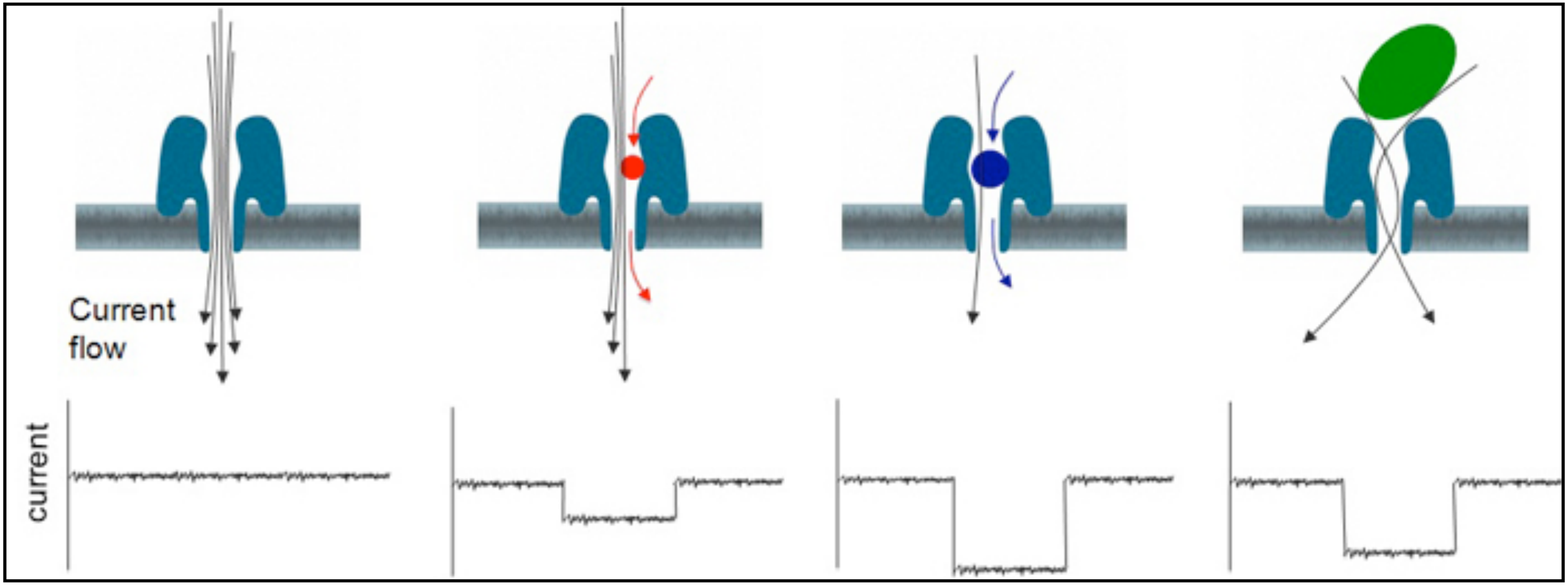


Figure 2. Schematic of PacBio's real-time single molecule sequencing. (A) The side view of a single ZMW nanostructure containing a single DNA polymerase (Φ29) bound to the bottom glass surface. The ZMW and the confocal imaging system allow fluorescence detection only at the bottom surface of each ZMW. (B) Representation of fluorescently labeled nucleotide substrate incorporation on to a sequencing template. The corresponding temporal fluorescence detection with respect to each of the five incorporation steps is shown below.

From Niedringhaus et al. Analytical Chemistry 83: 4327. 2011.

Single Molecule III: Oxford Nanopores



This diagram shows a protein nanopore set in an electrically resistant membrane bilayer. An ionic current is passed through the nanopore by setting a voltage across this membrane. If an analyte passes through the pore or near its aperture, this event creates a characteristic disruption in current. By measuring that current it is possible to identify the molecule in question. For example, this system can be used to distinguish the four standard DNA bases and G, A, T and C, and also modified bases. It can be used to identify target proteins, small molecules, or to gain rich molecular information for example to distinguish the enantiomers of ibuprofen or molecular binding dynamics.

From Oxford Nanopores Web Site

Single Molecule III: Oxford Nanopores

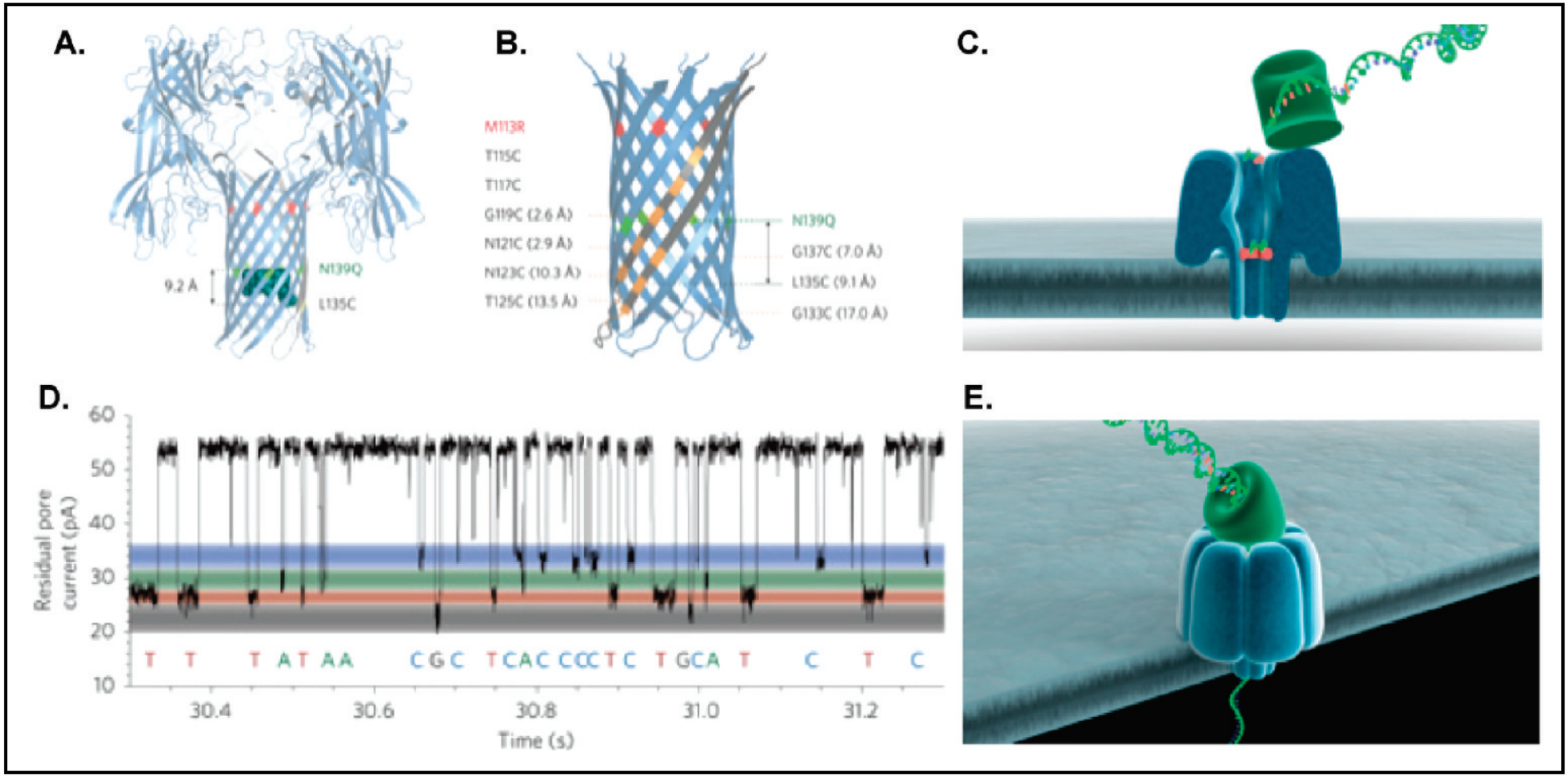
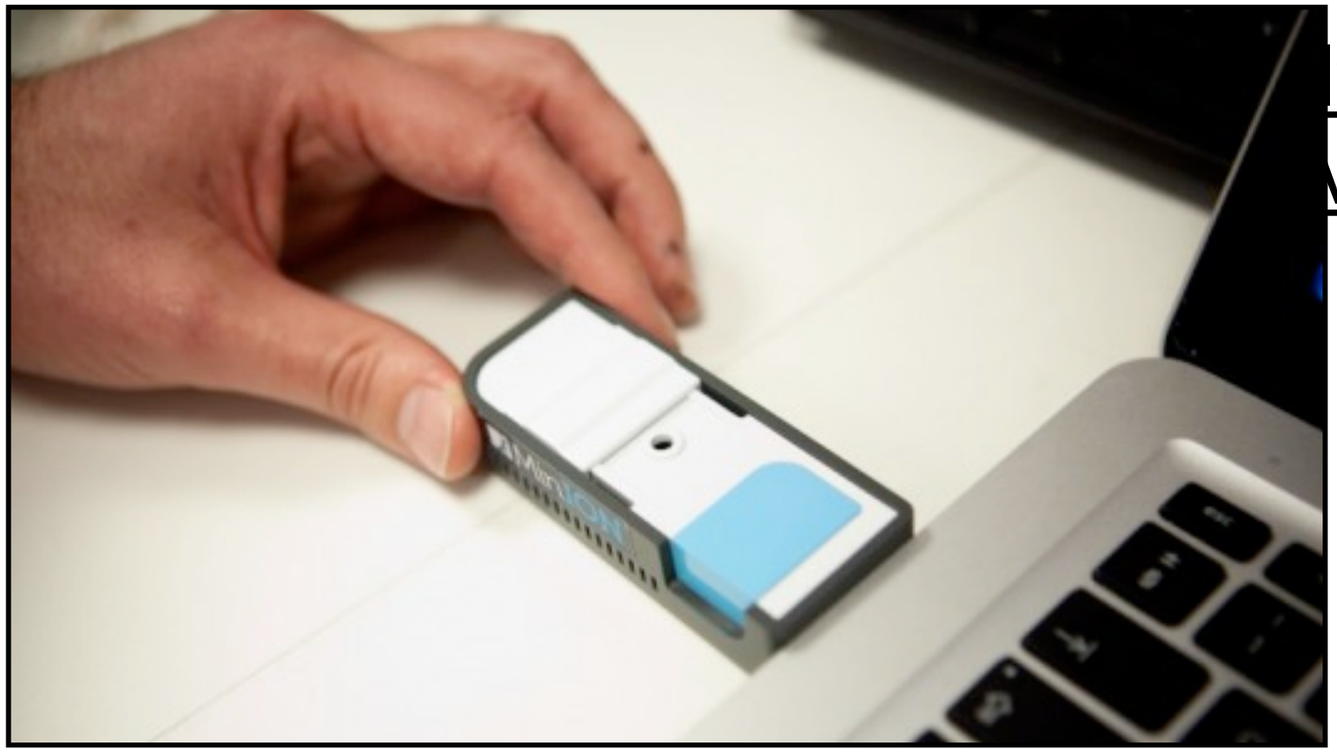


Figure6. Biological nanopores cheme employed by Oxford Nanopore.(A) Schematic of RHL protein nano pore mutant depicting the positions of the cyclodextrin (at residue 135) and glutamines (at residue 139). (B) A detailed view of the β barrel of the mutant nanopore shows the locations of the arginines (at residue 113) and the cysteines. (C) Exonuclease sequencing: A processive enzyme is attached to the top of the nanopore to cleave single nucleotides from the target DNA strand and pass them through the nanopore. (D) A residual current-vs-time signal trace from an RHL protein nanopore that shows a clear discrimination between single bases (dGMP, dTMP, dAMP, and dCMP). (E) Strand sequencing: ssDNA is threaded through a protein nanopore and individual bases are identified, as the strand remains intact. Panels A, B, and D reprinted with permission from ref 91. Copyright 2009 Nature Publishing Group. Panels C and E reprinted with permission from Oxford Nanopore Technologies (Zoe McDougall).

From Niedringhaus et al. Analytical Chemistry 83: 4327. 2011.

Single Molecule III: Oxford Nanopores



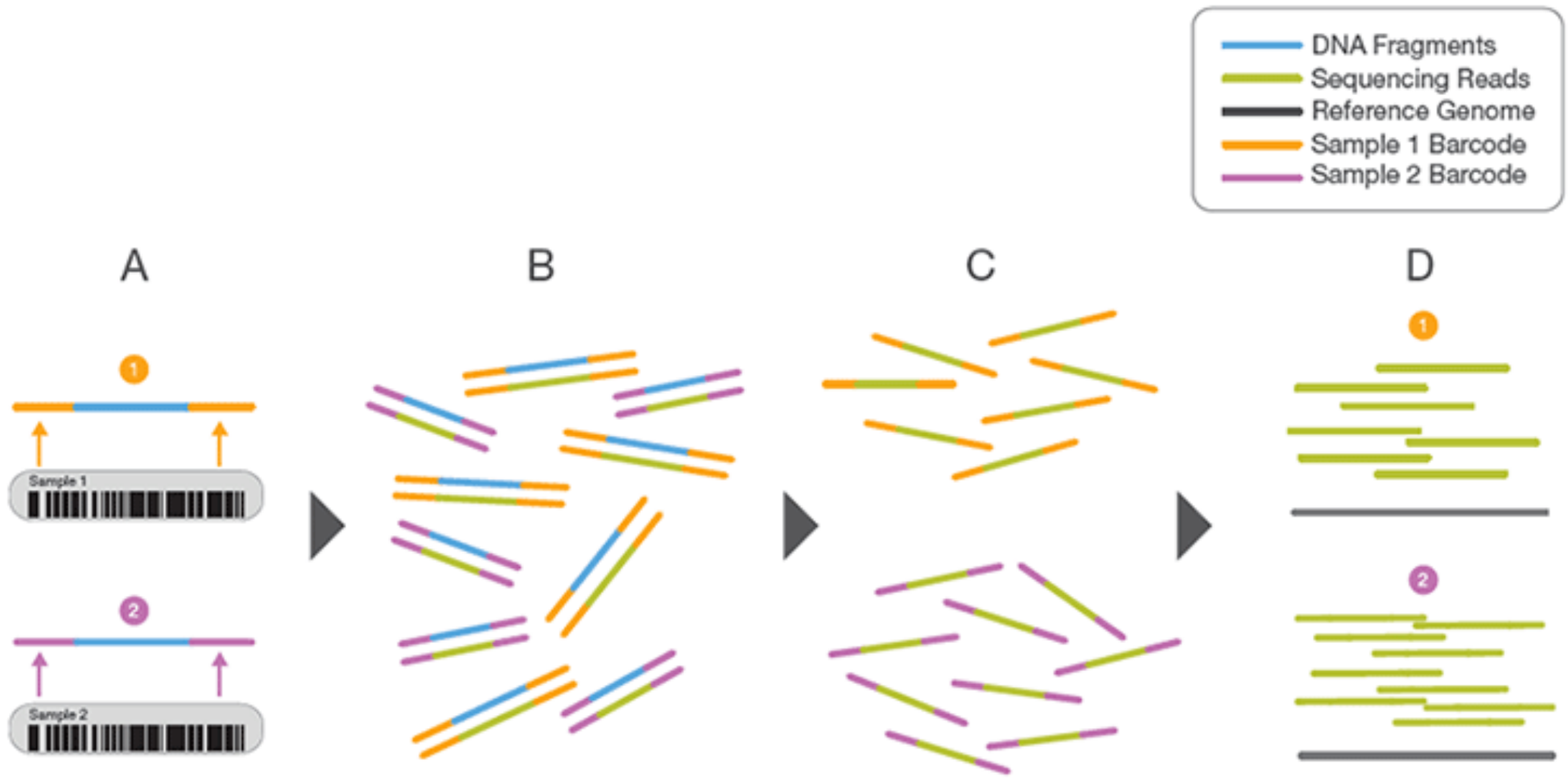
“It’s kind of a cute device,” says Jaffe of the MinION, which is roughly the size and shape of a packet of chewing gum. “It has pretty lights and a fan that hums pleasantly, and plugs into a USB drive.” But his technical review is mixed. From http://www.nature.com/news/data-from-pocket-sized-genome-sequencer-unveiled-1.14724

Bells and Whistles

- Multiplexing and barcoding
- Small amounts of DNA
- Capture methods
- Paired end
- HiC
- Modified bases

Multiplexing

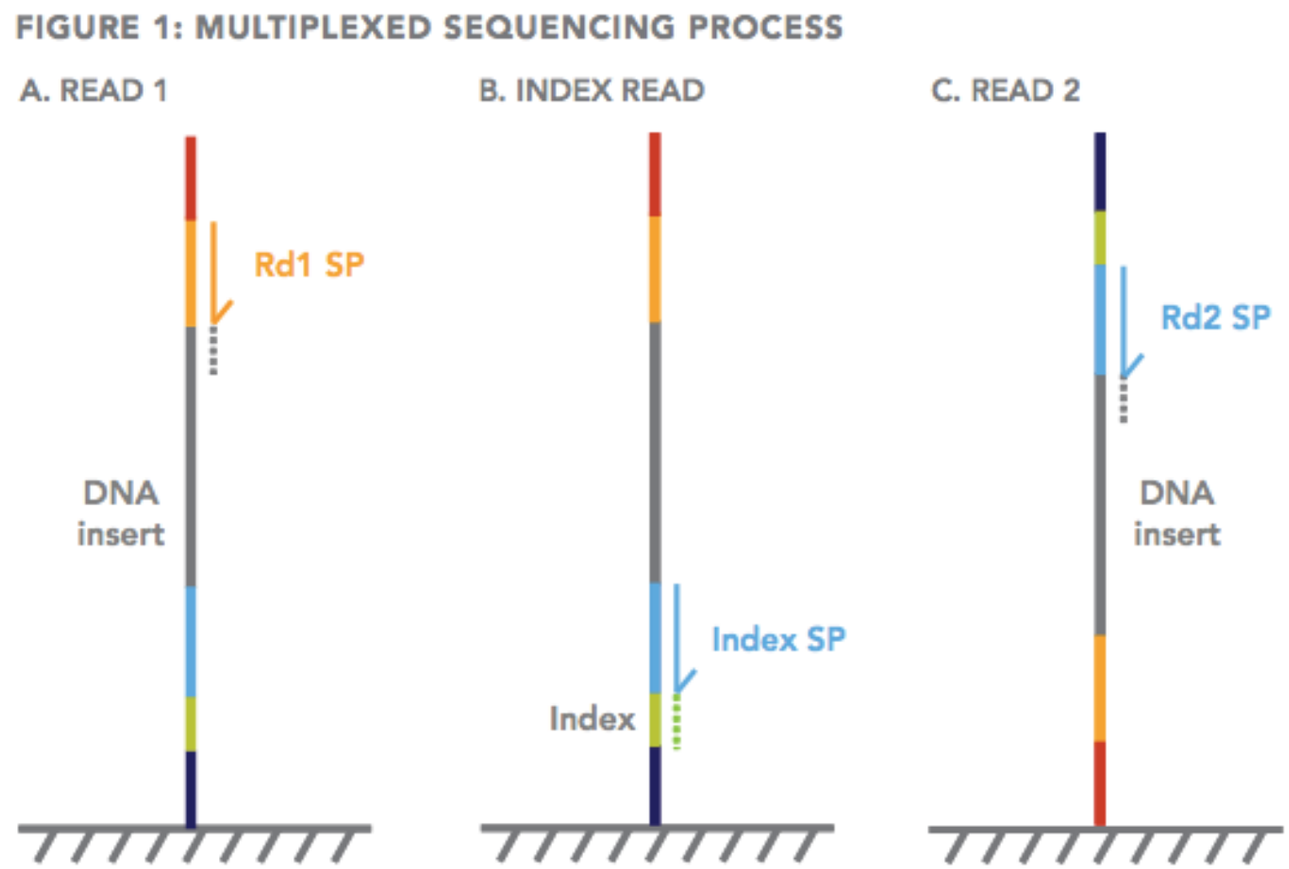
Figure 2: Conceptual Overview of Sample Multiplexing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

From http://www.illumina.com/technology/multiplexing_sequencing_assay.ilmn

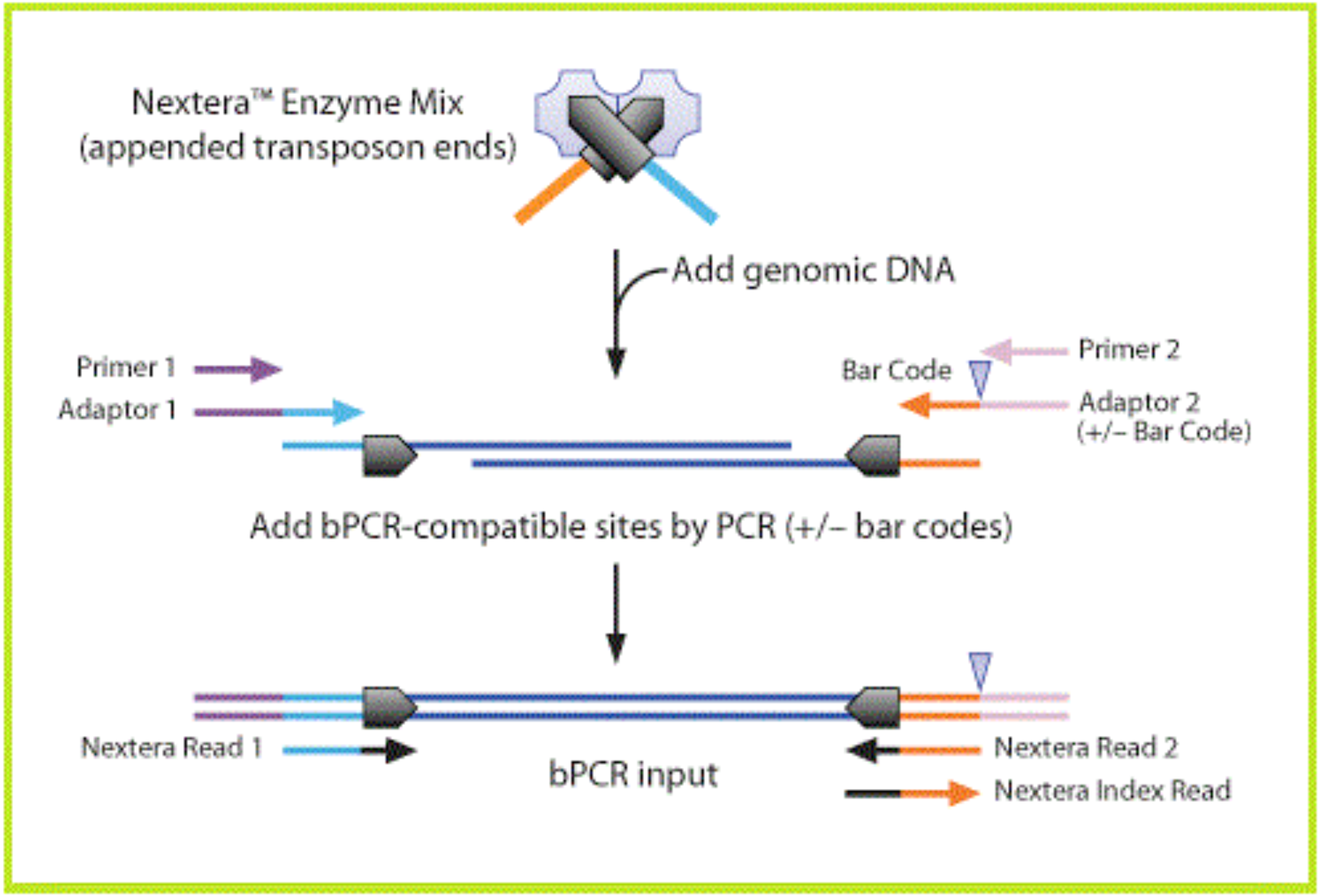
Multiplexing



Sample multiplexing involves a total of three sequencing reads, including a separate index read, which is generated automatically on the Genome Analyzer equipped with the Paired-End Module. A: Application read 1 (dotted line) is generated using the Read 1 Sequencing Primer (Rd1 SP). B: The read 1 product is removed and the Index Sequencing Primer (Index SP) is annealed to the same strand to produce the 6-bp index read (dotted line). C: If a paired-end read is required, the original template strand is used to regenerate the complementary strand. Then, the original strand is removed and the complementary strand acts as a template for application read 2 (dotted line), primed by the Read 2 Sequencing Primer (Rd2 SP). Pipeline Analysis software identifies the index sequence from each cluster so that the application reads can be assigned to a single sample. Hatch marks represent the flow cell surface.

http://res.illumina.com/documents/products/datasheets/datasheet_sequencing_multiplex.pdf

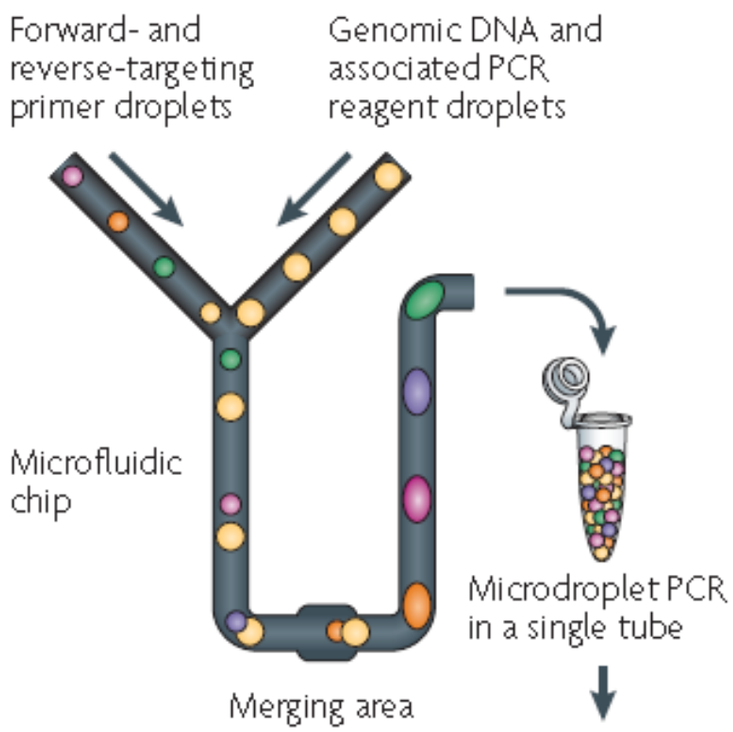
Small Amounts of DNA



http://www.epibio.com/docs/default-source/protocols/nextera-dna-sample-prep-kit-(illumina--compatible).pdf?sfvrsn=4

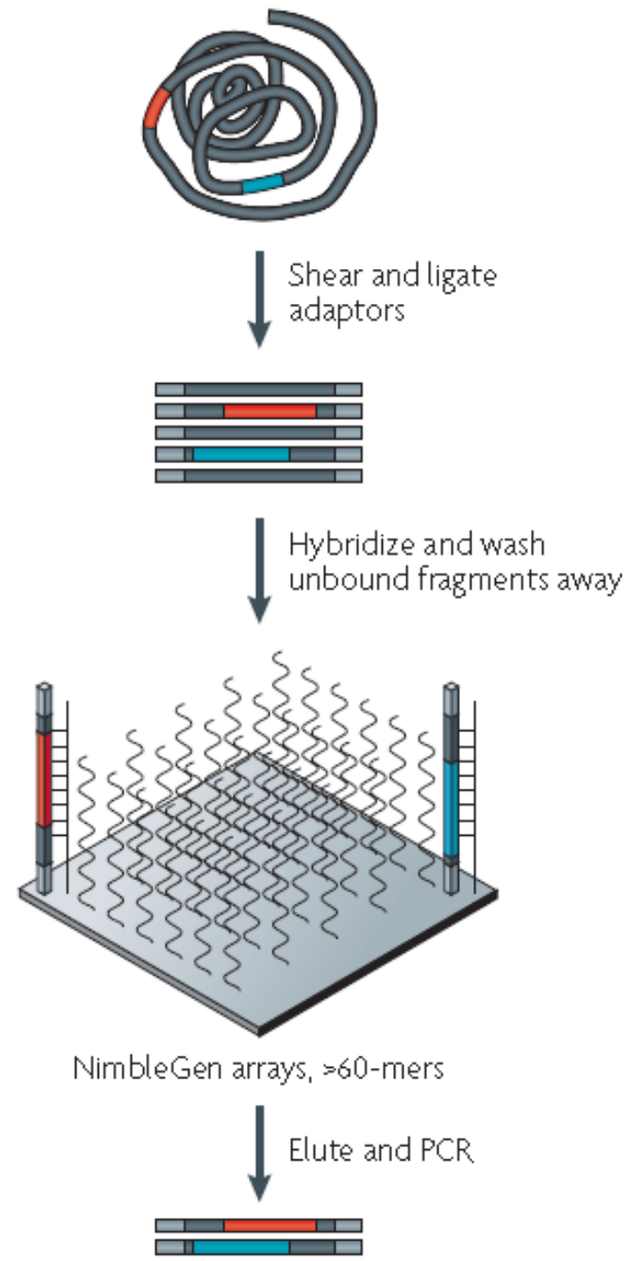
Capture Methods

RainDance
Microdroplet PCR



Reported 84% of capture efficiency

Roche Nimblegen
Salid-phase capture with custom-designed oligonucleotide microarray

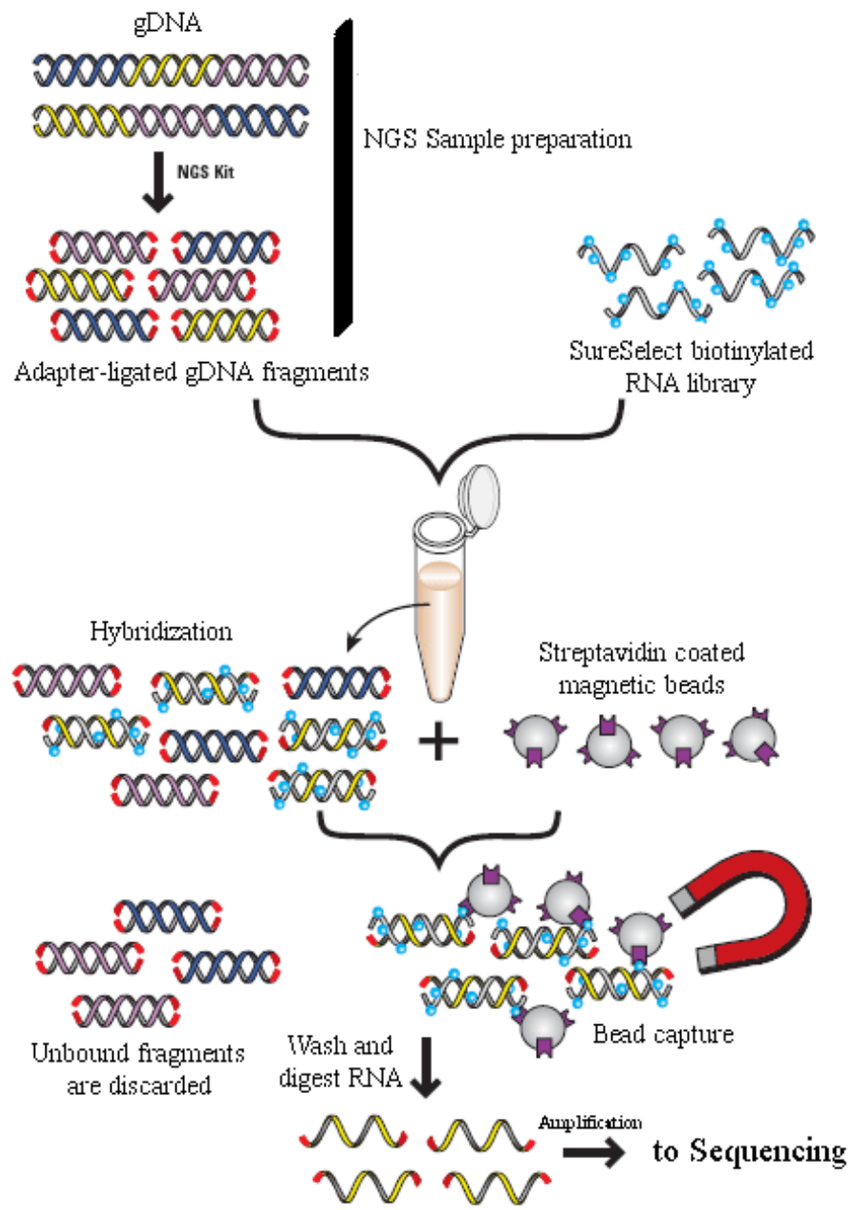


Reported 65-90% of capture efficiency

From Slideshare presentation of Cosentino Cristian
http://www.slideshare.net/cosentia/high-throughput-equencing

Capture Methods

Agilent SureSelect
Solution-phase capture with
streptavidin-coated magnetic beads

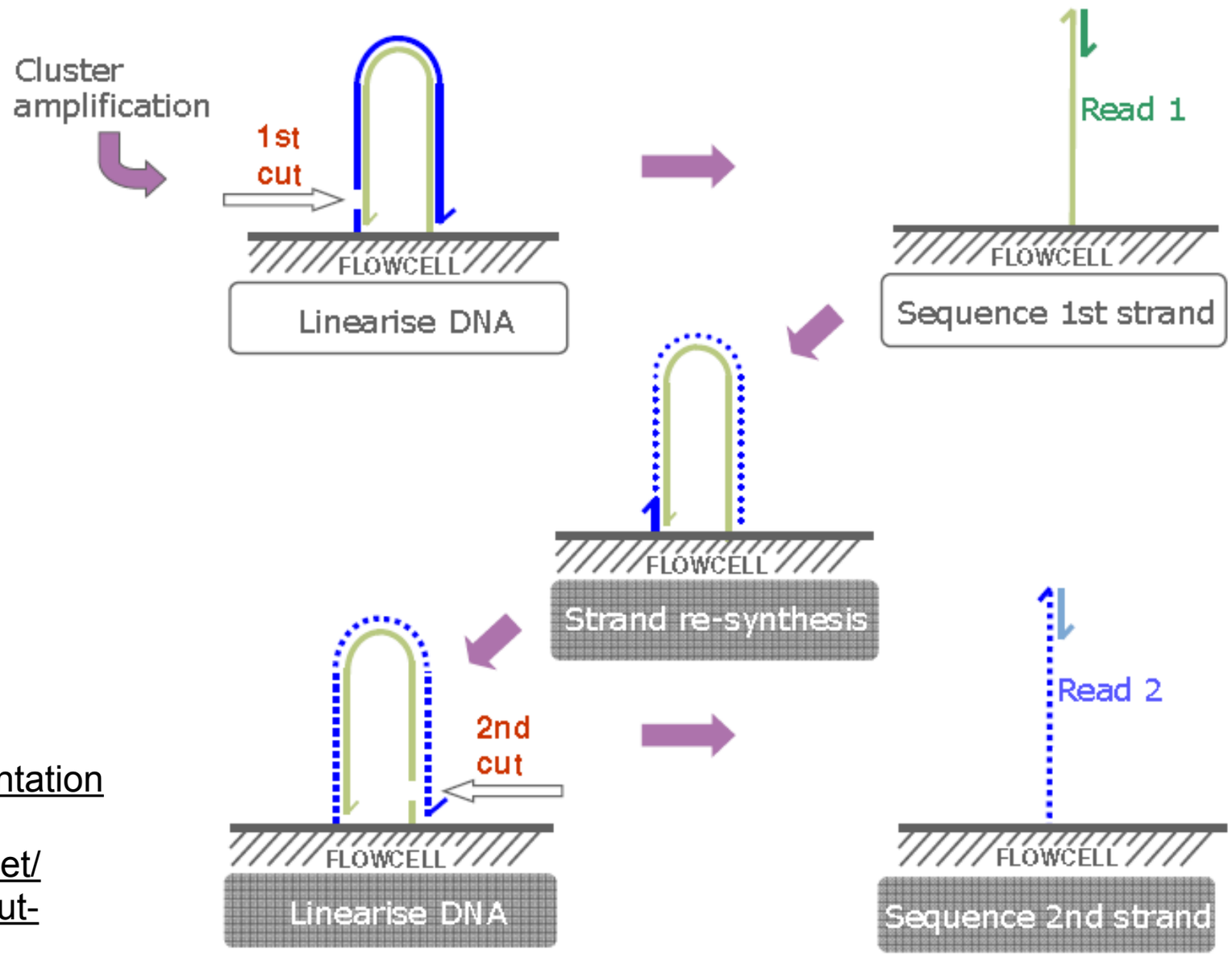


Reported 60-80% of capture efficiency

From Slideshare presentation of Cosentino Cristian http://www.slideshare.net/cosentia/high-throughput-equencing

Illumina Paired Ends

Paired-end sequencing works into GA and uses chemicals from the PE module to perform cluster amplification of the reverse strand



From Slideshare presentation of Cosentino Cristian http://www.slideshare.net/cosentia/high-throughput-equencing

HiC

Published in final edited form as:
Methods. 2012 November ; 58(3): . doi:10.1016/j.ymeth.2012.05.001.

Hi-C: A comprehensive technique to capture the conformation of genomes

Jon-Matthew Belton[1], **Rachel Patton McCord**[1], **Johan Gibcus**[1], **Natalia Naumova**[1], **Ye Zhan**[1], and **Job Dekker**[1,*]

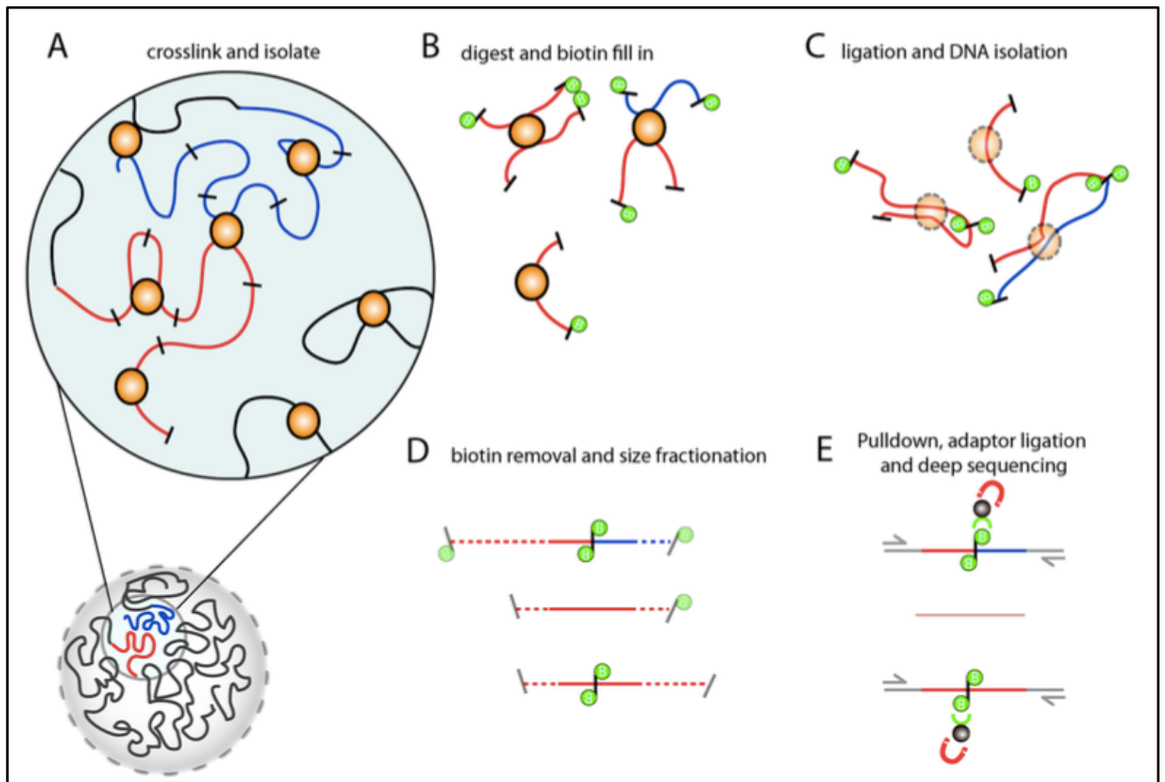


Figure 1. Overview of Hi-C technology
A) Hi-C detects chromatin interaction both within and between chromosomes by covalently crosslinking protein/DNA complexes with formaldehyde. **B)** The chromatin is digested with a restriction enzyme and the ends are marked with a biotinylated nucleotide. **C)** The DNA in the crosslinked complexes are ligated to form chimeric DNA molecules. **D)** Biotin is removed from the ends of linear fragments and the molecules are fragmented to reduce their overall size. **E)** Molecules with internal biotin incorporation are pulled down with streptavidin coated magnetic beads and modified for deep sequencing. Quantitation of chromatin interactions is achieved through massively parallel deep sequencing.

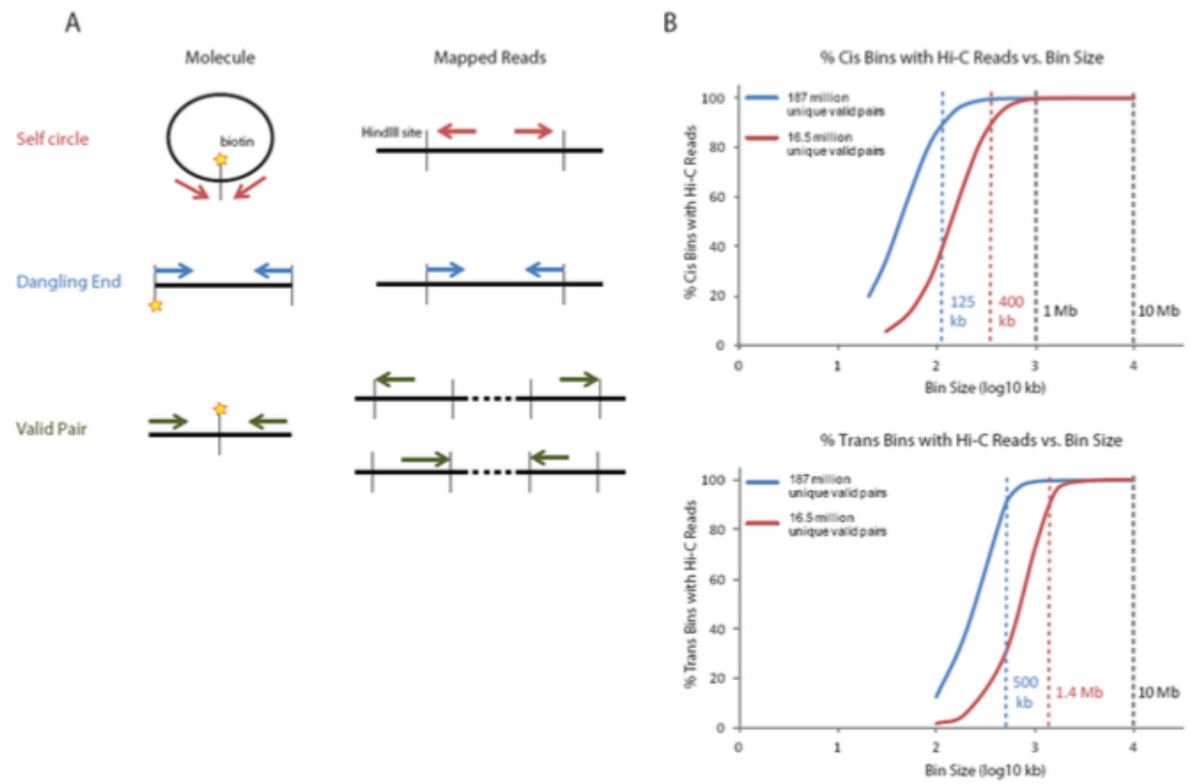
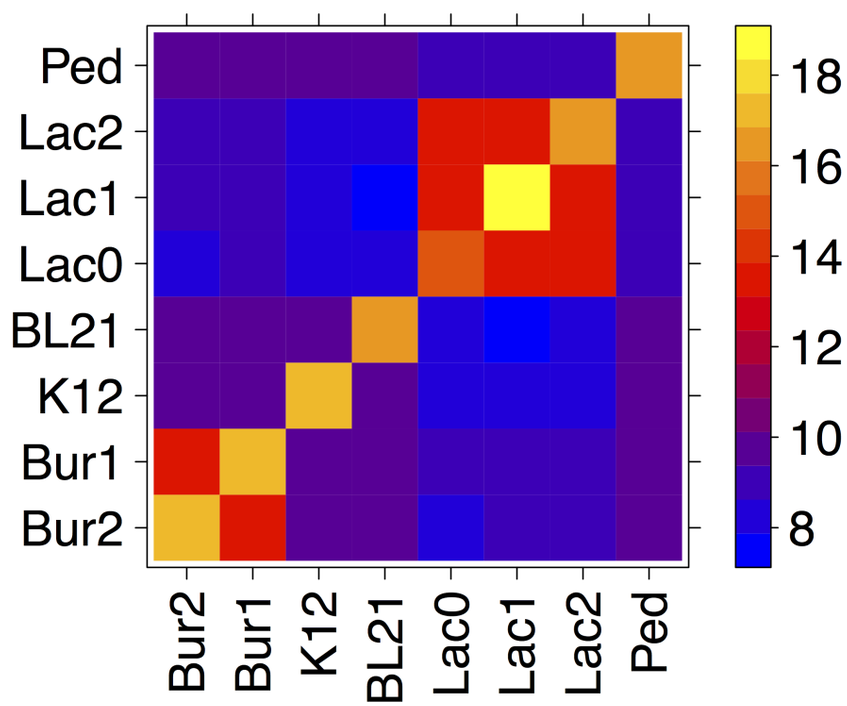


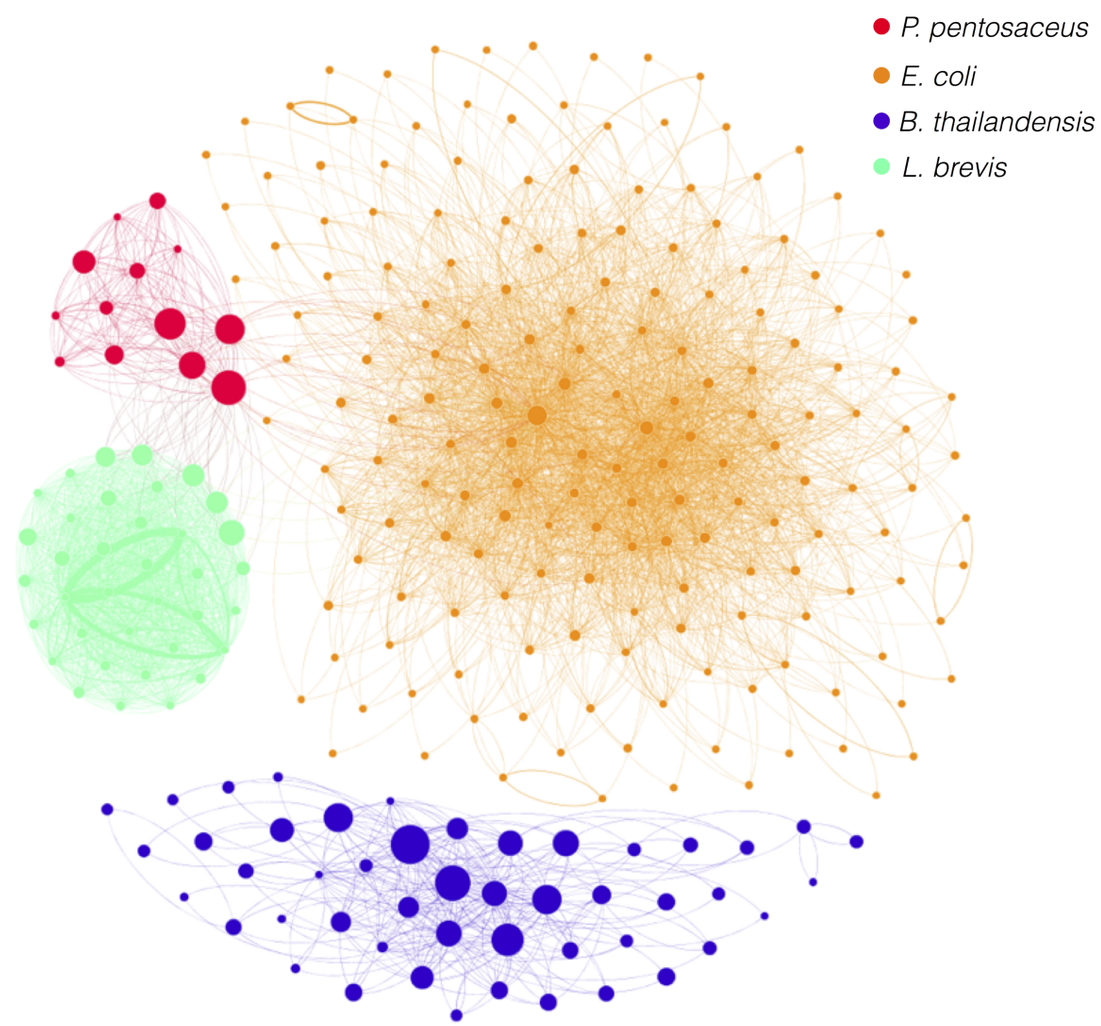
Fig 3. Hi-C sequence mapping and binning considerations
A) Different types of molecules in the Hi-C library (left) lead to different orientations of mapped reads relative to restriction sites (right). Mapped reads (colored arrows) facing outward in the same fragment come from self-circles (top); Reads facing inward in the same fragment arise from dangling ends (middle); Reads from different restriction fragments and facing toward a restriction site arise from valid interaction pairs (bottom). **B)** Relationship between sequencing depth and choice of bin size. Each graph shows the percentage of *cis* (top) or *trans* (bottom) bins that contain at least one mapped read from a valid interaction pair (y-axis) for each different bin size (x-axis). Colored dotted lines indicate the bin size at which 90% of bins contain at least one valid pair read for a Hi-C library with a high (blue) or low (red) number of total unique valid pairs after sequencing.

Suggested by Carlos Bustamante and Keith Bradnam

HiC Crosslinking & Sequencing

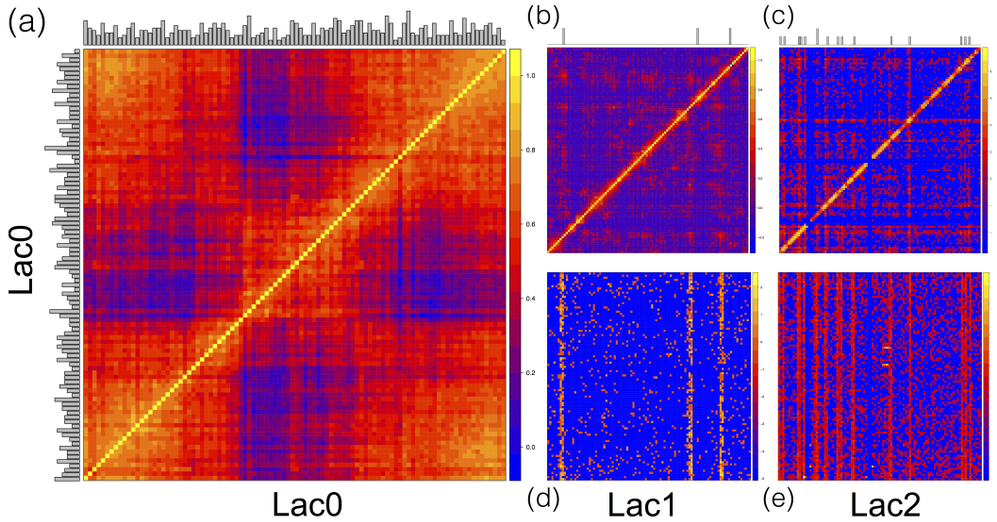
Sequence	Alignment	% of Total	Filtered	% of aligned	Length	GC	#R.S.
Lac0	10,603,204	26.17%	10,269,562	96.85%	2,291,220	0.462	629
Lac1	145,718	0.36%	145,478	99.84%	13,413	0.386	3
Lac2	691,723	1.71%	665,825	96.26%	35,595	0.385	16
Lac	11,440,645	28.23%	11,080,865	96.86%	2,340,228	0.46	648
Ped	2,084,595	5.14%	2,022,870	97.04%	1,832,387	0.373	863
BL21	12,882,177	31.79%	2,676,458	20.78%	4,558,953	0.508	508
K12	9,693,726	23.92%	1,218,281	12.57%	4,686,137	0.507	568
E. coli	22,575,903	55.71%	3,894,739	17.25%	9,245,090	0.51	1076
Bur1	1,886,054	4.65%	1,797,745	95.32%	2,914,771	0.68	144
Bur2	2,536,569	6.26%	2,464,534	97.16%	3,809,201	0.672	225
Bur	4,422,623	10.91%	4,262,279	96.37%	6,723,972	0.68	369





Chris Beitel
@datscimed

Aaron Darling
@koadman



Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen JA, Darling AE. (2014) Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ 2:e415 http://dx.doi.org/10.7717/peerj.415

Detecting Modified Bases

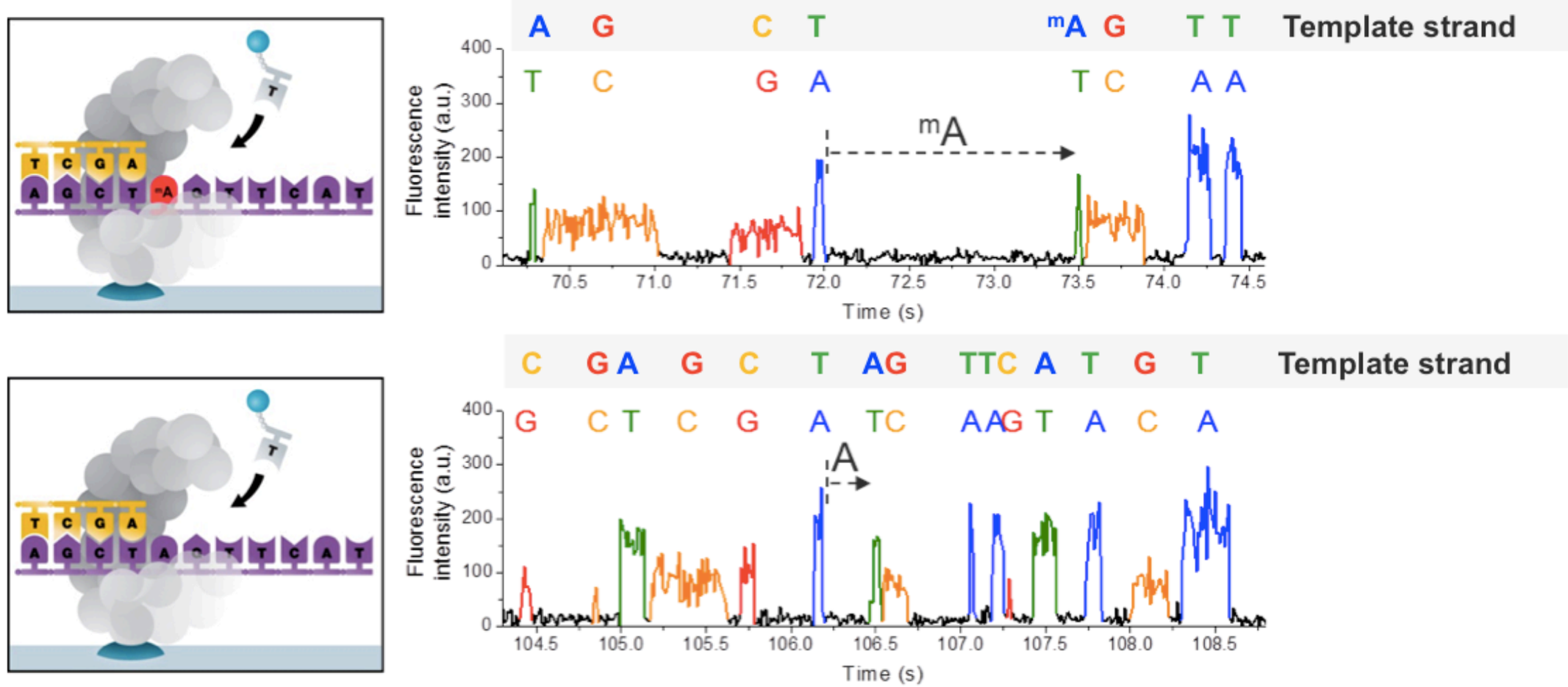


Figure 2. Principle of detecting modified DNA bases during SMRT sequencing. The presence of the modified base in the DNA template (top), shown here for 6-mA, results in a delayed incorporation of the corresponding T nucleotide, i.e. longer interpulse duration (IPD), compared to a control DNA template lacking the modification (bottom).[3]

http://www.pacificbiosciences.com/pdf/microbial_primer.pdf

Key Issues

- Cost / bp
- Read length
- Paired end approaches
- Ease of feeding
- Error profiles
- Barcoding and multiplexing potential

Evolution of Sequencing