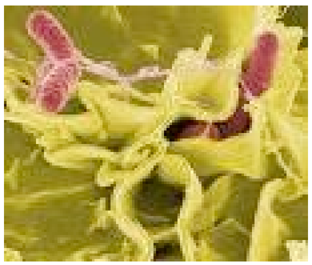
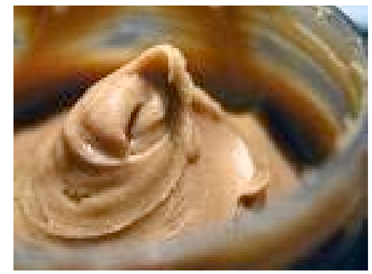
Introduction to Genome Sequencing, Metagenomics and Synthetic Biology

Bioinformatics and Genomics
Tara Gianoulis
Feb. 23, 2009

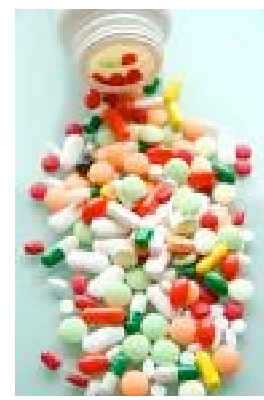
Basic Introduction



ATACGTGTTG
ATCTGTCGAT
ACGCTAGCTA
ACGTCAGTCA
ATACATCACG



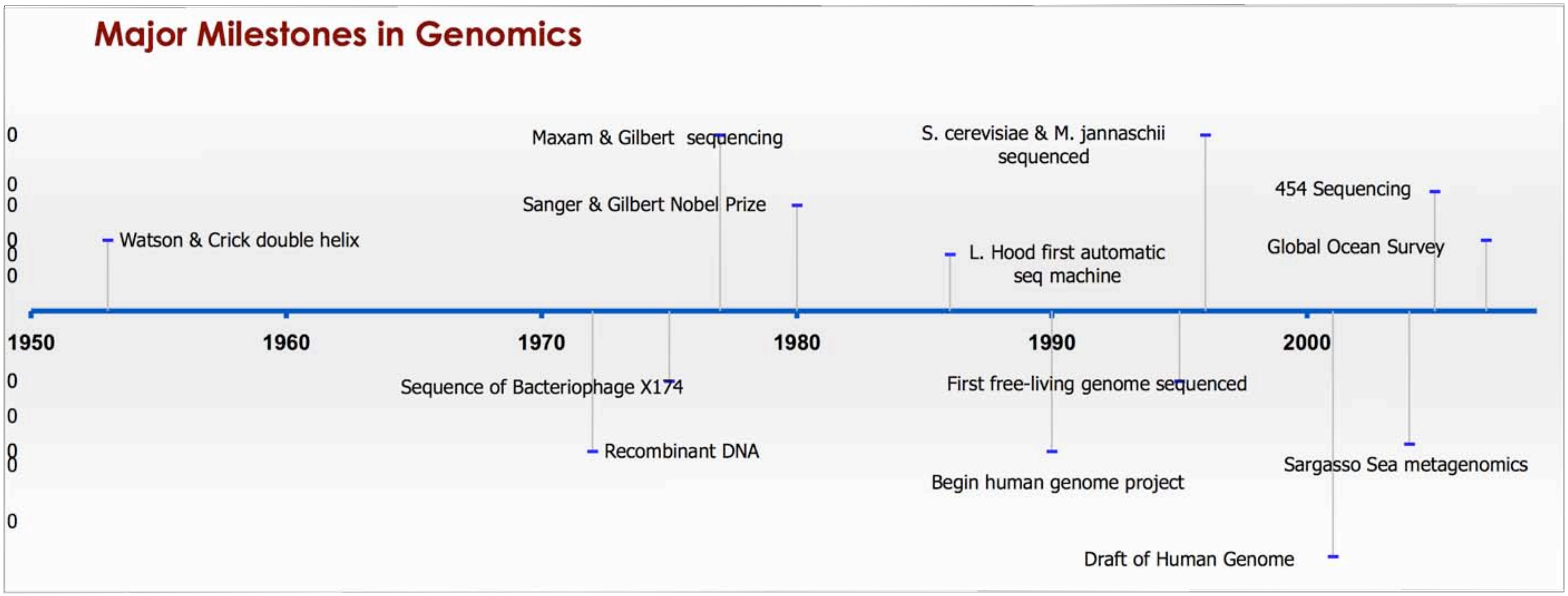
TACTGTGTGTTA
ATTACGTGTTGT
ATGCTGTCGATA
ACGGCTAGCTAA
ACGTCCCAGTCA
ATACATCGTCAC



Outline

- Overview
- Genome Sequencing
  - Strategies: Bottom Up, Top Down
  - Next Gen and Assembly (will be covered in considerable detail later in the course, Snyder and Noonan)
  - Examples
- Metagenomics
  - Why do we need it and what is it?
  - Comparative Metagenomics
  - Examples
- Synthetic Biology
  - Definitions
  - Applications to biofuels
- Summary and Additional Resources

Overview: Major Scientific Discoveries

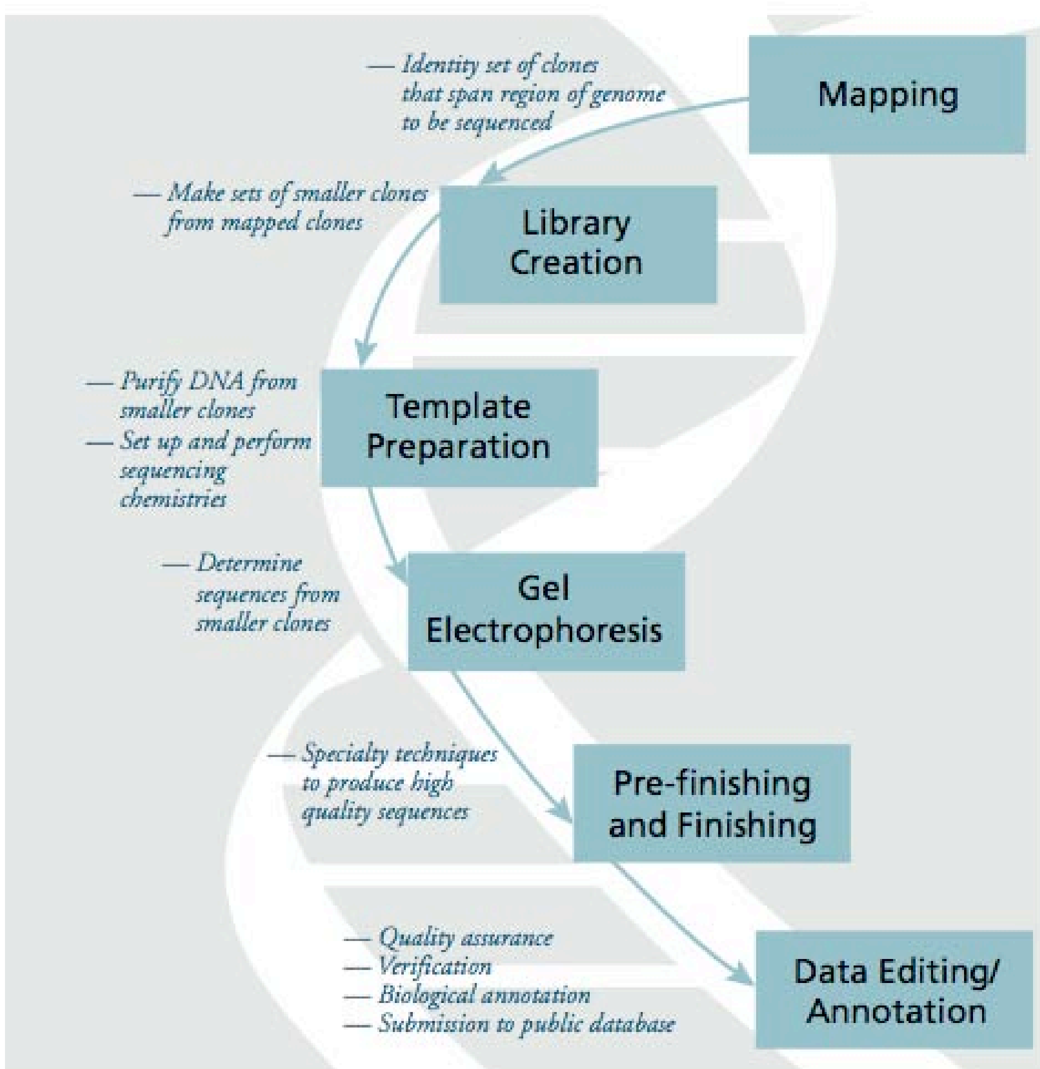


Gianoulis, 2009

Genome Sequencing: Main Strategies

What is the anatomy of a genome sequence project? What are the main strategies? What is the historical context? What types of questions can we answer with it? What are the computational challenges?

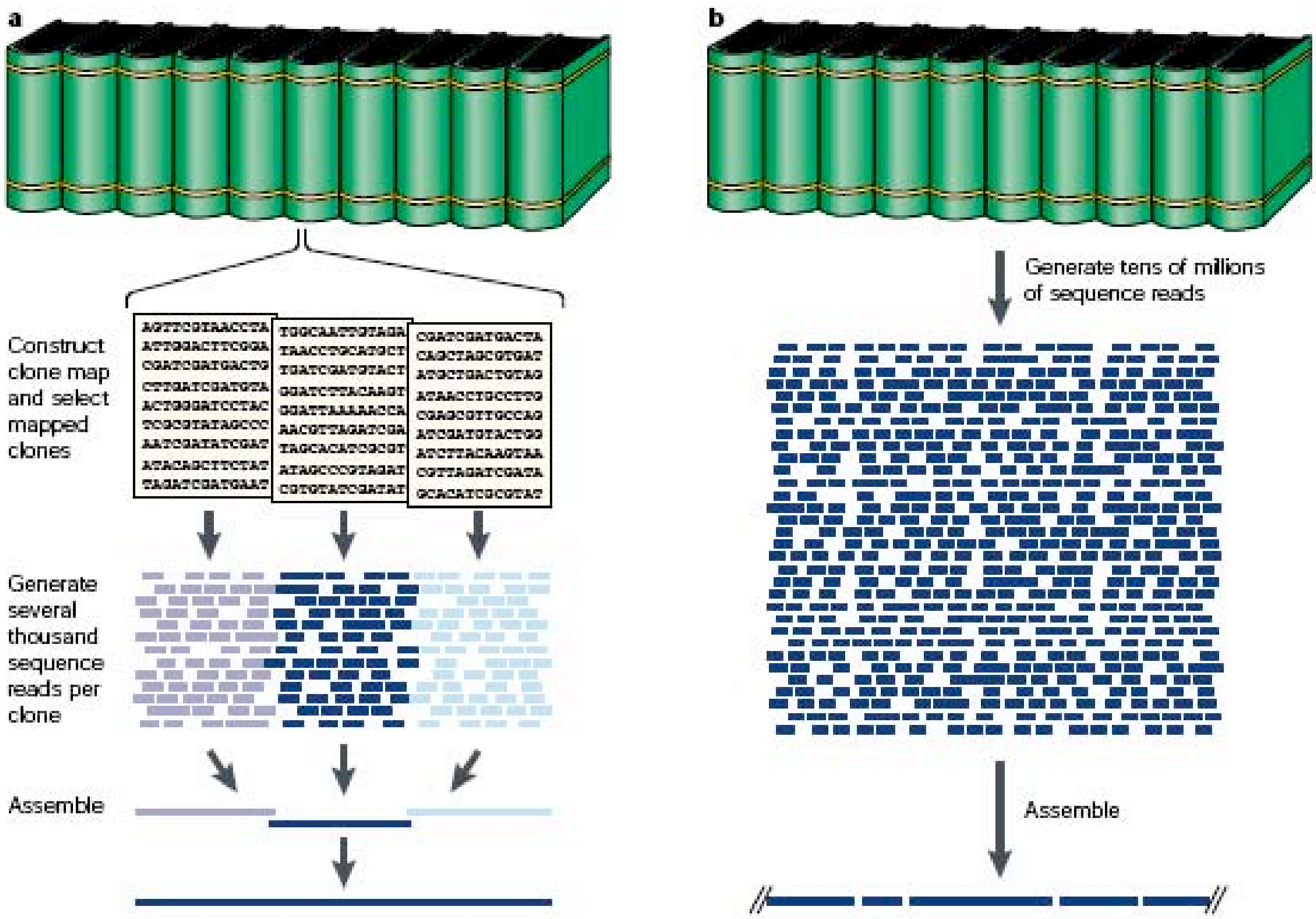
Anatomy of a Traditional Sequencing Project



DOE JGI

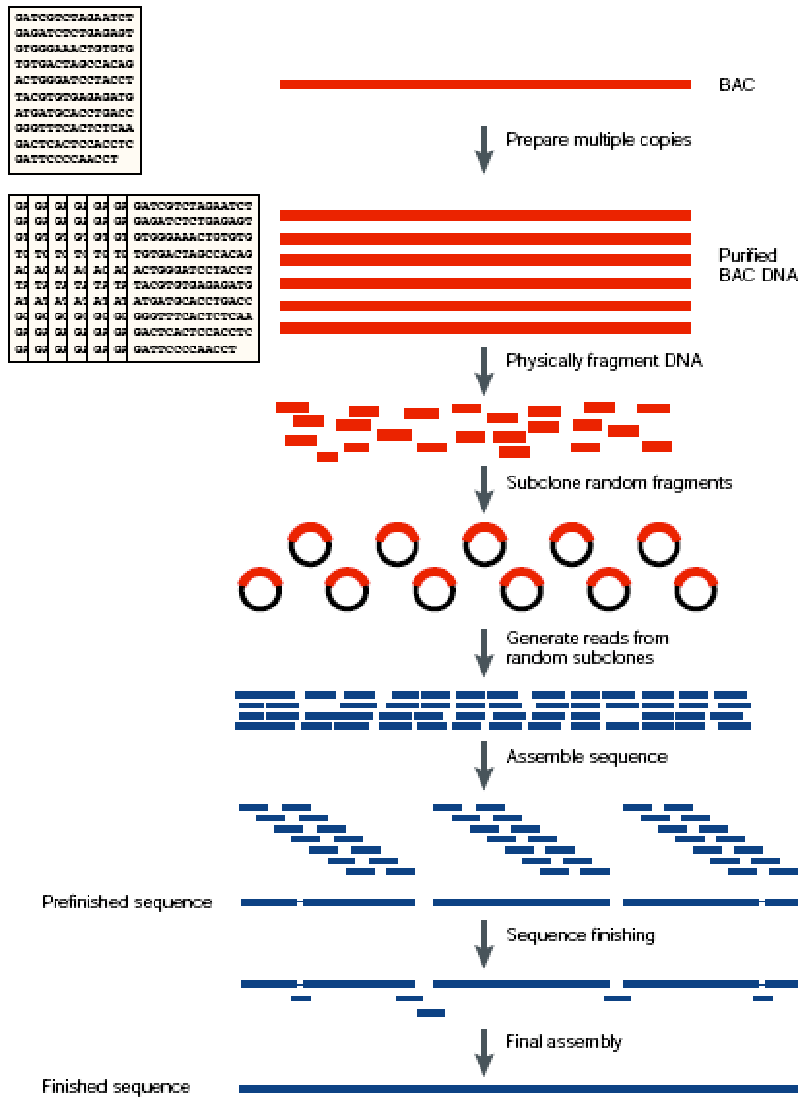
Conceptual: Two Sequencing Strategies

Top Down | Bottom Up



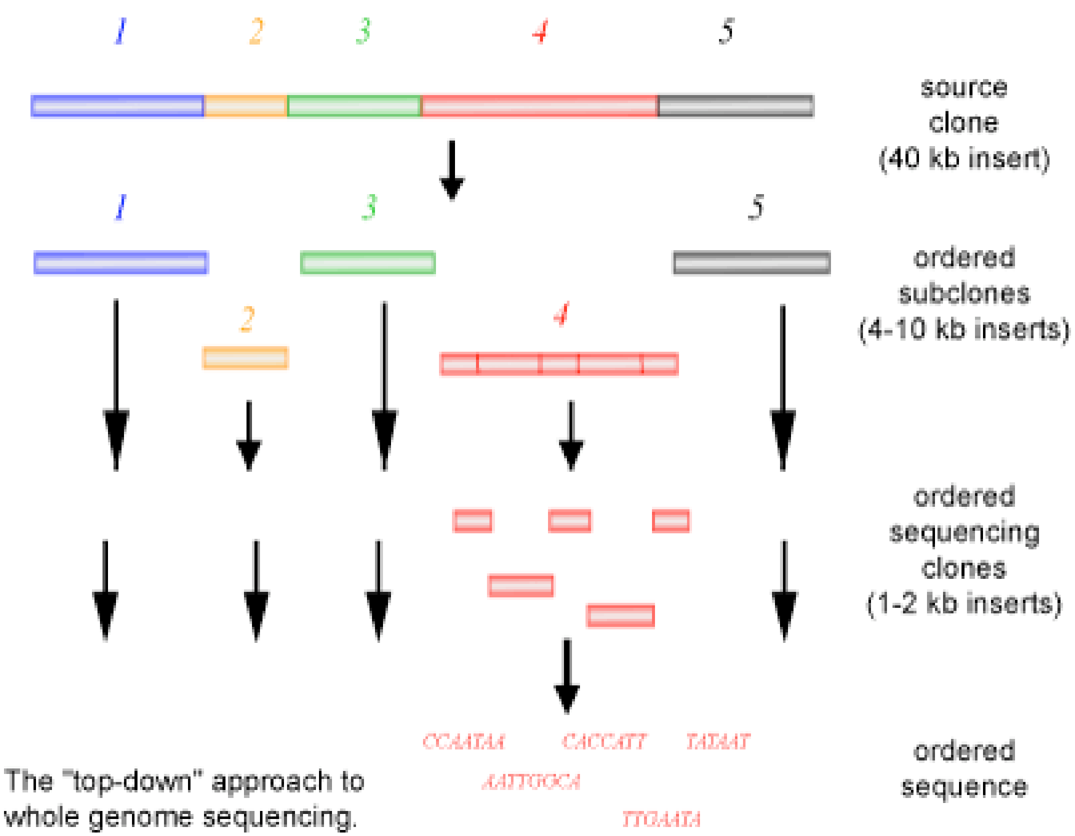
Green et al, Nat Rev 2001

Technical: Sequencing Strategies



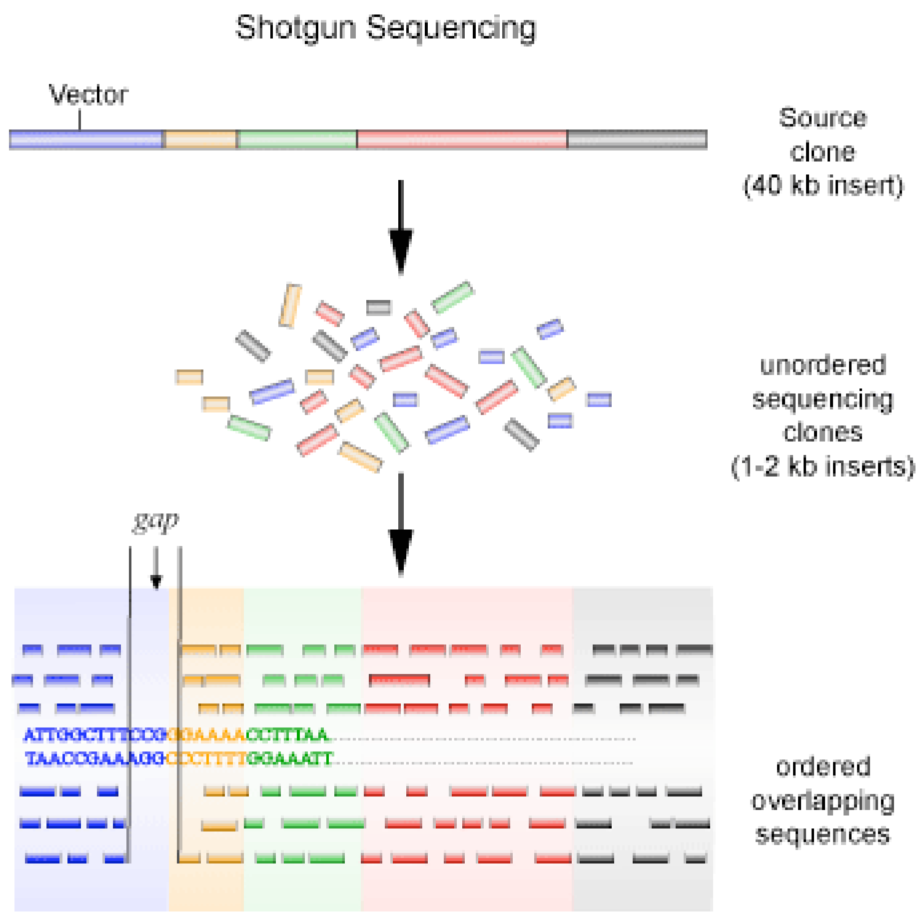
Green et al, Nat Rev 2001

Top-Down Approach



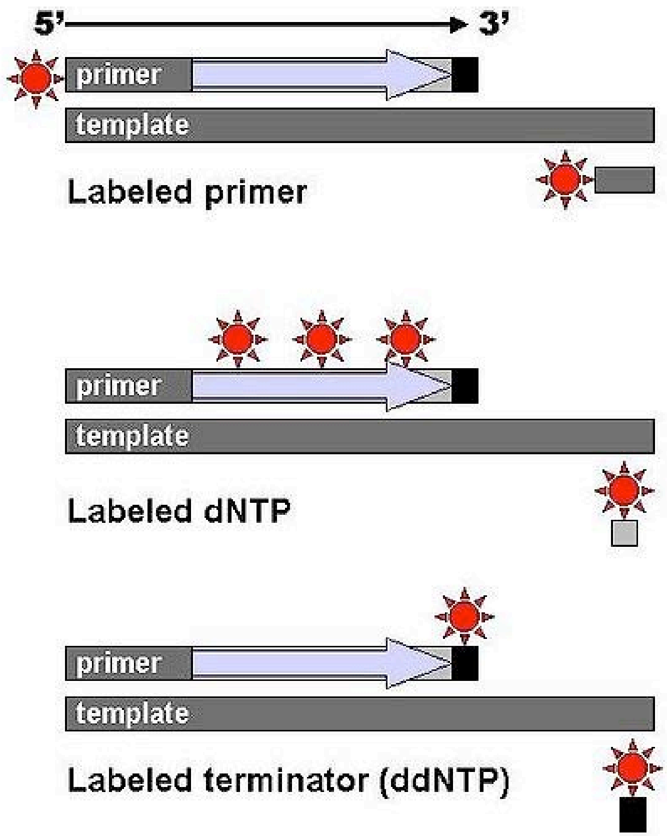
The "top-down" approach to whole genome sequencing.

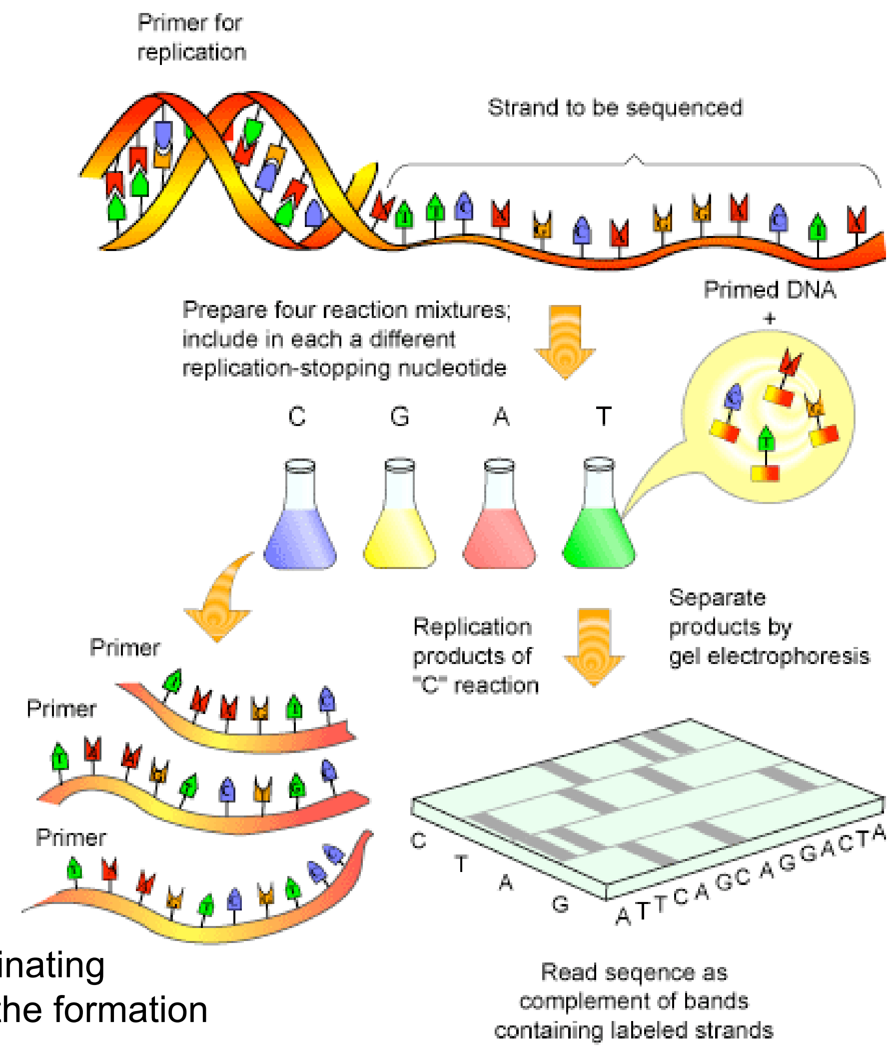
Bottom-Up



Shotgun Sequencing
Vector
Source clone (40 kb insert)
unordered sequencing clones (1-2 kb inserts)
gap
ATTGGCTTTCCGGGAAAACCTTTAA
TAACCGAAAGGCCCTTTTGGAAATT
ordered overlapping sequences

Sanger Sequencing



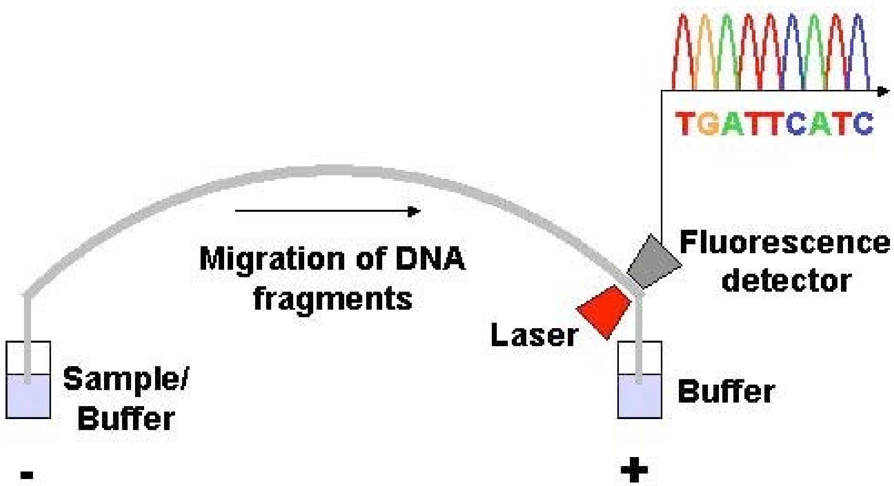


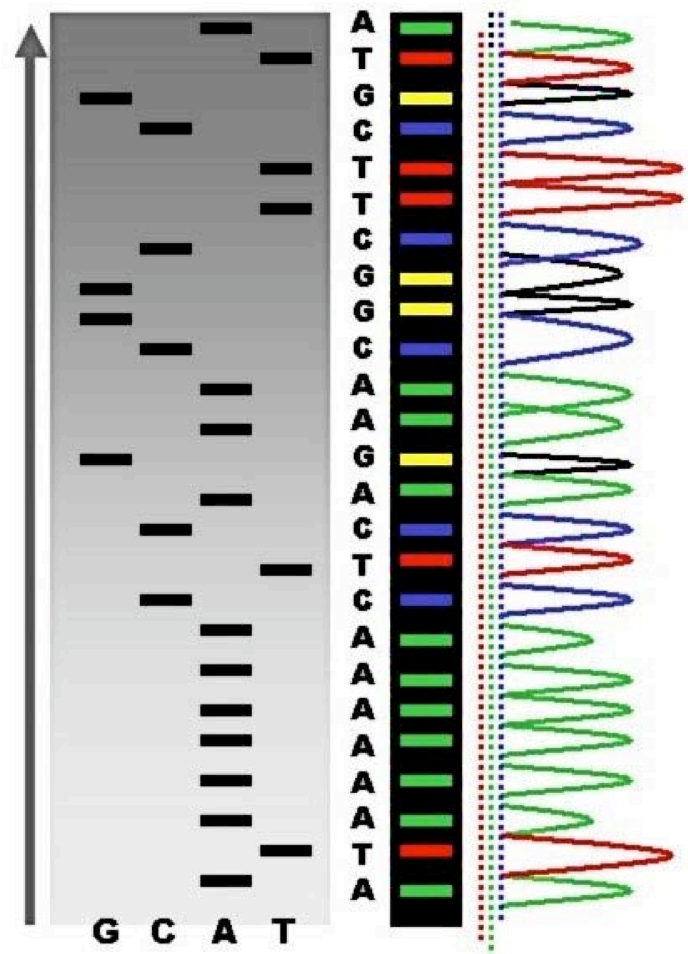
To each reaction is added **only one** of the four dideoxynucleotides which are the chain-terminating nucleotides, **lacking a 3'-OH group** required for the formation of a phosphodiester bond between two nucleotides, thus **terminating DNA strand extension** and resulting in various DNA fragments of varying length.

http://www.scq.ubc.ca/wp-content/uploads/2006/08/sequencing2.gif

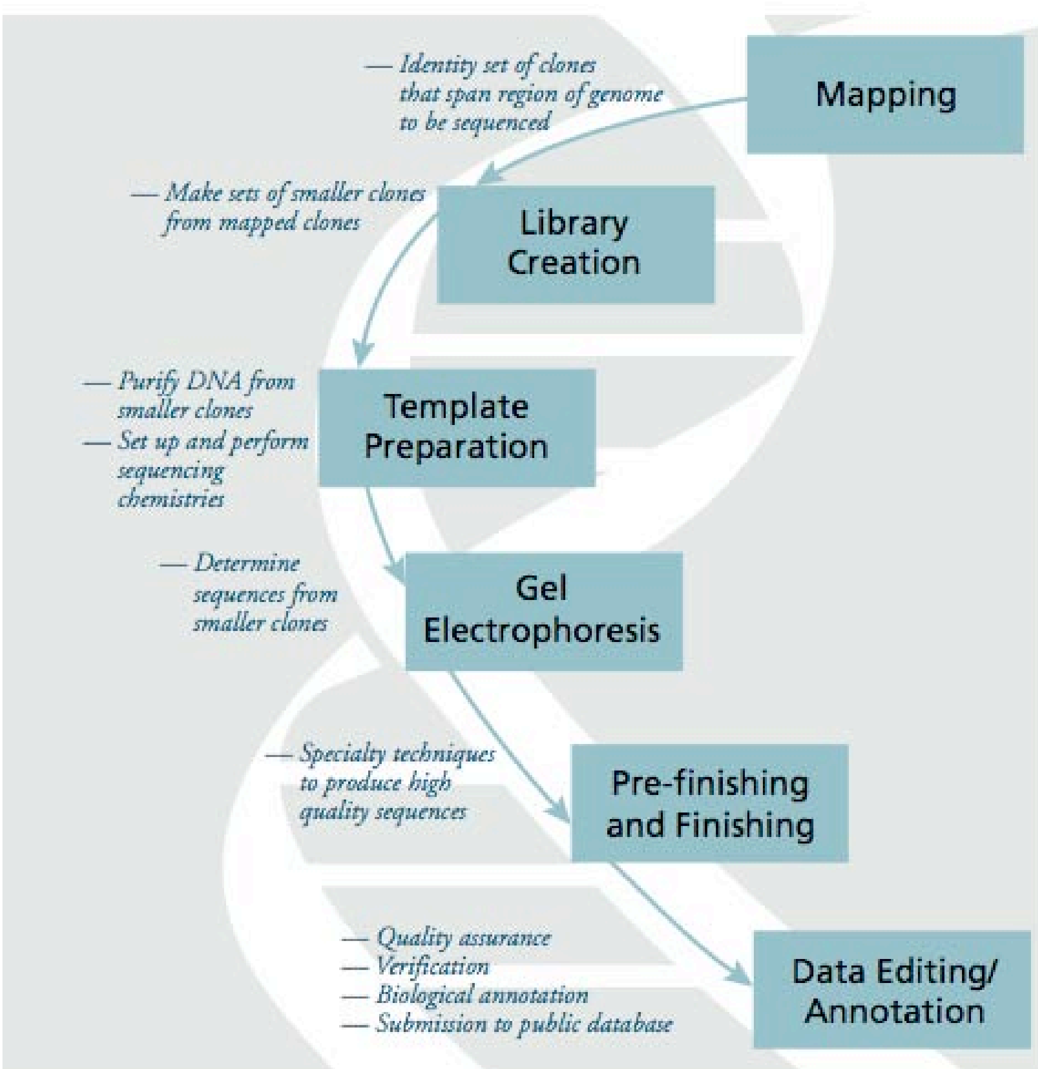
Sanger Sequencing II

DNA fragments visualized on a gel with each of the four reactions run in one of the lanes (G,C,A,T). DNA can either be visualized either via exposing the gel to UV and reading from the X-ray film or if the ddNTPs were fluorescently labeled via a laser

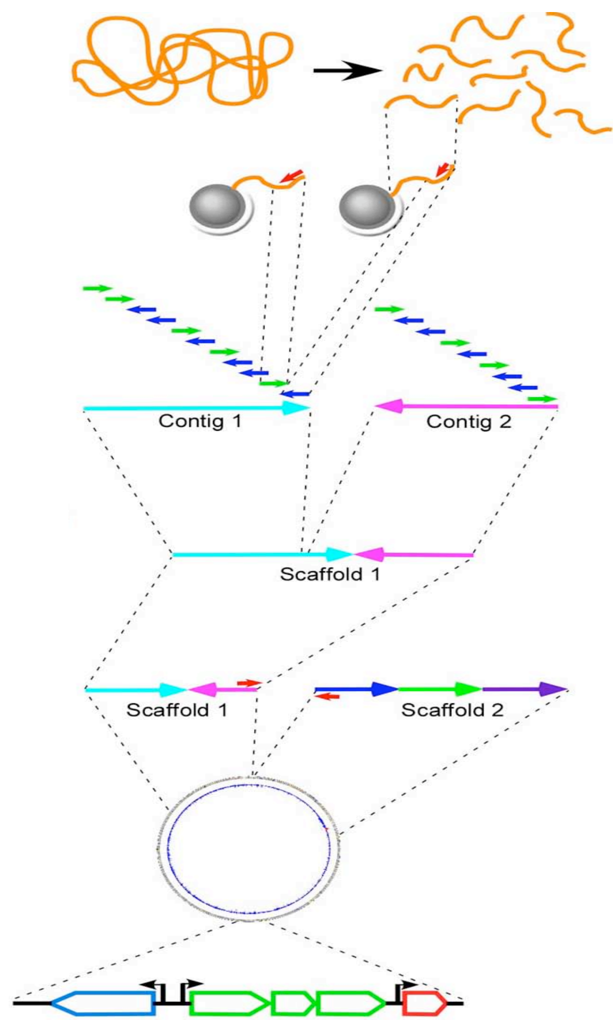




Comparison of Next Gen Seq Project Set up

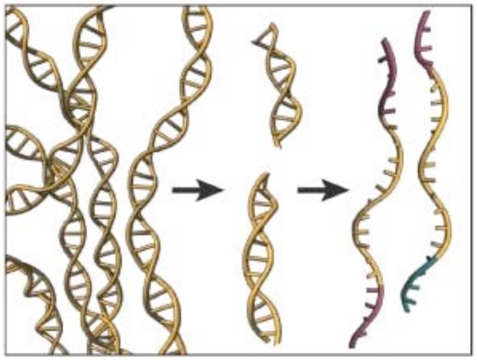


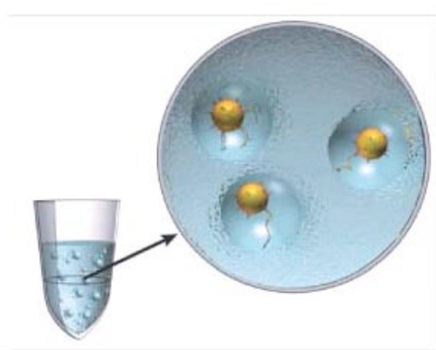
DOE JGI

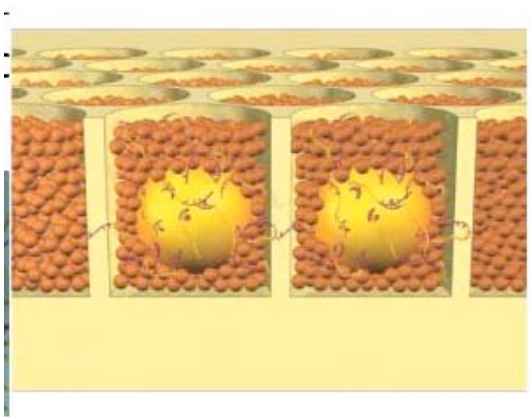


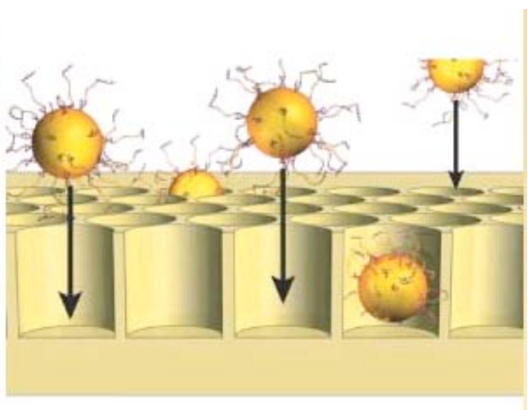
Smith, Gianoulis, et al., (2007) Genes Dev

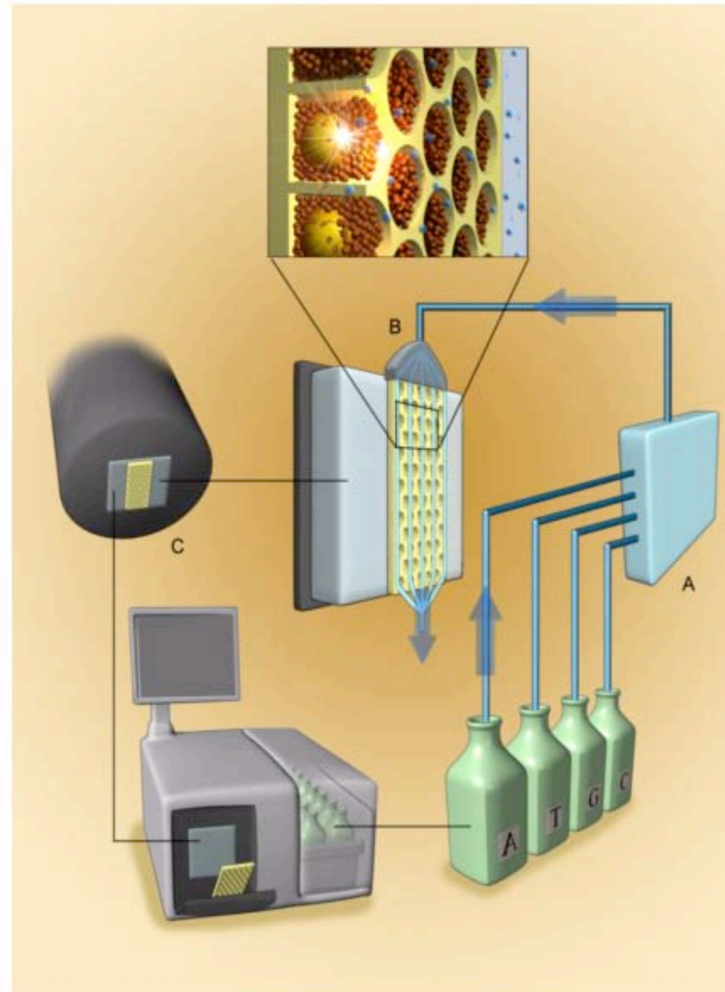
Next-Gen Sequencing





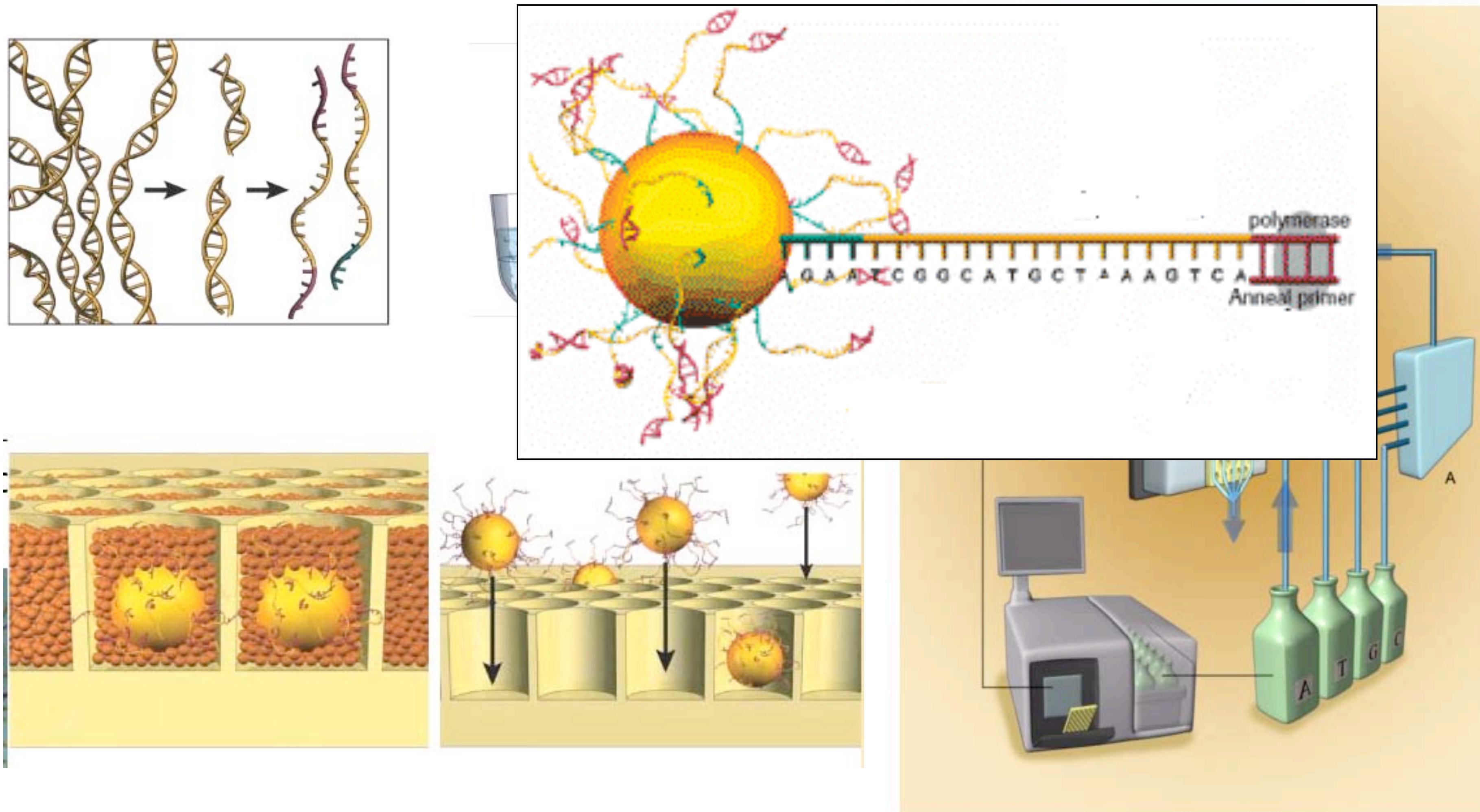






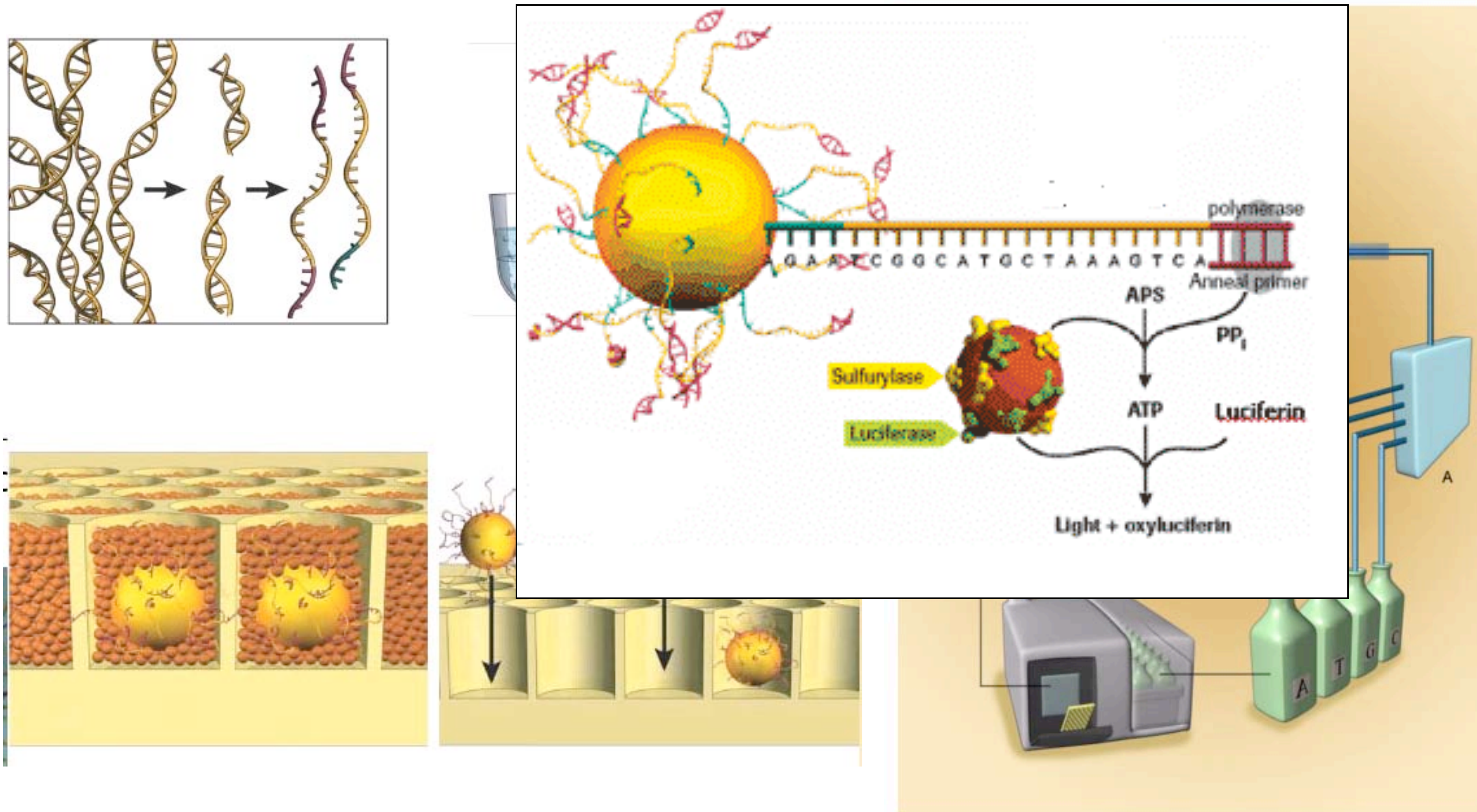
Margulies et al., (2005) Nature

Next-Gen Sequencing



Margulies et al., (2005) Nature

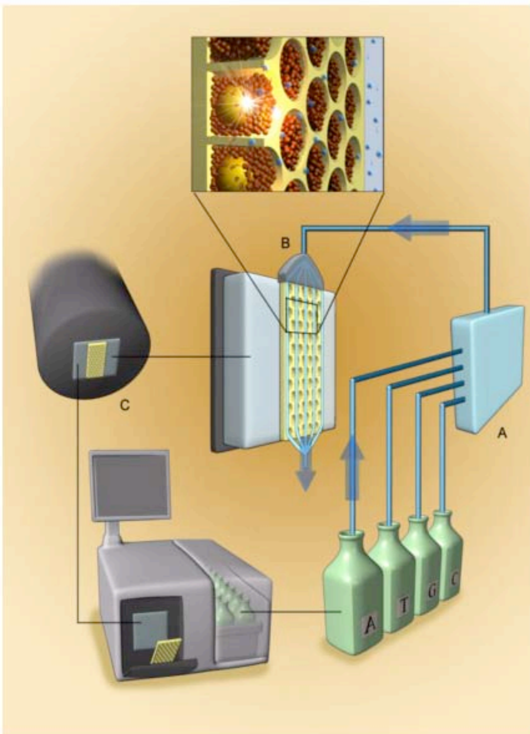
Next-Gen Sequencing



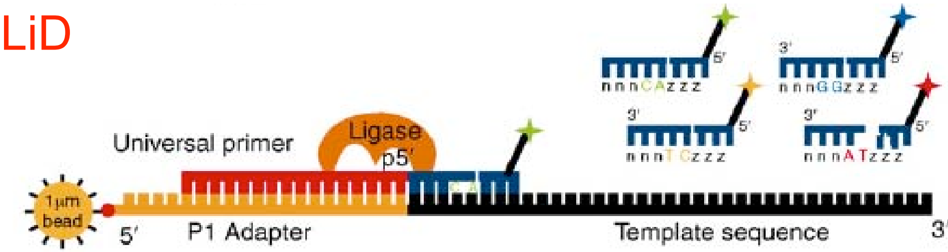
Margulies et al., (2005) Nature

Next-Gen Sequencing

454 Sequencing



SOLiD



Solexa

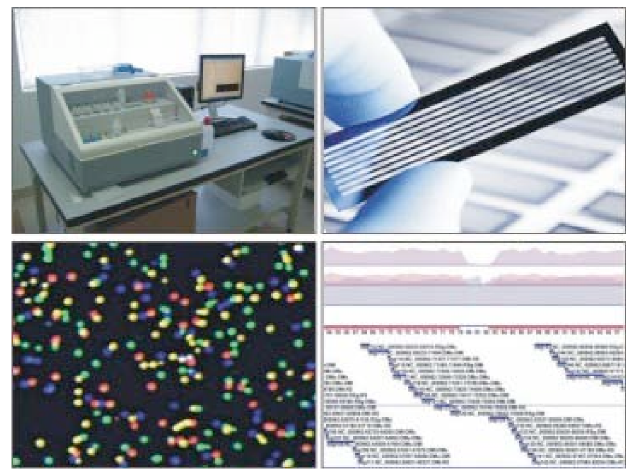


Table 1. Comparing metrics and performance of next-generation DNA sequencers

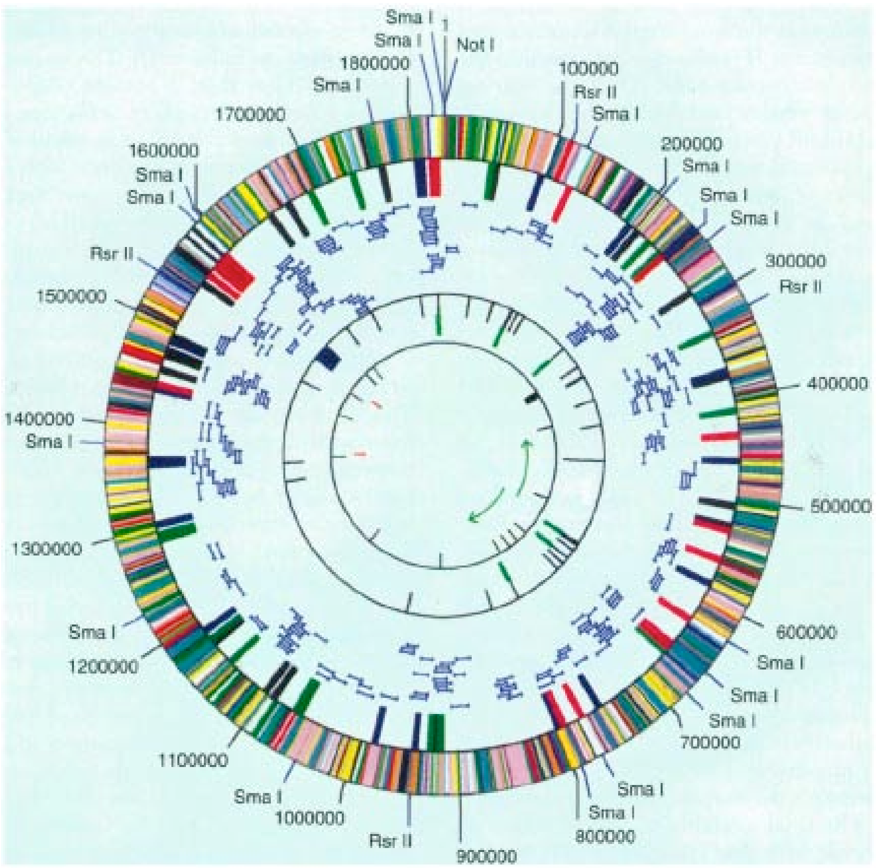
	Platform		
	Roche(454)	Illumina	SOLiD
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/separation	Yes/3 kb	yes/200 bp	Yes/3 kb
Mb/run	100 Mb	1300 Mb	3000 Mb
Time/run (paired ends)	7 h	4 days	5 days
Read length	250 bp	32–40 bp	35 bp
Cost per run (total direct[a])	$8439	$8950	$17 447
Cost per Mb	$84.39	$5.97	$5.81

[a]Total direct costs include the reagents and consumables, the labor, instrument amortization cost and the disc storage space required for data storage/access.

GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACAT
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCAT
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTT
GGTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGT
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAAT
ATGAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTT
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAT
AGGAAAGGTAGAAGAGTTTAAGAAAAATTGCCGGTCTCTAGCTAGCATGCTACGTACTACGTACACGATCAGCATTATATATCGCGCGCTACGATCGATCTACGTA
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACAT
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTC
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTGT
GTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAAT
CCAAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTA
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAT
AGGAAAGGTAGAAGAGTTTAAGAAAAATTGCCGGGTCAGTCGATAGTGATCTATATATATGCGCGTAGCTACTAGCTAGCTAGCTAGCTAGCTAGTCAGTAGCTAC
TAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGCC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAAT
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACAC
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCTA
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGATT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGAA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGCA
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGGT
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTAC
GTGGATAACGTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
CTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCT
TCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGCCGCGGGTGCGTCAGGTTGAAATTTTGGTAGA
CTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCGTGTGTTCCTATACCTCAACCTGCTAAGGTTAAAAGCT
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAAC
AATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATC
AGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGGTGTTGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAA
AAAGGTAGAAGAGTTTAAGAAAAATTGCCGGAAAGGTAGAAGACGCGCGCGCGGCGCCGGCAAGAAAAAATTGCCGGATATTATATATAGCATCGATCGATGCATCC
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACAC
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCTA
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGATT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGAA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGCA
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGGT
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTAC

GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCA
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTT
GGTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAG
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
ATGAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTT
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AGGAAAGGTAGAAGAGTTTAAGAAAAATTGCCGGTCTCTAGCTAGCATGCTACGTACTACGTACACGATCAGCATTATATATCGCGCGCTACGATCGATCTACGT
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCT
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTG
GTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAG
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
CCAAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTT
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AGGAAAGGTAGAAGAGTTTAAGAAAAATTGCCGGGTCAGTCGATAGTGATCTATATATATGCGCGTAGCTACTAGCTAGCTAGCTAGCTAGCTAGTCAGTAGCTA
TAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCT
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGAT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGC
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGG
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTA
GTGGATAACGTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGAC
CTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCC
TCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGCCGCGGGTGCGTCAGGTTGAAATTTTGGTAG
CTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCGTGTGTTCCTATACCTCAACCTGCTAAGGTTAAAAGC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
AATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAAT
AGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGGTGTTGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AAAGGTAGAAGAGTTTAAGAAAAATTGCCGGAAAGGTAGAAGACGCGCGCGCGGCGCCGGCAAGAAAAAATTGCCGGATATTATATATAGCATCGATCGATGCATC
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCT
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGAT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGC
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGG
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTA

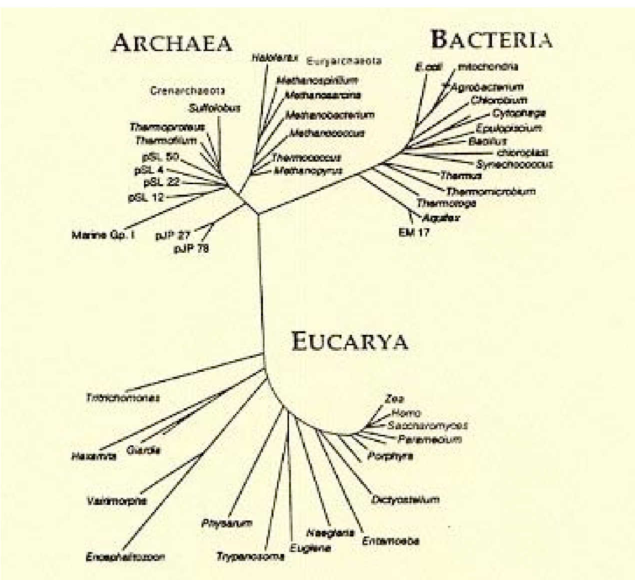
Example 1: First free-living organism sequenced *H. influenzae*



Fleischmann, RD et al Science (1995)

Example 2: Confirmation of the third branch of life, *M. jannischii* ancient and WEIRD

• In the 1970s, Carl Woess hypothesized a third domain of life.
• Sequencing of M jannischii showed Archaeal DNA was no more similar to bacteria than to fish or flowers!

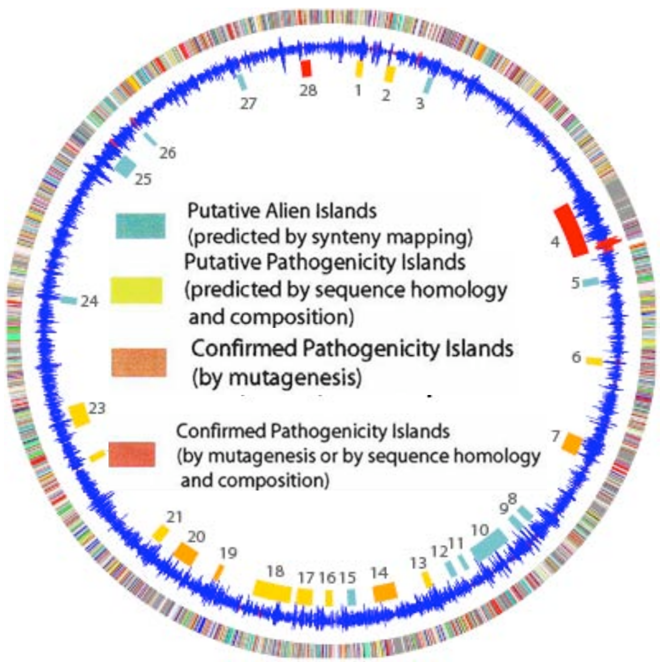


Features of *M. jannaschii*
• Lives 2600 meters below sea level
• Temperature near boiling point
• Pressure high enough to crush a submarine
• Methanogens - Survives on carbon dioxide,
hydrogen, and some mineral
• Their proteins were found to crystallize better

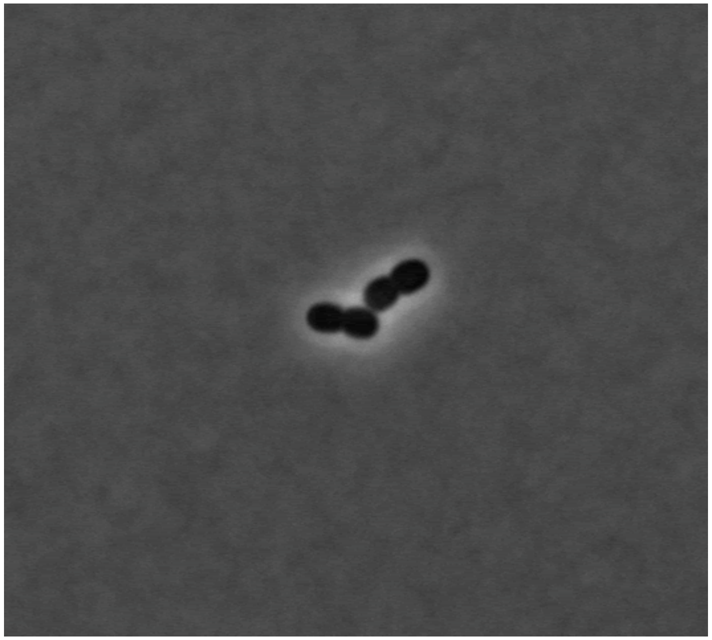
Science 1996

Example 3: Identifying Pathogenicity factors, *Acinetobacter baumannii*

- Multidrug resistant, opportunistic pathogen
- Causative agent of pneumonia, meningitis, septicemia, etc

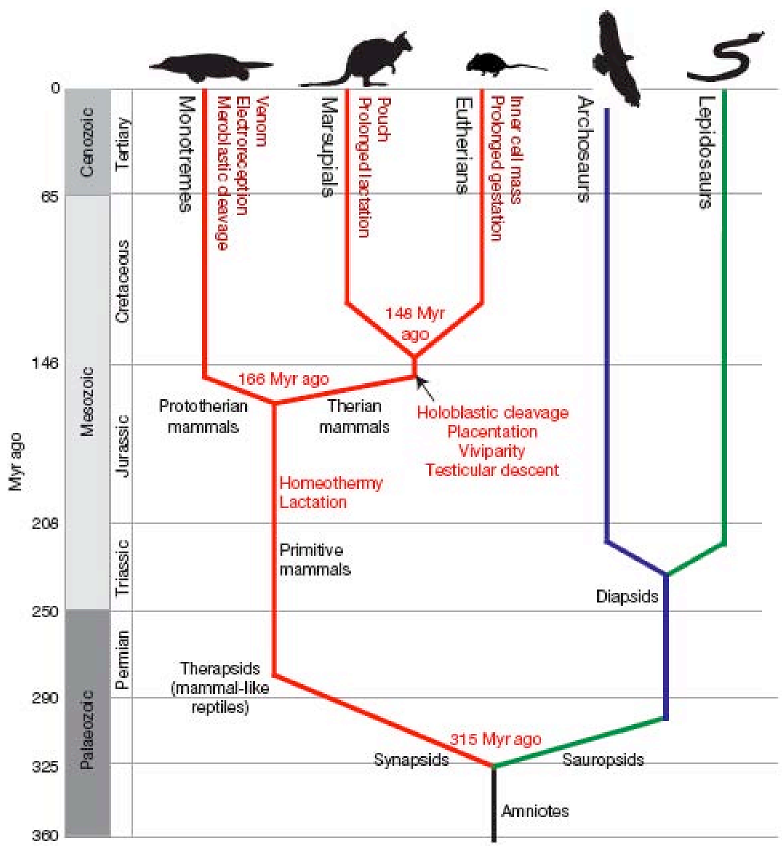


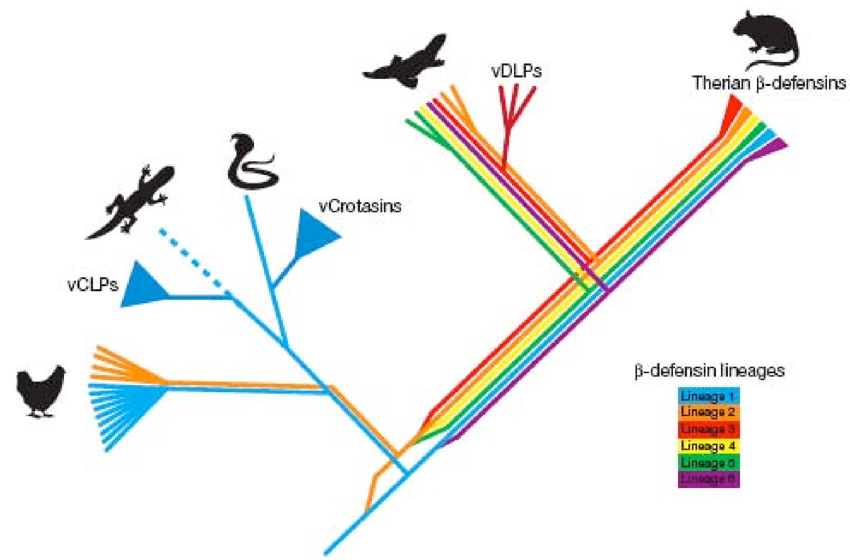
Smith et al, Genes and Dev (2006)



Gianoulis and Schofield (2007) NOVA

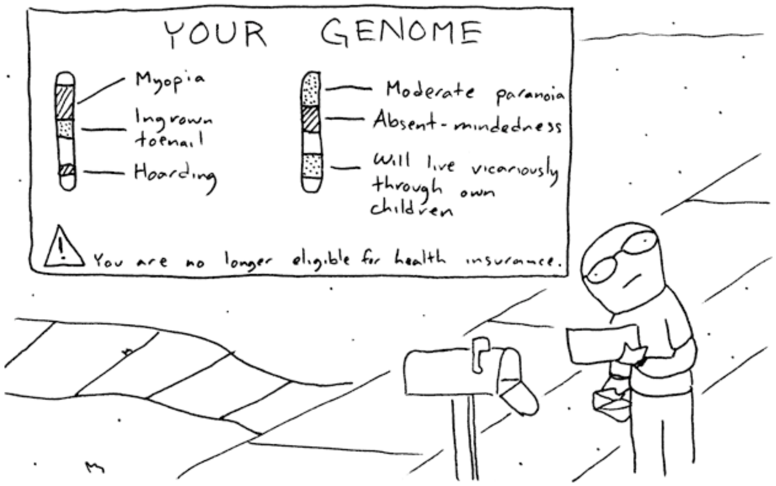
Example 4: Unique evolutionary signatures, The Platypus





Platypus consortium, Nature 2008

Example 5: Human Genome Project



Surprising findings from the Human Genome Project

- Just about 1.5% of the genome was exons
- There appear to be only about 20,000 human genes

NOTE: To put this in perspective, rice has 37,000 predicted genes. What does this mean? What are the other ~99% doing? (will be covered in detail by M. Snyder)

Metagenomics: Environments

What is metagenomics? What types of questions can we answer with it? What are the computational challenges?

Why do we care?

- Microbes are found everywhere: in the soil, the air, the sea, in deep thermal vents, our skin and gut.
- They are incredibly adaptable. Able to beg, steal, or borrow DNA components from other microbes making their evolutionary lineage difficult to determine.
- Microbes have been identified that can thrive in extreme heat, pH. They can survive without water, without sunlight, some can even withstand nuclear radiation.
- What can we learn about these adaptation strategies through studying marine microbes in a continuum of habitats?

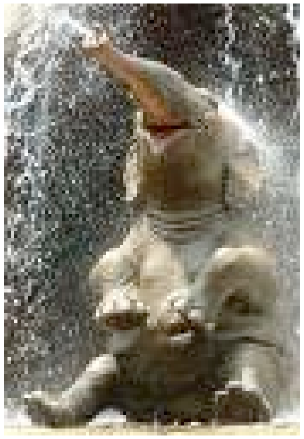
Prokaryotes: The “unseen” majority

Prokaryotes catalyze unique and indispensable transformations in the biogeochemical cycles of the biosphere, produce important components of the earth’s atmosphere, and represent a large portion of life’s genetic diversity. (Whitman et al, PNAS 1998).

- Estimated number of prokaryotes = 4-6x10^30 cells
- Estimated amount of cellular carbon = 350-550x10^15 g
- Estimated amount of nitrogen = 85-130x10^15 g
- Estimated amount of phosphorous = 9 -14 x 10^15 g
- 60% of the ocean’s biomass is prokaryotic

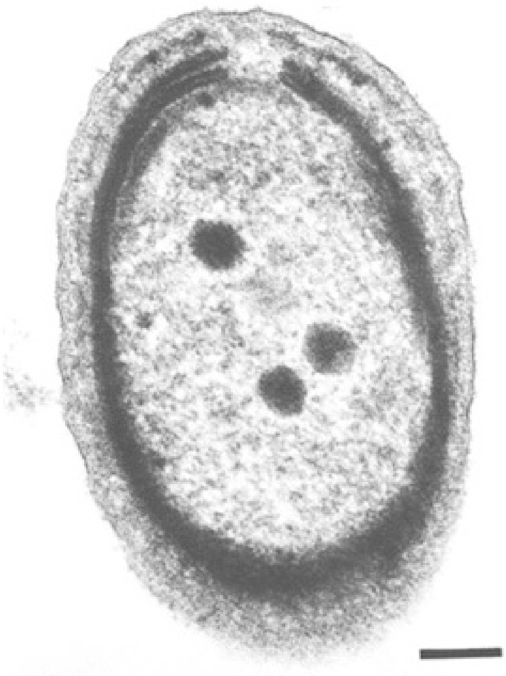
Note: 1x10^15 is 1 trillion grams or 1102311 tons or 44,000 elephants!

Additional Note: Earth is estimated to weight 6x10^24 kg



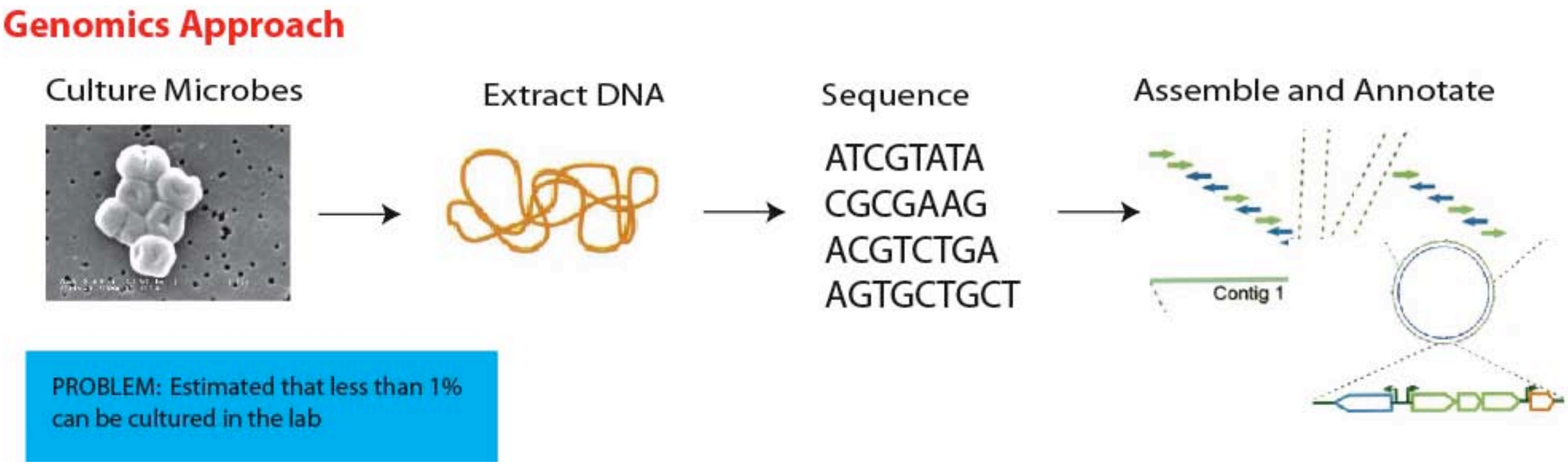
What can we learn?

- The tiny, just 0.6 micrometer, Prochlorococcus has just under 1500 genes. It also happens to account for 30-80% of the primary productivity of the world's oceans.
- A single milliliter of water can contain 100,000 cells or more.
- Despite, its significance in world-wide carbon cycling and climate regulation, it wasn't discovered until 1986 by Sallie Chisholm.



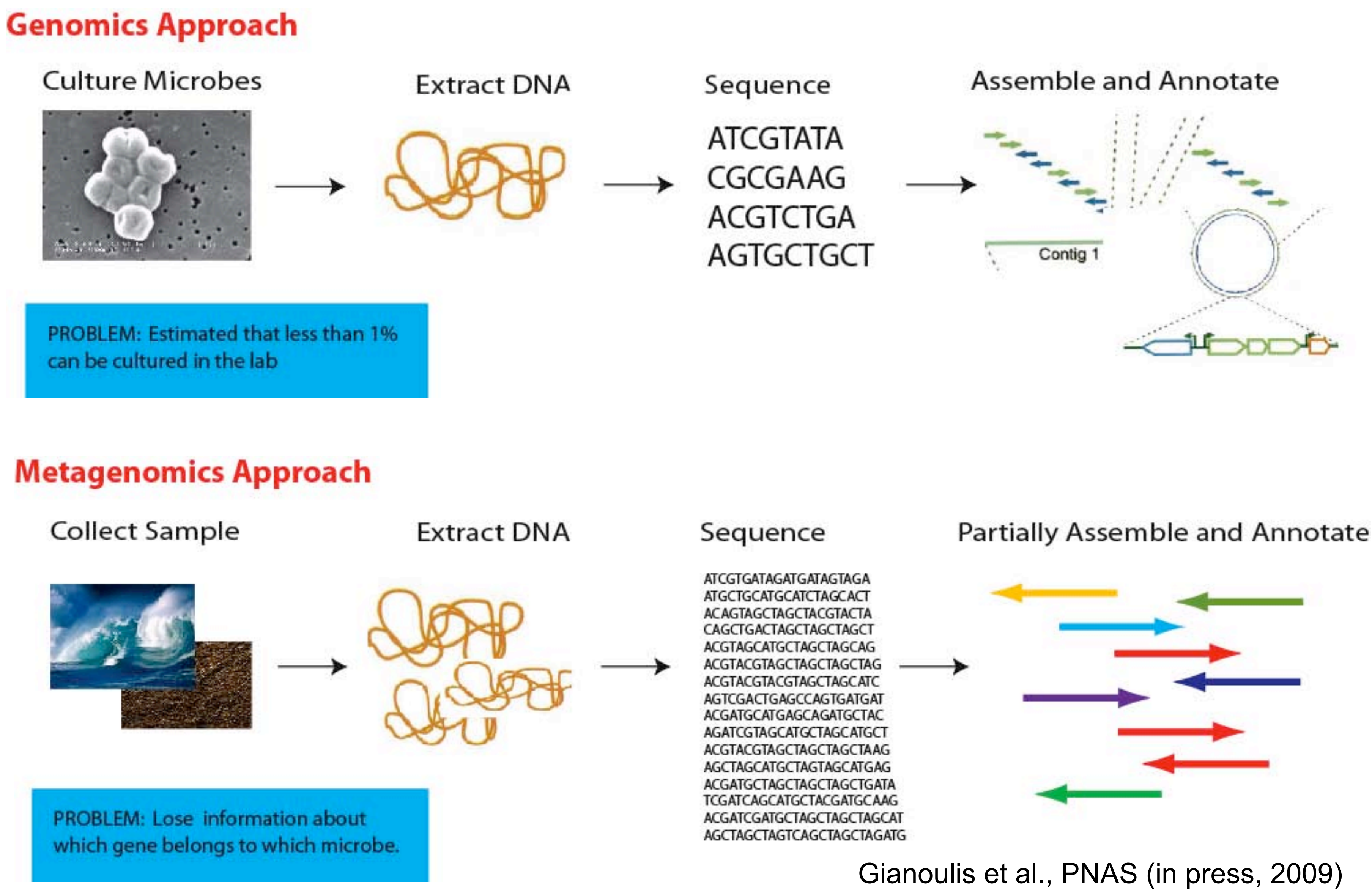
http://mitworld.mit.edu/video/421/

What is metagenomics?



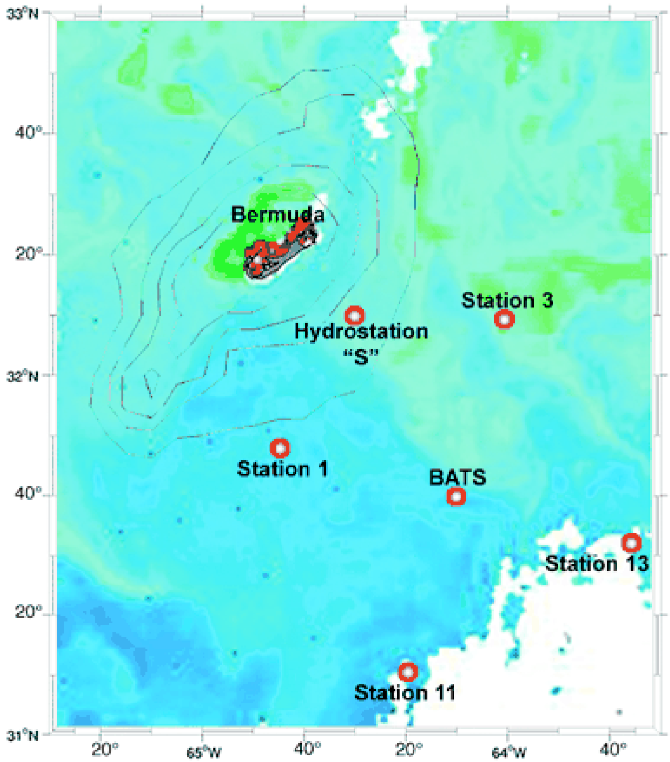
Gianoulis et al., PNAS (in press, 2009)

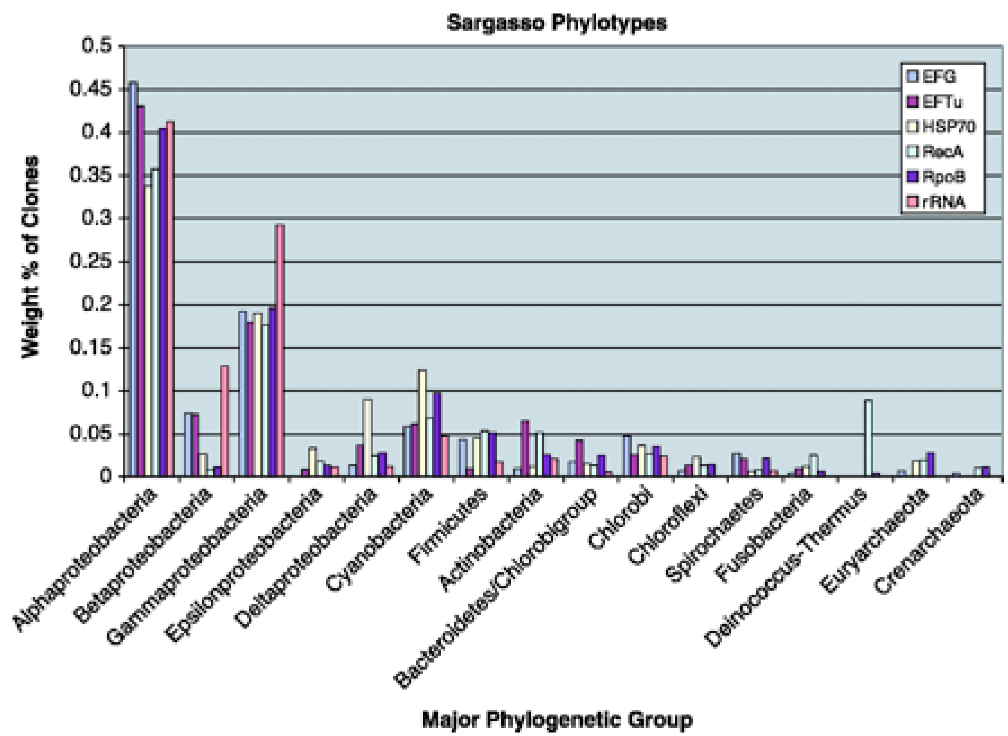
What is metagenomics?



Gianoulis et al., PNAS (in press, 2009)

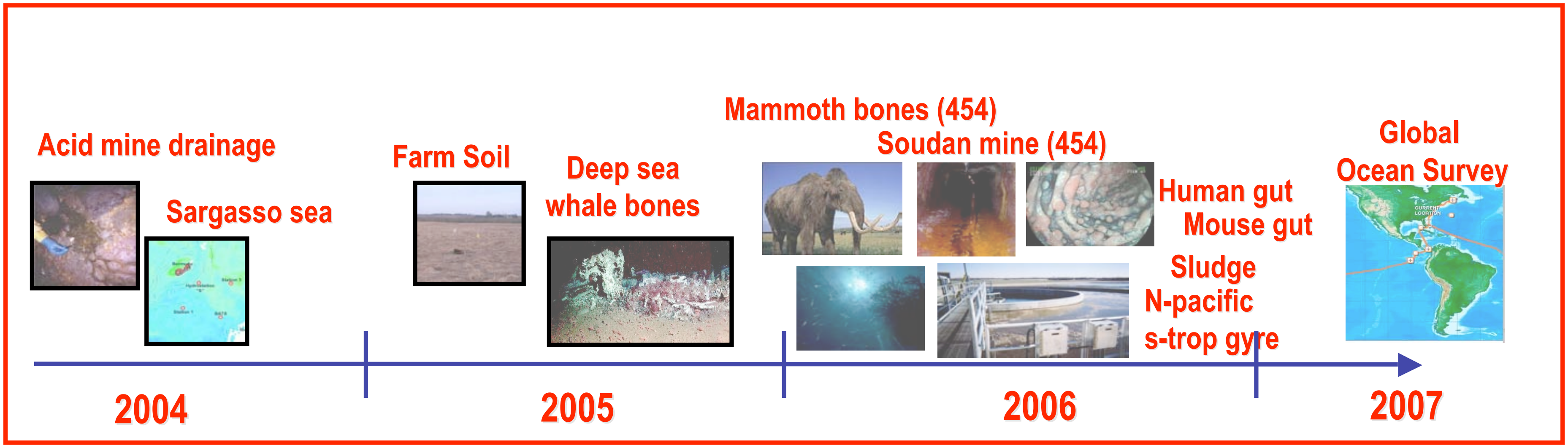
Seeing the "unseen": Sargasso Sea is NOT a desert





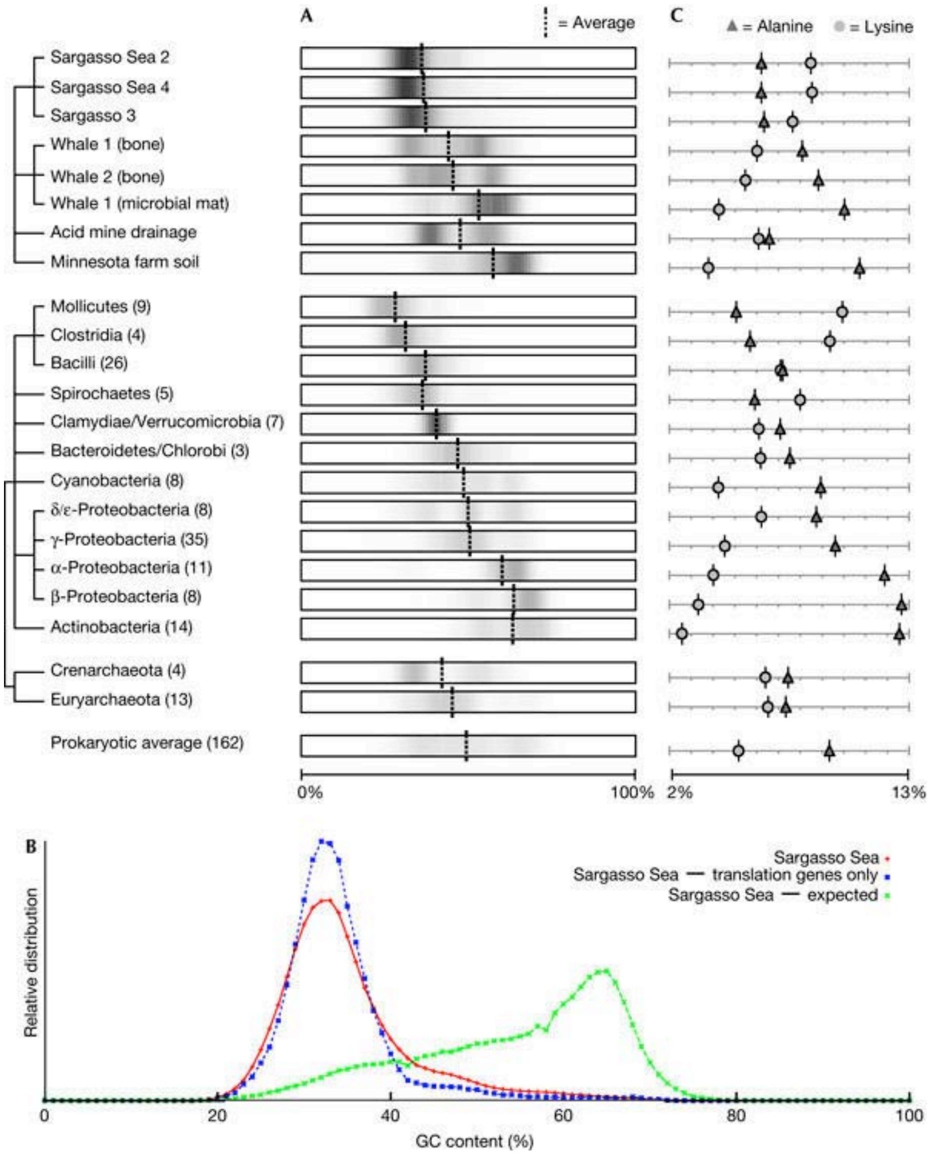
Venter et al. Science (2004)

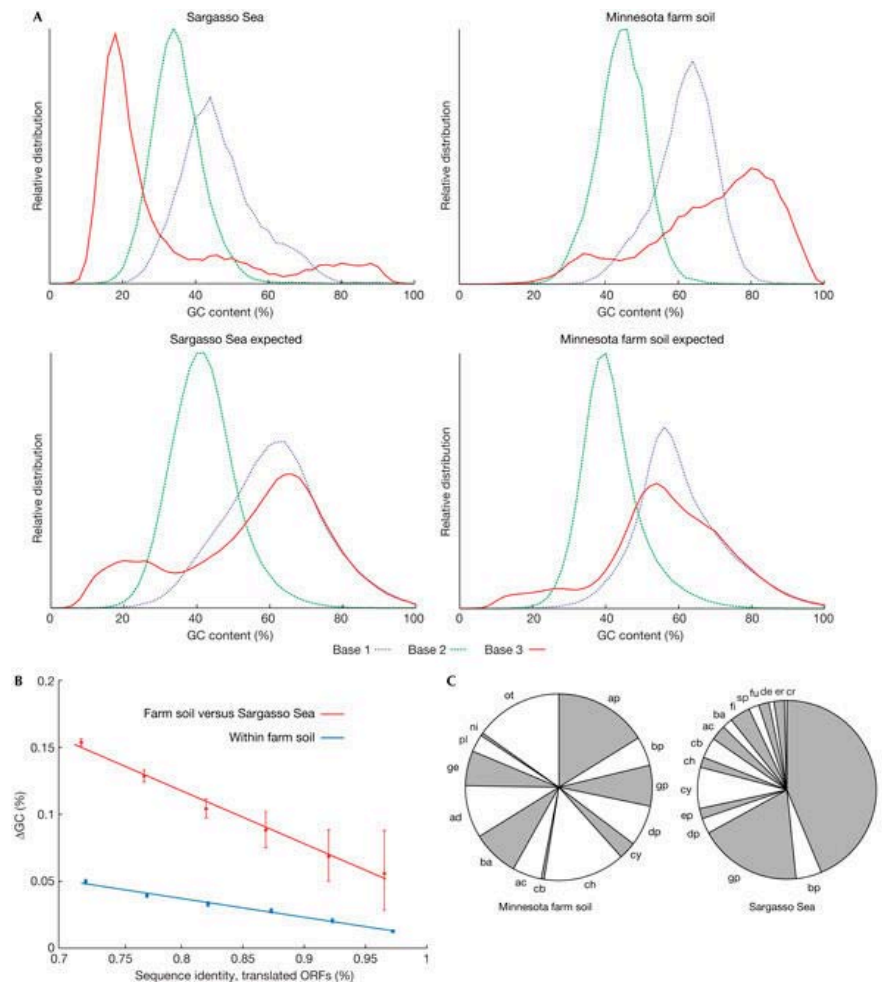
Comparative Metagenomics



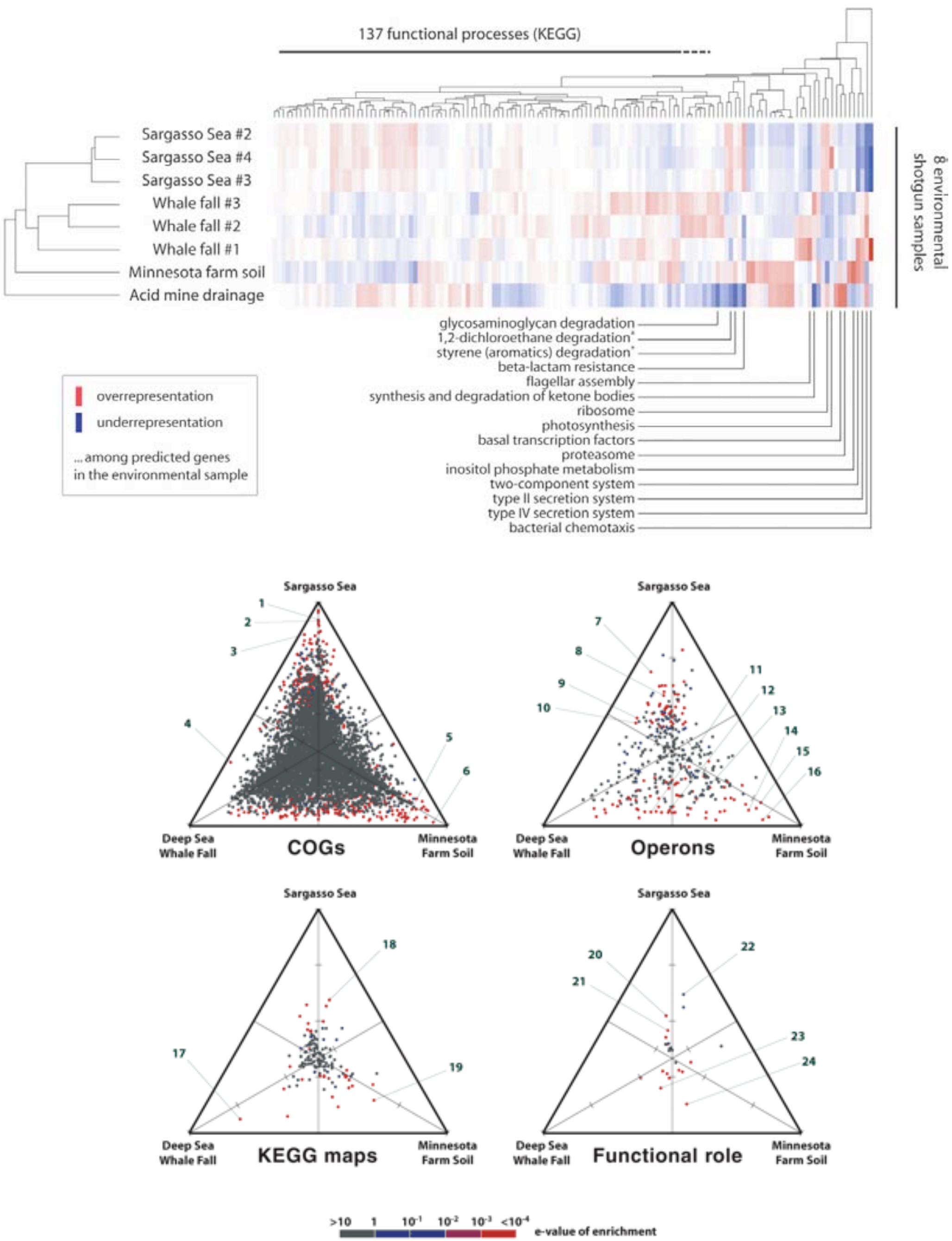
Adapted from P. Bork

Do environments shape nucleotide composition?





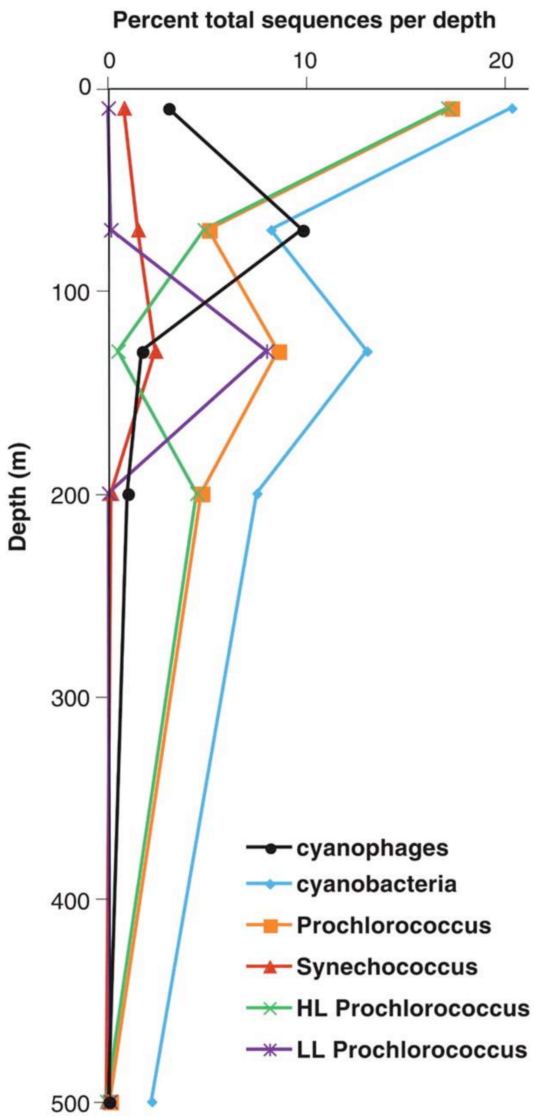
Foerstner KU EMBO 2006

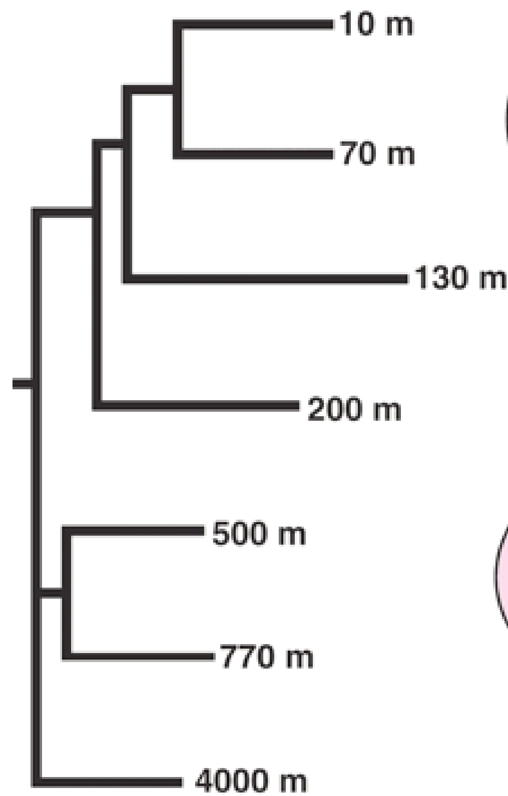


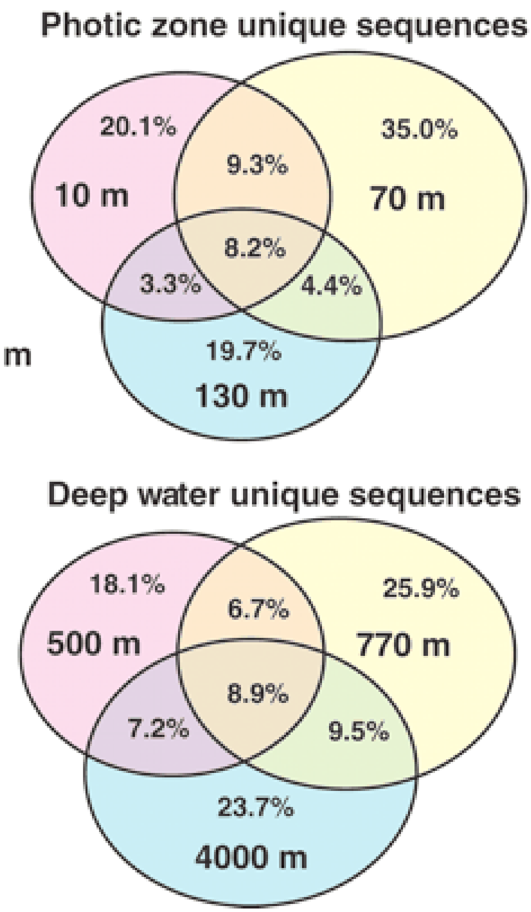
The predicted metaproteome, based on fragmented sequence data, is sufficient to identify functional fingerprints that can provide insight into the environments from which microbial communities originate. . . . Just as the incomplete but information-dense data represented by expressed sequence tags have provided useful insights into various organisms and cell types, EGT-based eco-genomic surveys represent a practical and uniquely informative means for understanding microbial communities and their environments.

Tringe, et al
Science 2005

HOT ALOHA





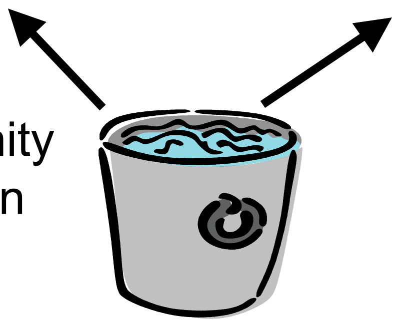


Delong et al, Science 2006

Global Ocean Survey (GOS)

Sequences

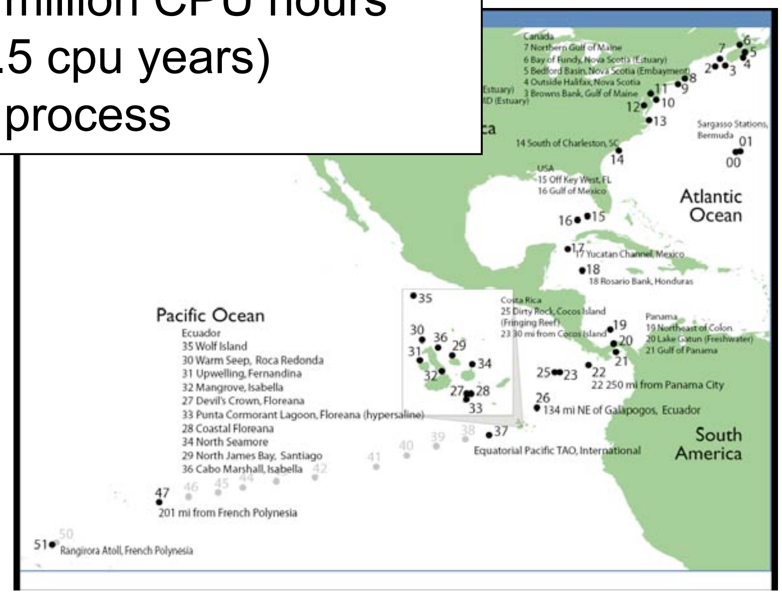
Environmental
Features



Community
Function

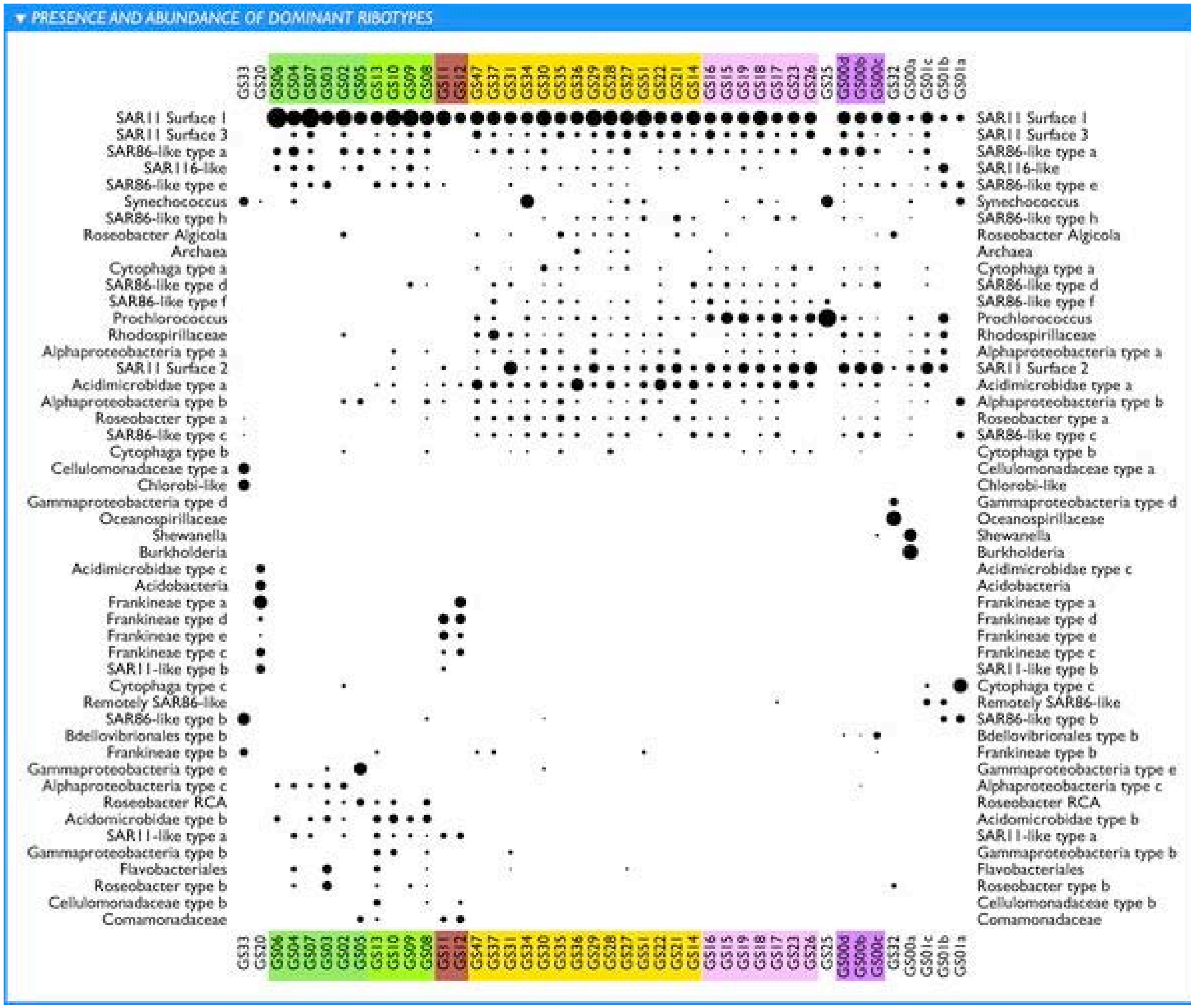
- Temperature
- Salinity
- Chlorophyll
- SampleDepth
- WaterDepth

6.25 GB of data
7.7M Reads
1 million CPU hours
(0.5 cpu years)
to process



Rusch et. al., (2007) PLOS Biology

Knowing who is in the community doesn't always tell you what they are doing.

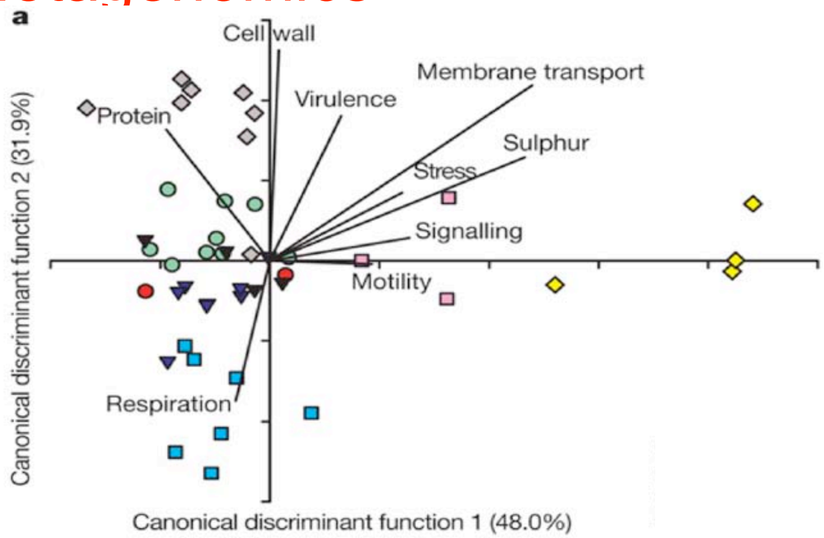


Rusch DB, PLOS Bio 2007

Comparative Metagenomics

Water

Soil



Dinsdale et. al., Nature 2008

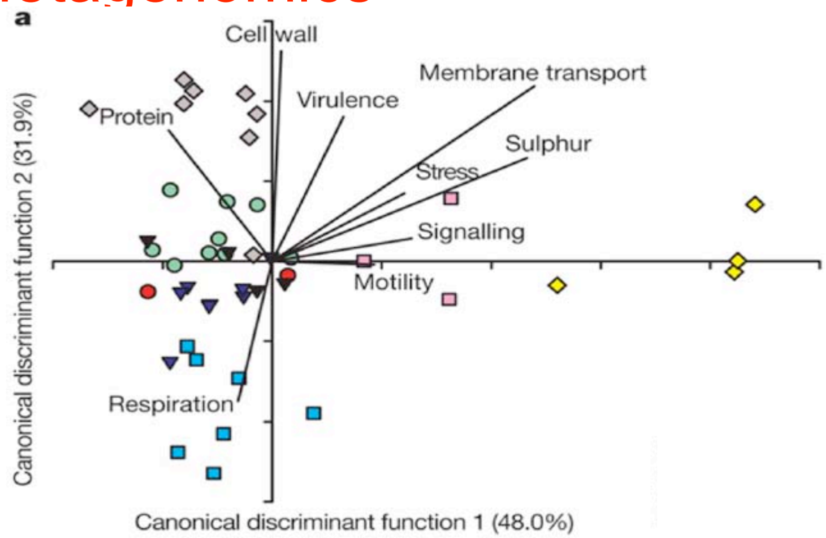
Comparison of different environmental samples revealed changes in:

- AA composition and GC content (Foerstner et al., EMBO, Tringe et al., Science)
- Metabolic capabilities (Tyson et. al.,Nature; Rusch et. al. PLOS Biology)
- Genome size ("streamlining") (Tringe et al., Science)

Comparative Metagenomics

Water

Soil



Dinsdale et. al., Nature 2008

Do the proportions of pathways represented in these two samples differ?

Comparison of different environmental samples revealed changes in:

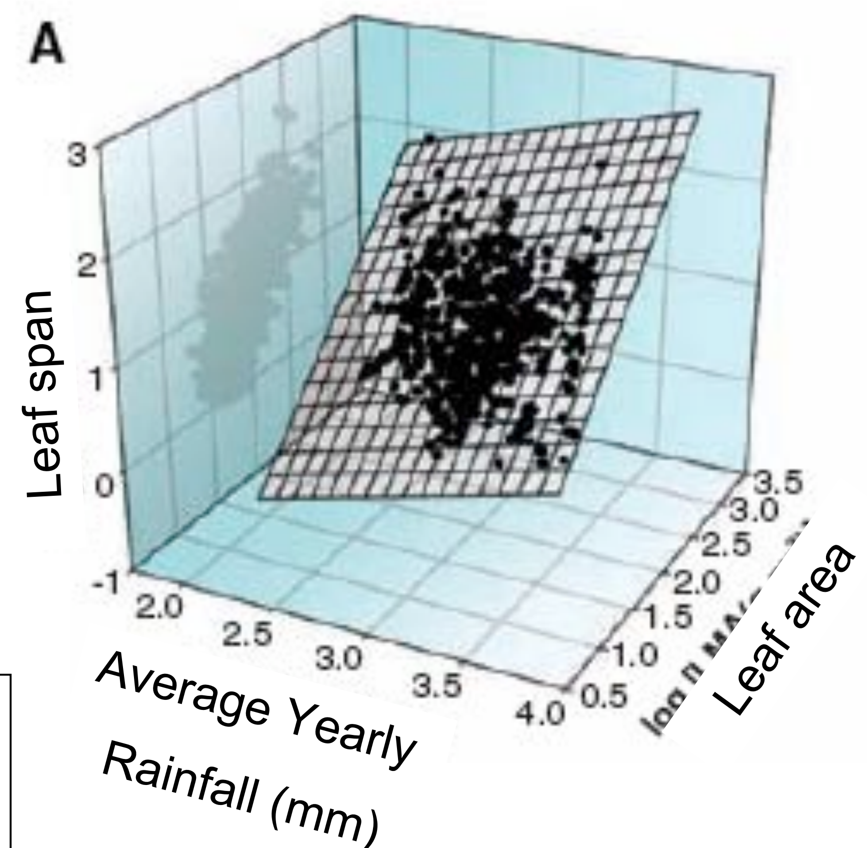
- AA composition and GC content (Foerstner et al., EMBO, Tringe et al., Science)
- Metabolic capabilities (Tyson et. al.,Nature; Rusch et. al. PLOS Biology)
- Genome size ("streamlining") (Tringe et al., Science)

Trait-based Biogeography: Quantitative Description of the Environment

Charles River, MA

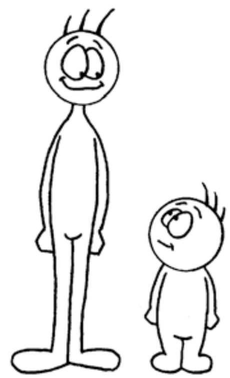
Long Island Sound, CT

Do the proportions of pathways represented in these two samples CHANGE as a function of their environments?

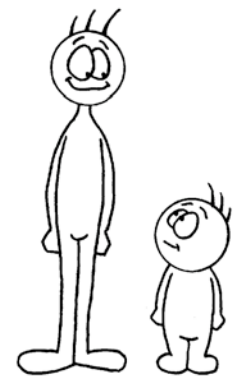


Green et. al., Science 2008

Qualitative to Quantitative



Tall Short



6ft5in 4ft9in

Water

Terrestrial

Global Ocean Survey (GOS)

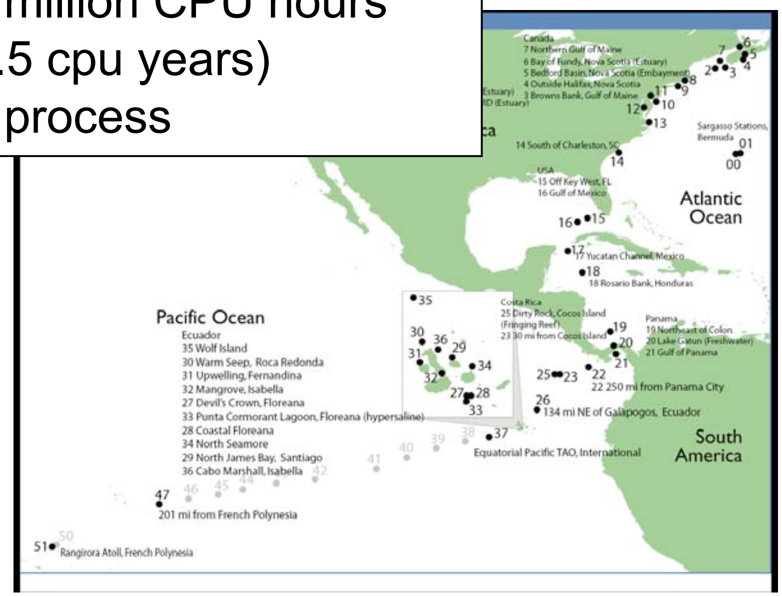
Sequences

Environmental Features

Community Function

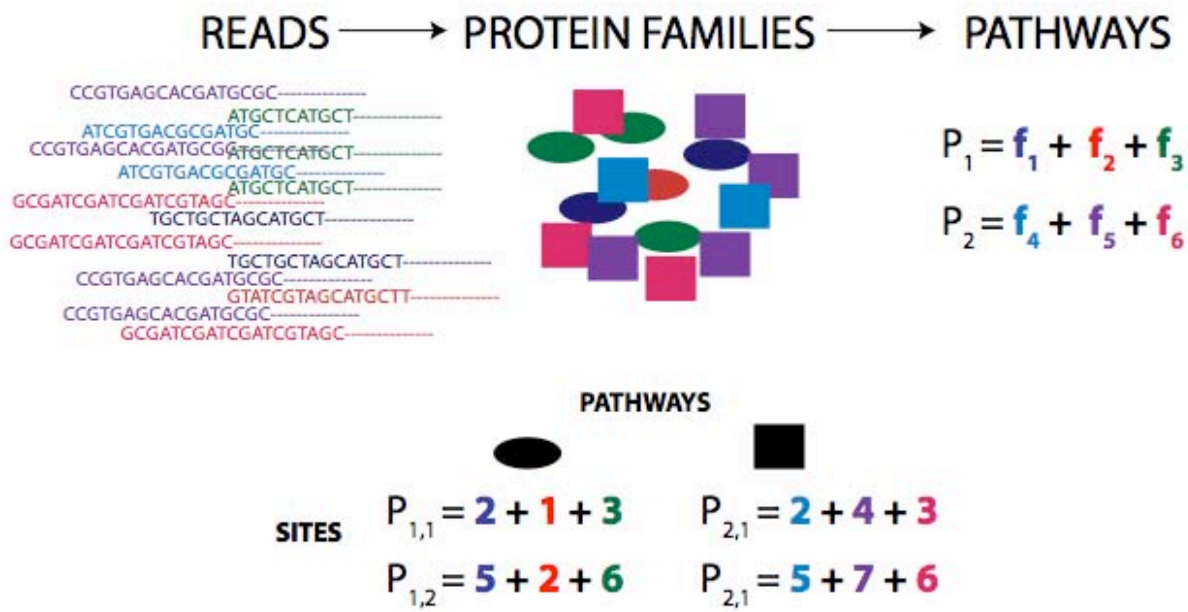
- Temperature
- Salinity
- Chlorophyll
- SampleDepth
- WaterDepth

6.25 GB of data
7.7M Reads
1 million CPU hours
(0.5 cpu years)
to process



Rusch et. al., (2007) PLOS Biology

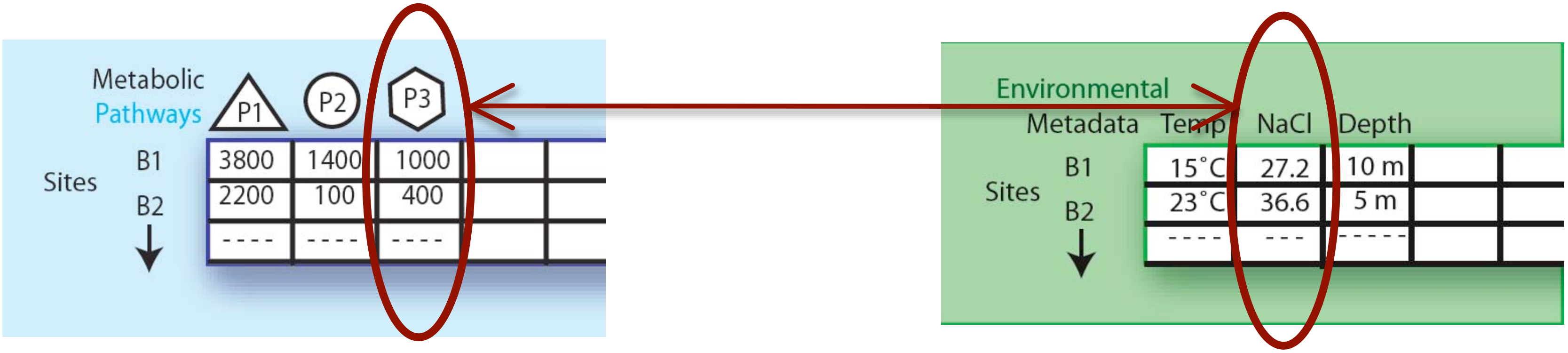
Pathwav Scores

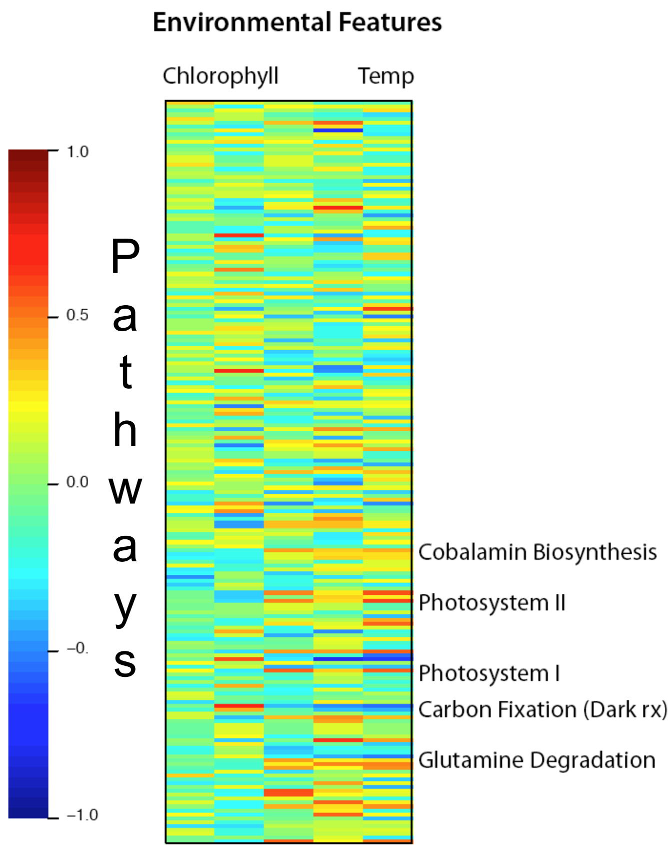


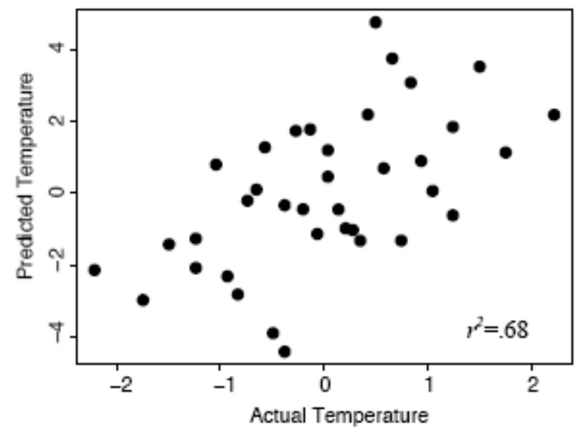
Divide each pathway sum by the "mappable portion" of the site.

(Sum of all f for the site).

Simple Relationships: Pairwise Correlations

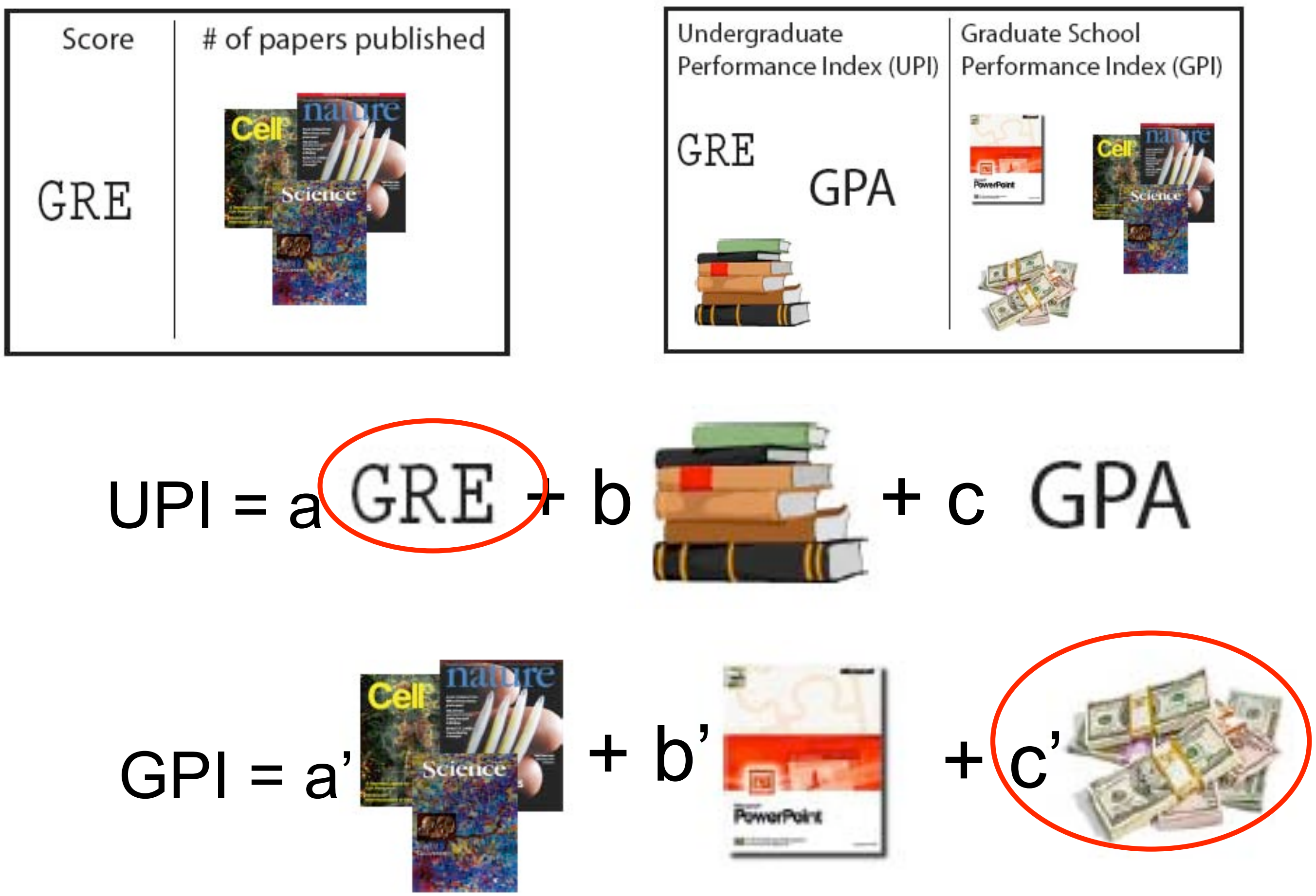




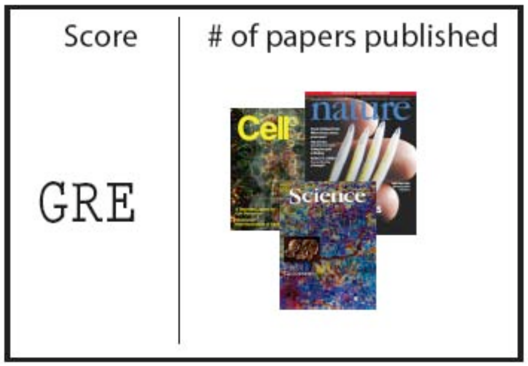


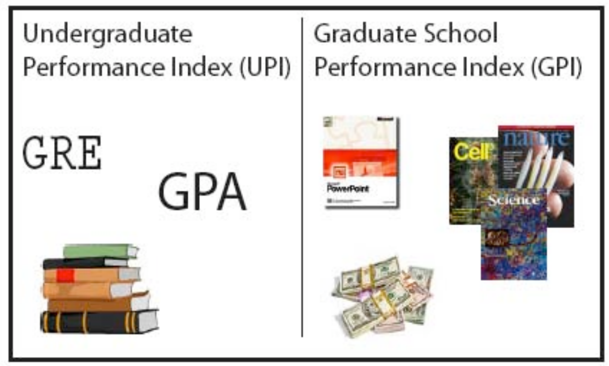
Gianoulis et al., PNAS (2009)

Canonical Correlation Analysis: Simultaneous weighting



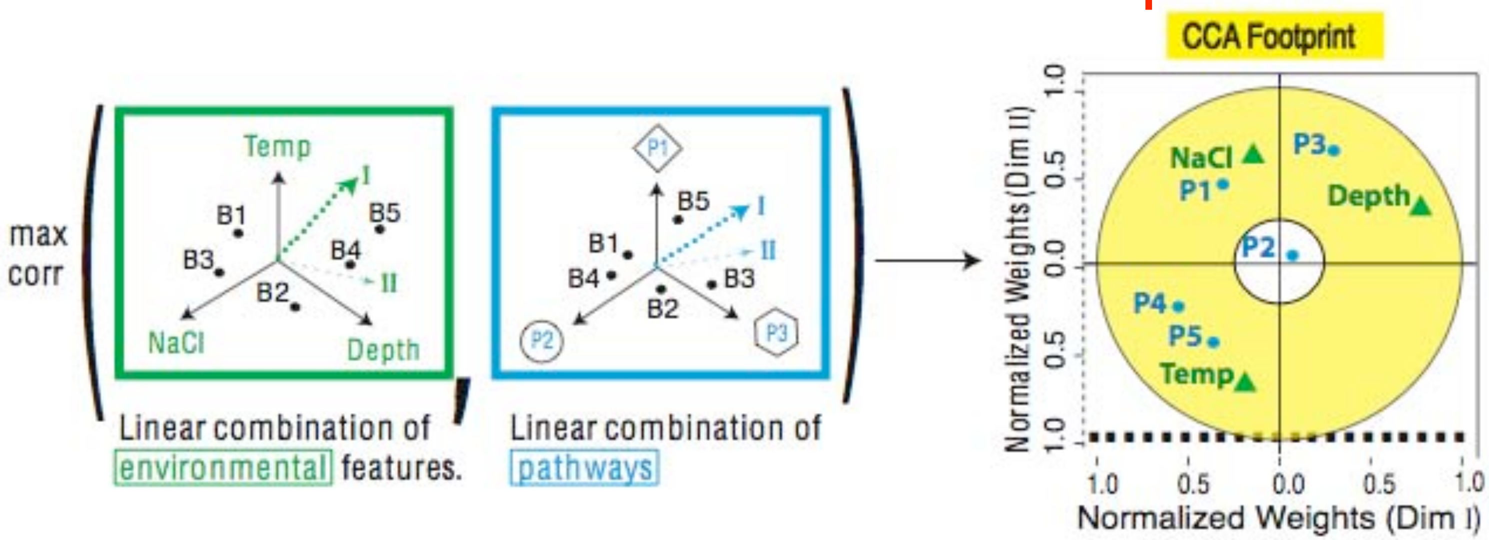
Canonical Correlation Analysis: Simultaneous weighting





Environmental Features	**Metabolic Pathways**
Temp	Photosynthesis
etc	etc
Chlorophyll	Lipid Metabolism

Environmental-Metabolic Space

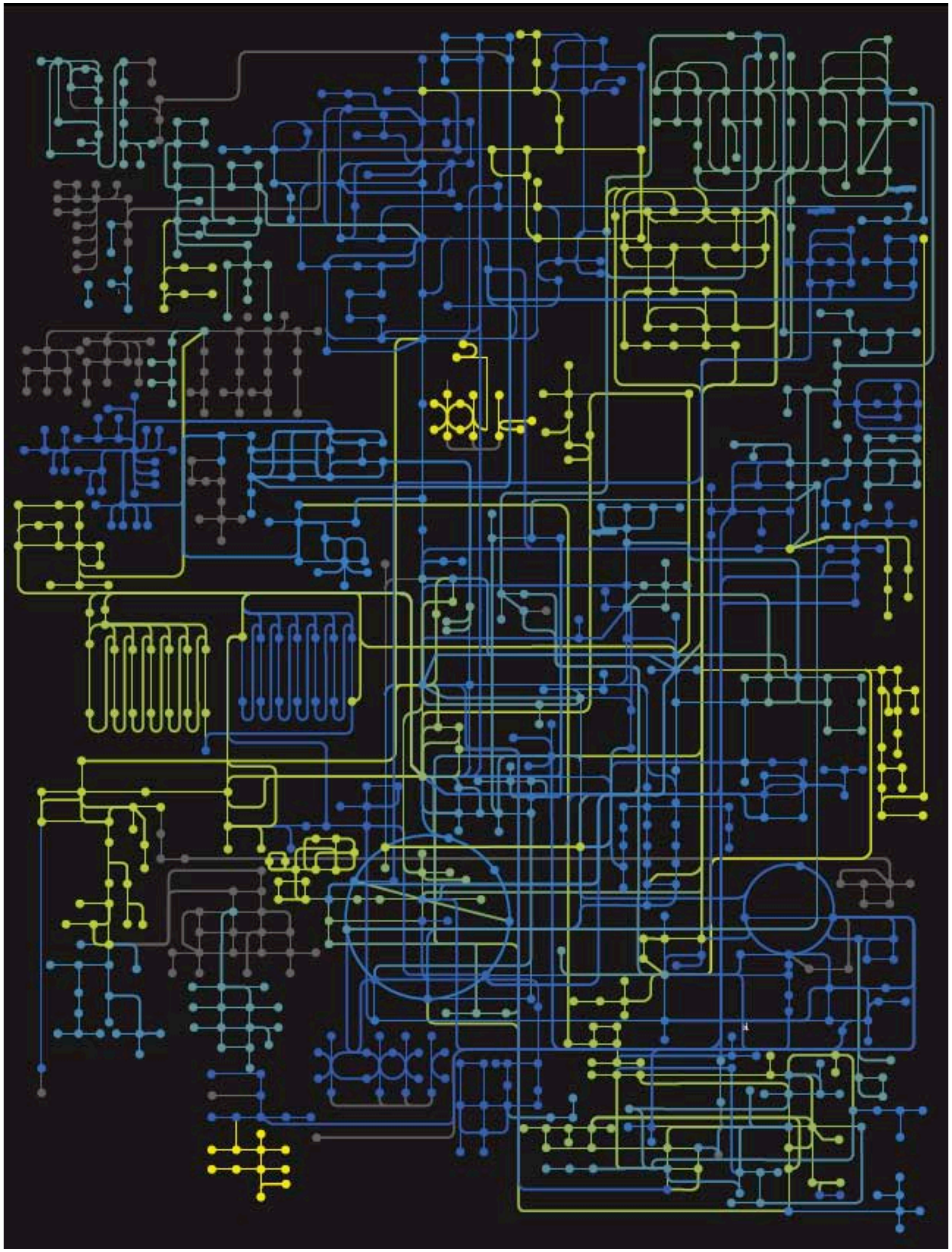


The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

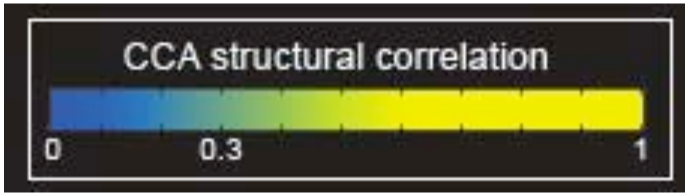
$$\text{Given } X = \{x_1, x_2, \ldots, x_n\} \text{ and } Y = \{y_1, y_2, \ldots, y_m\}$$

$$C = \begin{matrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_Y & \Sigma_{Y,X} \end{matrix} \qquad \max_{a,b} Corr(U,V) = \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a} \sqrt{b' \Sigma_{22} b}}$$

Gianoulis et al., PNAS (2009)

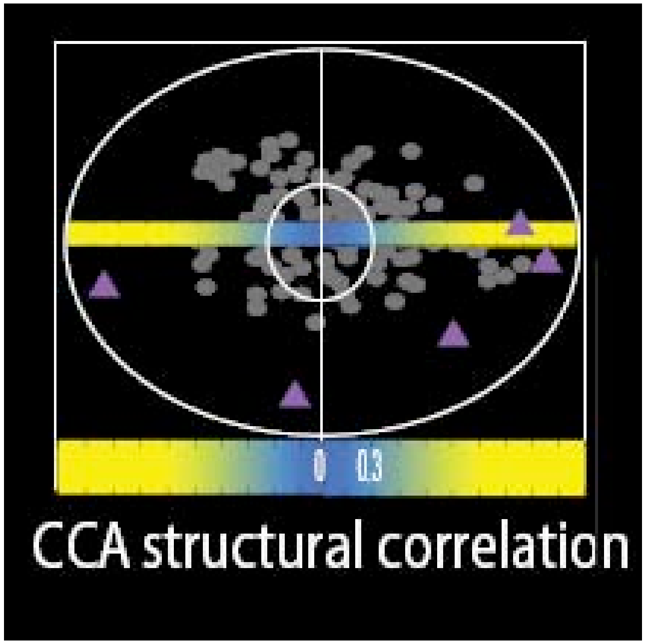


Strength of Pathway co-variation with environment



Environmentally invariant

Environmentally variant



Gianoulis et al., PNAS (2009)

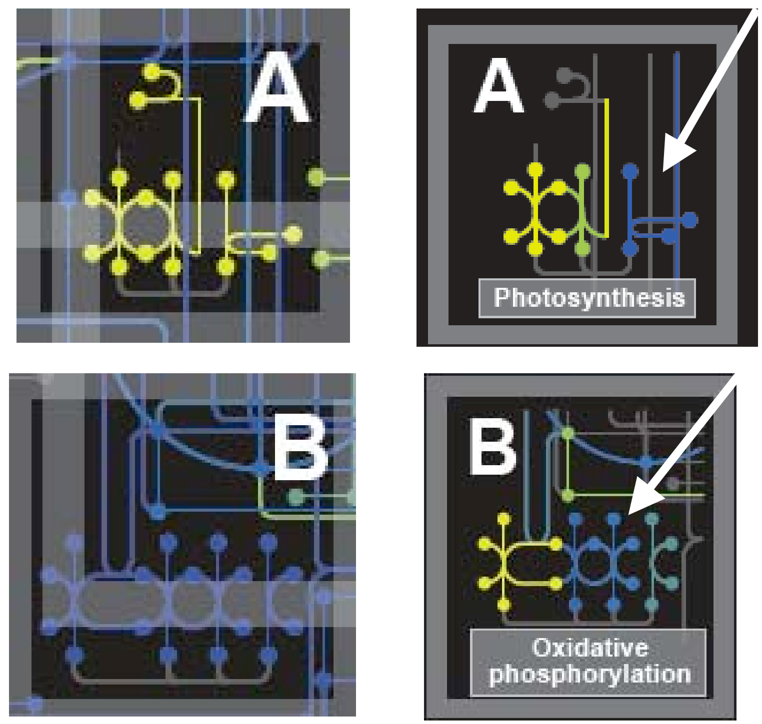
Conclusion #1: Differences in Energy conversion

KEGG

Modules

Photosynthesis

Oxidative Phosphorylation

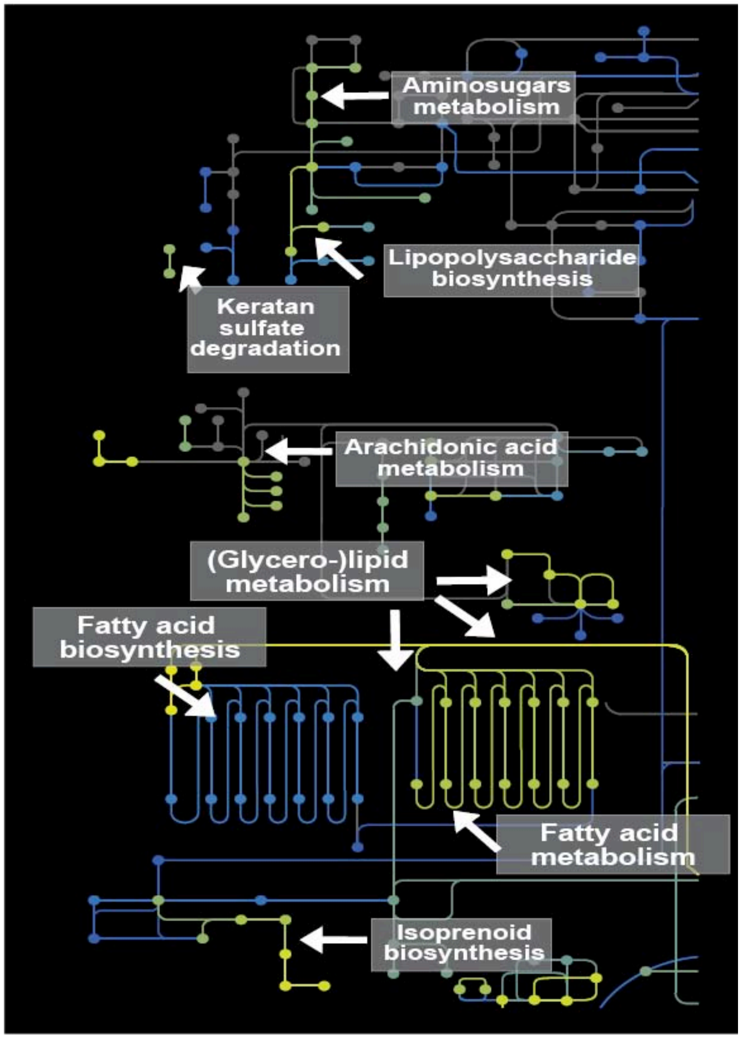


ATP synthase complex

ATP synthase complex

Gianoulis et al., PNAS (2009)

Conclusion #2: Outer Membrane Components Vary with the Environment

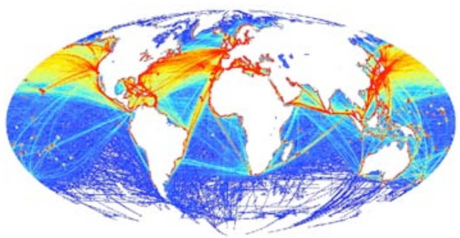


Gianoulis et al., PNAS (i2009)

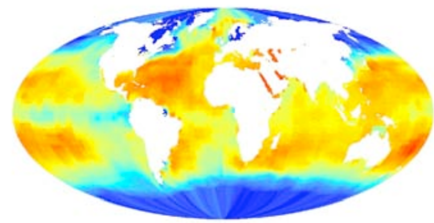
Biosensors: Beyond Canaries in a Coal Mine

Gianoulis et al., PNAS (i2009)

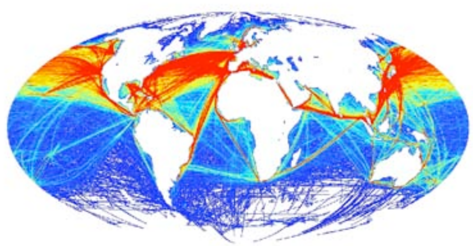
Metadata Integration



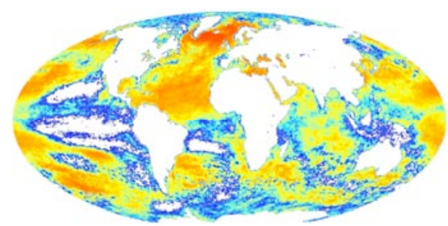
Pollution



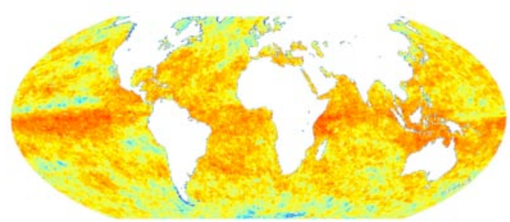
Ocean Acidification



Shipping



Climate Change (SST)



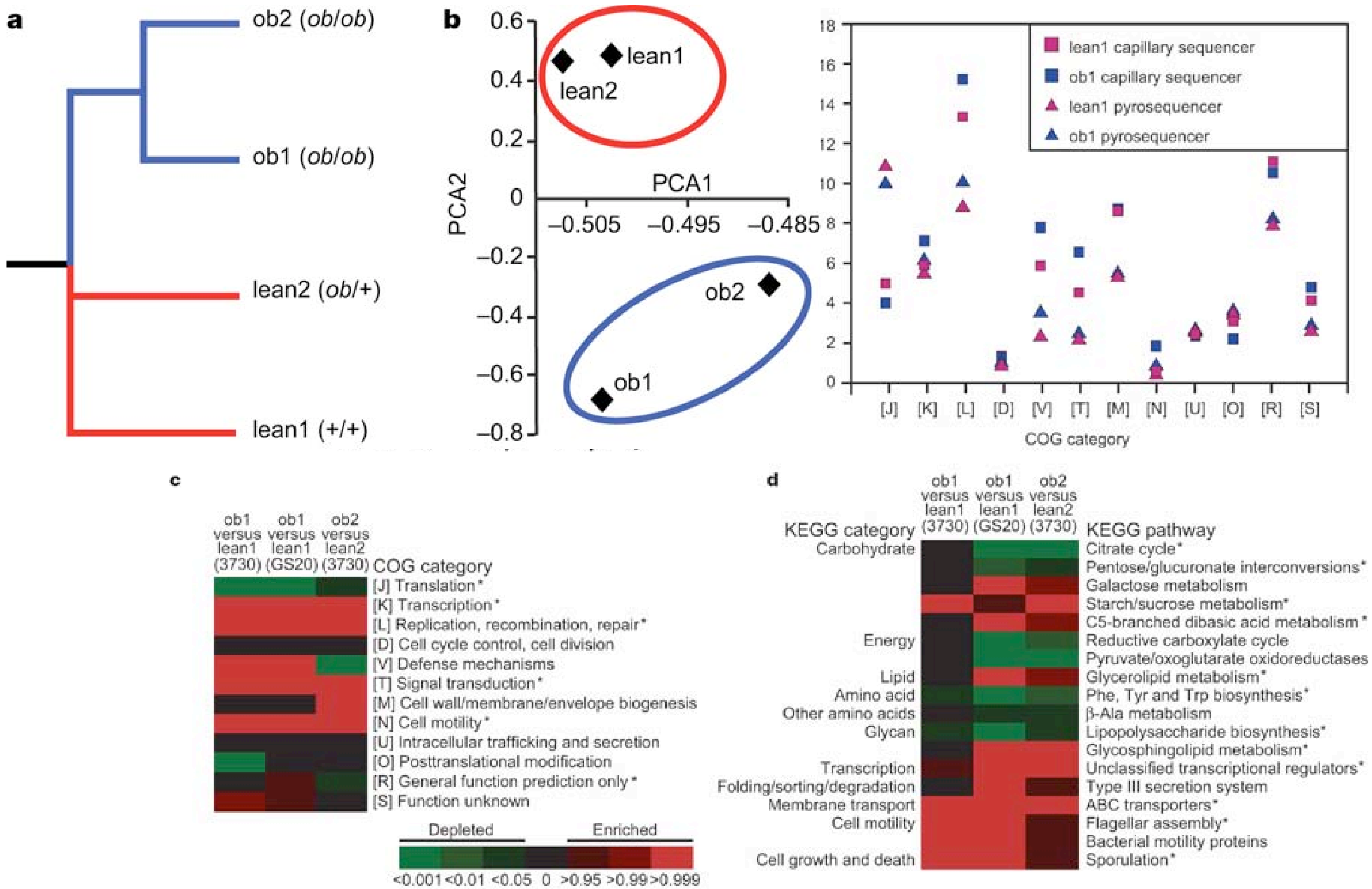
UV

Halperin et. al., (2008) *Science*

Some additional examples

- Human microbiome
- Paleogenomics

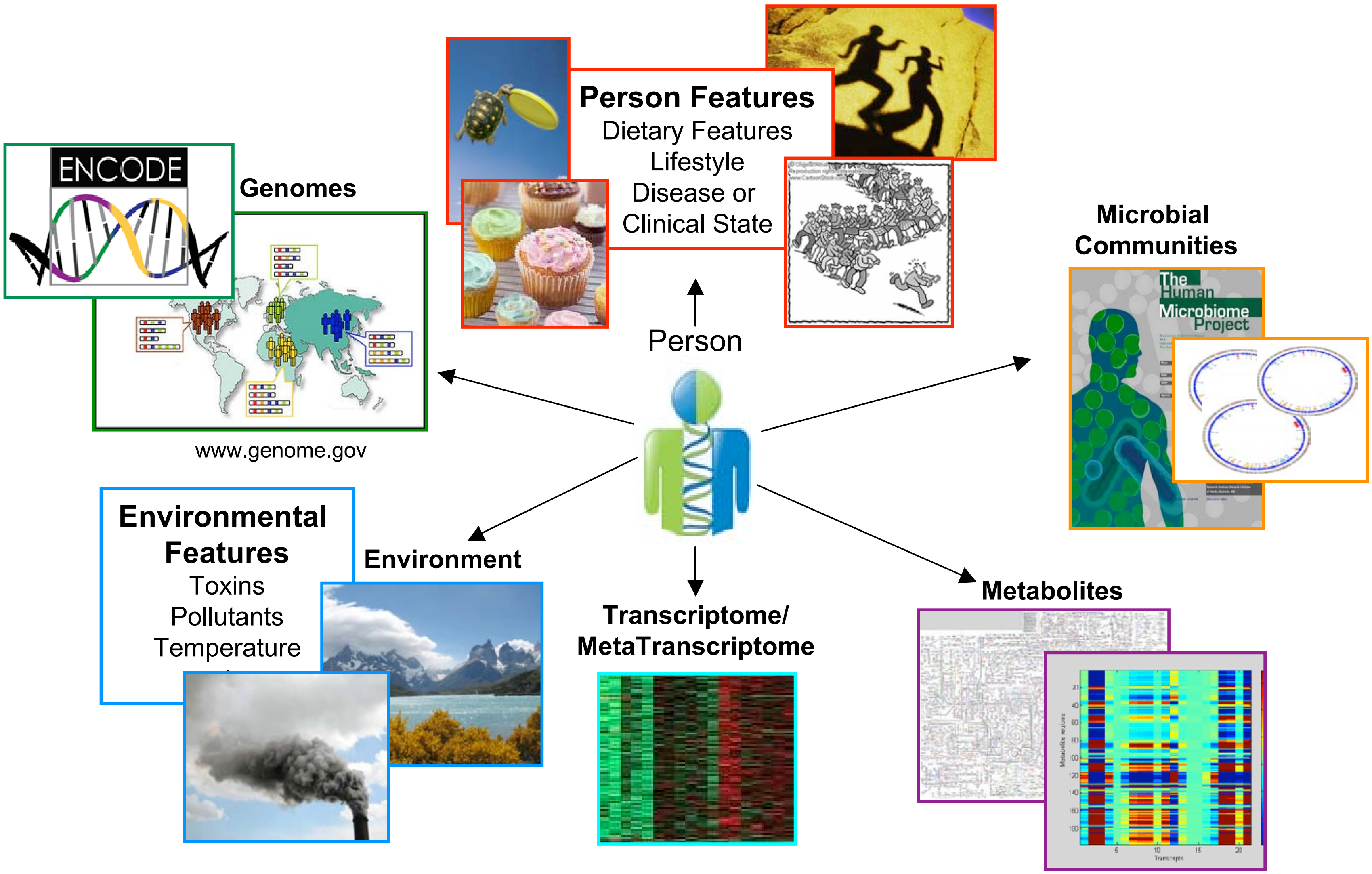
The Human (mouse) Microbiome: not much of you is you, what does that mean?



Changes in relative proportions of bacteriodites and firmicutes

Turnbaugh PJ et al, Nature 2006

Mining Complexity



Paleogenomics: dead provide new clues for the living

Stone Age genomics. Neandertal skull and femur fragments that are approximately 40,000 years old were isolated from sites of Neandertal habitation in the Altai region of Siberia.

Mackelprang & Rubin, Science 2008

Synthetic Biology: Making stuff and Learning from it

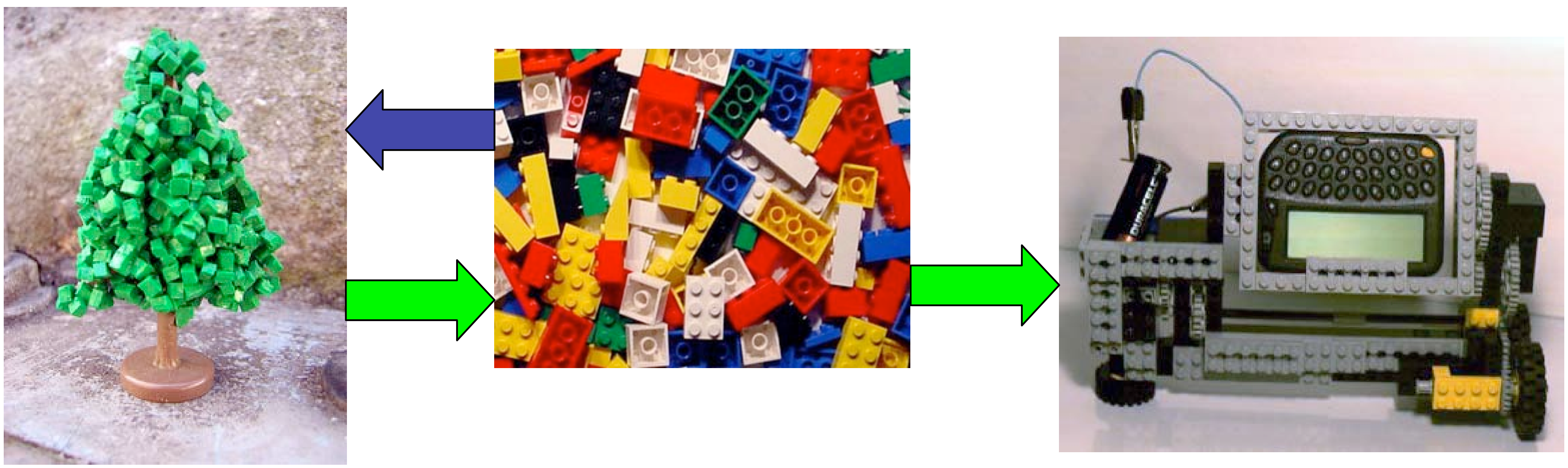
What is synthetic biology? Why do we want to do ‘it’? How far along are we? Can we learn basic scientific principles? What are some example success stories? What are some future goals and challenges?

Introduction to Synthetic Biology

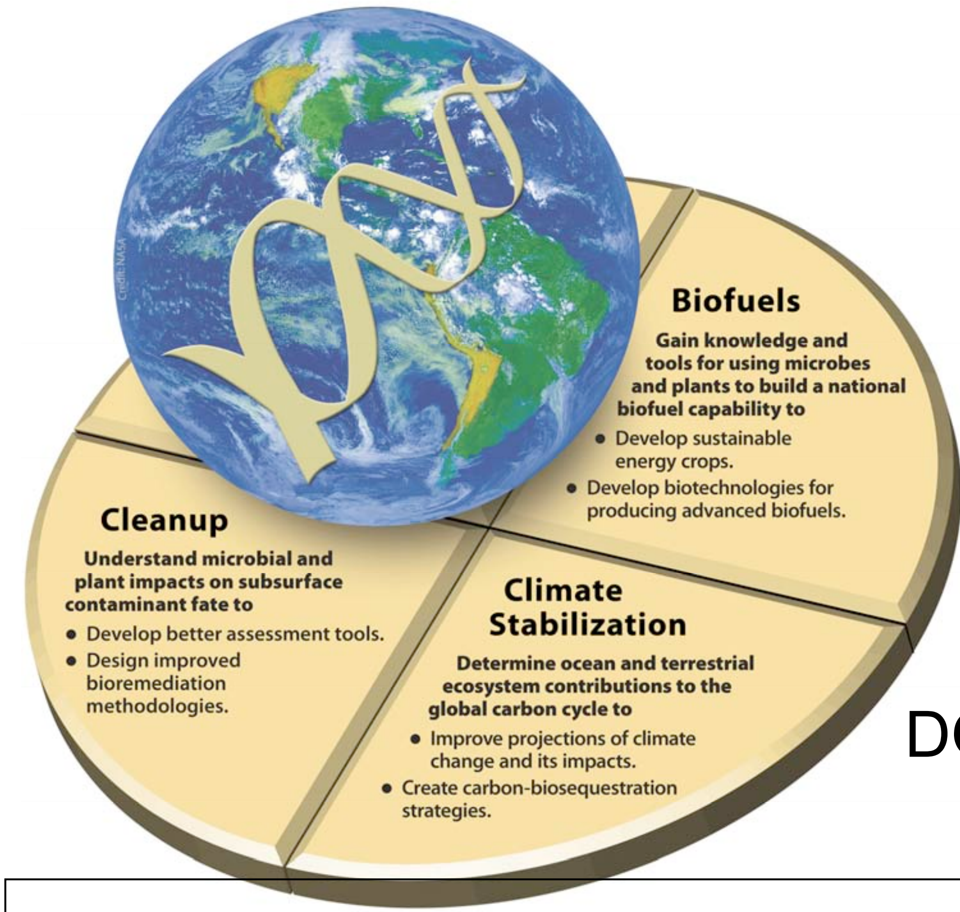
Two Inter-related Types and Goals

Synthetic Systems: Identify interchangeable parts from the **natural** world and **assemble** into new **(synthetic)** systems or functions.

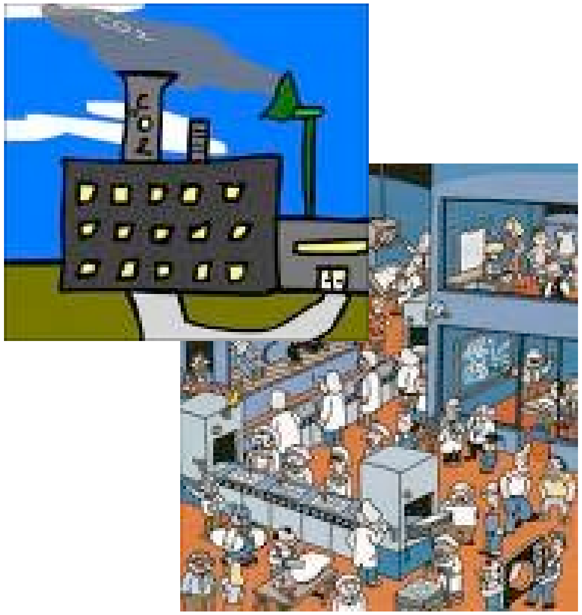
Synthetic Components: Use **unnatural** "synthetic" molecules to **reproduce** emergent behaviors from **natural** biology, with the goal of creating artificial life.



Synthesis to make "Stuff" and learn "Stuff"



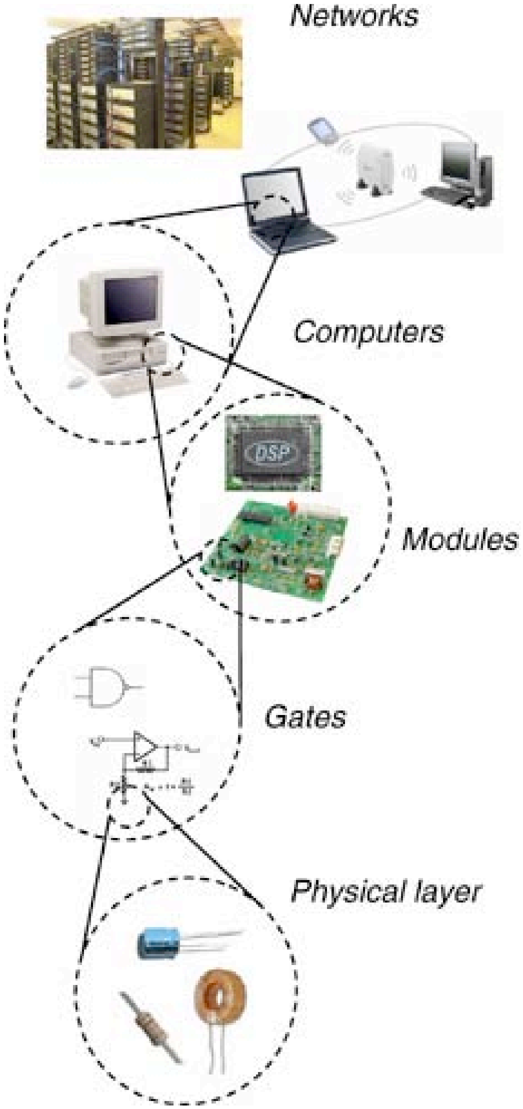
DOE

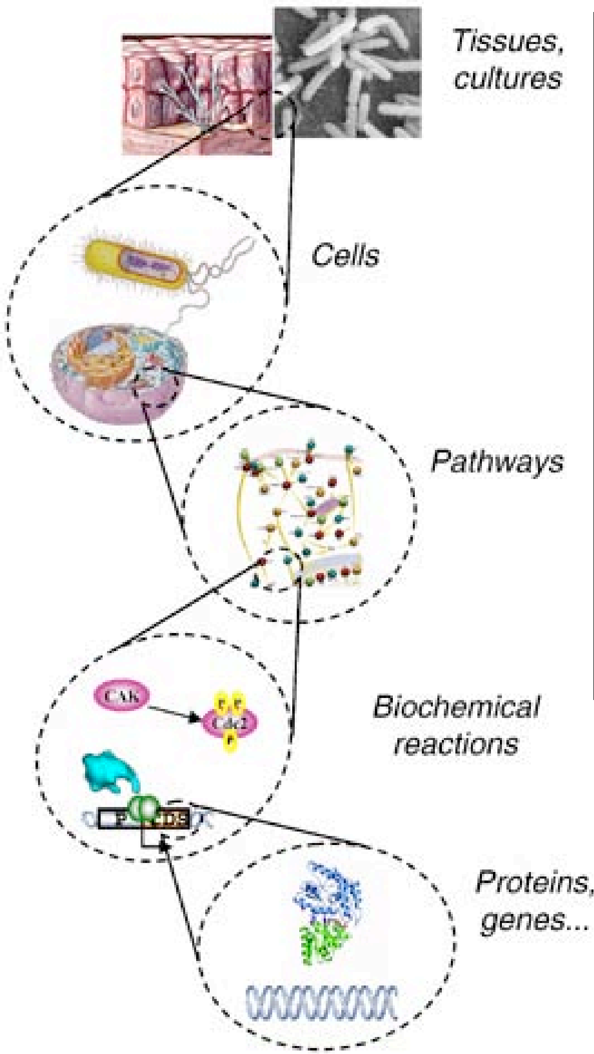


"One of the metrics of the **success of synthetic biology** will be how well the effort to assemble **existing** biological parts into **machines**, and how well the effort to create **artificial** systems that r**eproduce** the emergent properties of living systems **drives new discoveries and new theories**."

Benner, Nat Rev Gen 2006

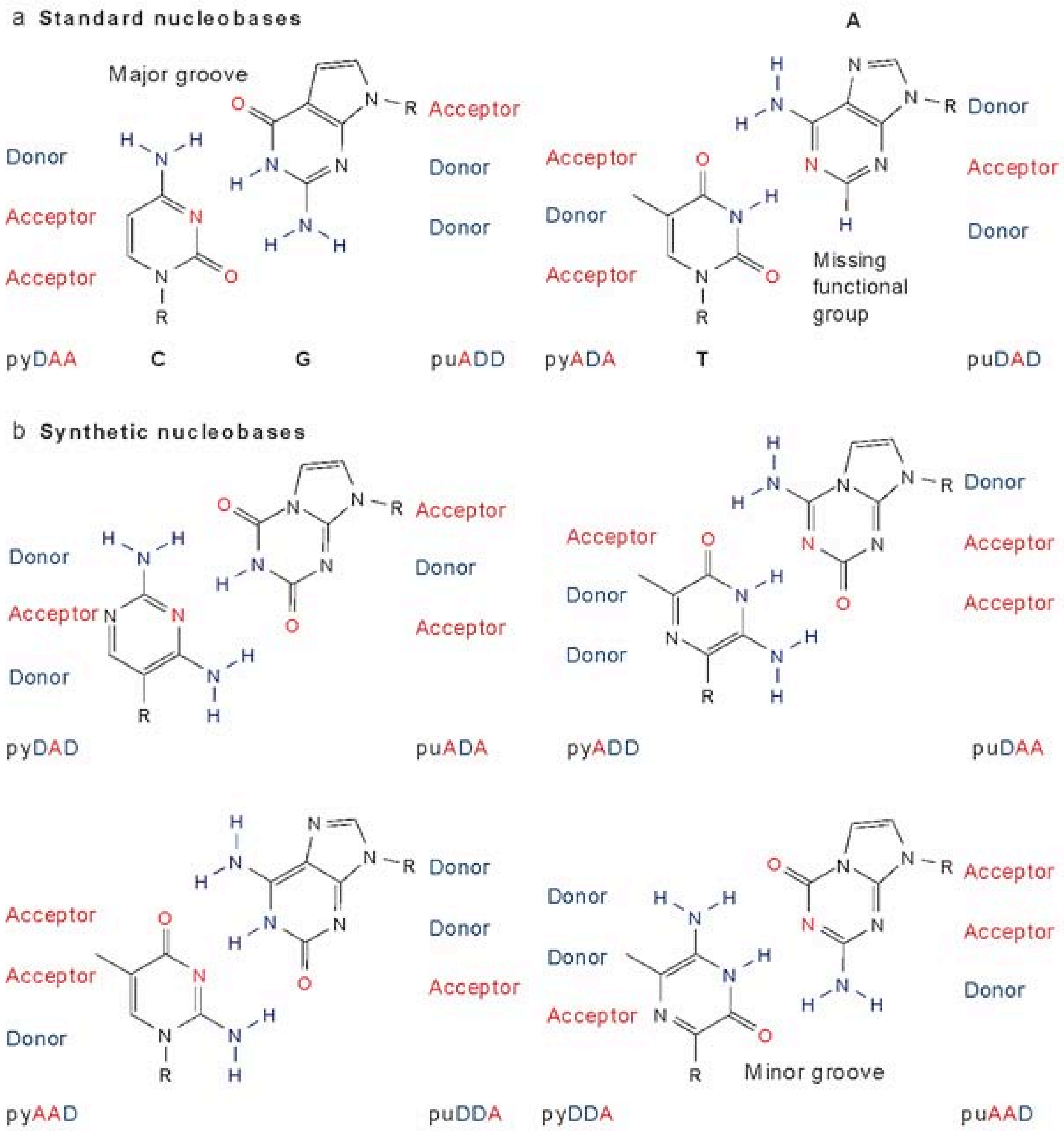
Synthetic Systems





http://www.nature.com/nature/comics/
syntheticbiologycomic/index.html

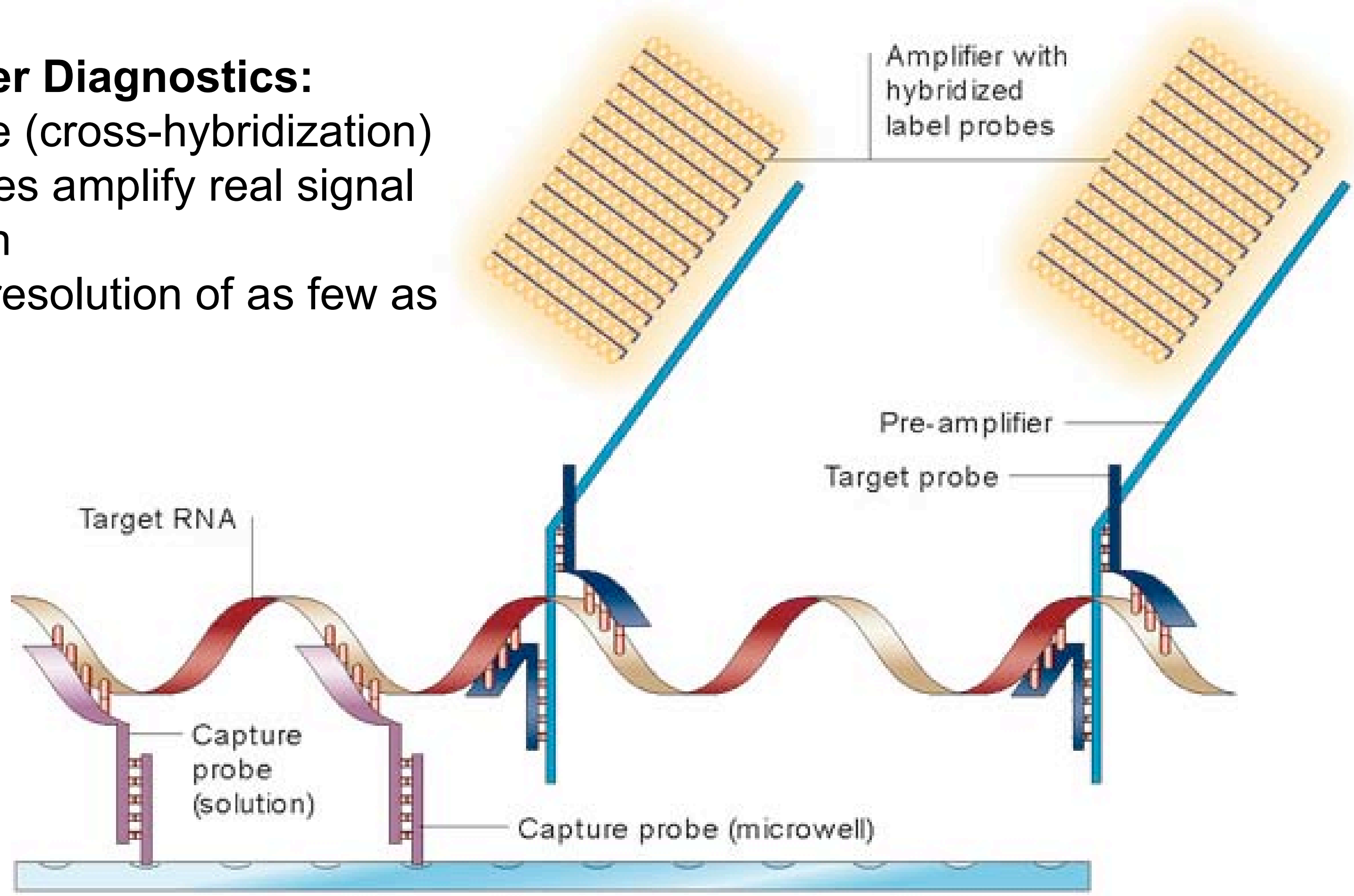
Synthetic Components



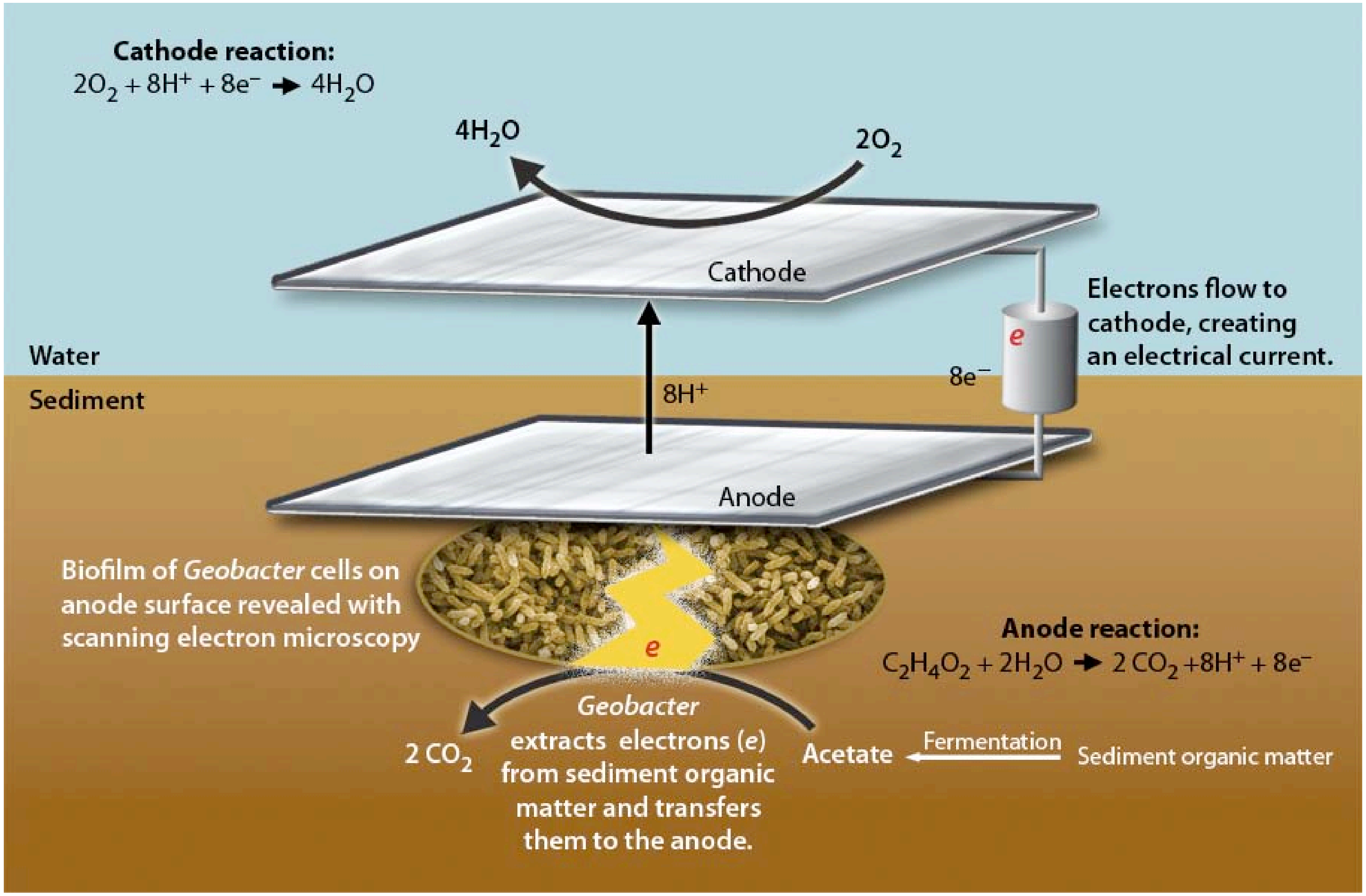
Nature Reviews | Genetics

Example 1: Clinical Diagnostics Using Expanded Genetic Alphabet more Sensitive Assays

Chiron and Bayer Diagnostics:
By reducing noise (cross-hybridization) the synthetic bases amplify real signal allowing detection
of HIV virus at a resolution of as few as 8 molecules.



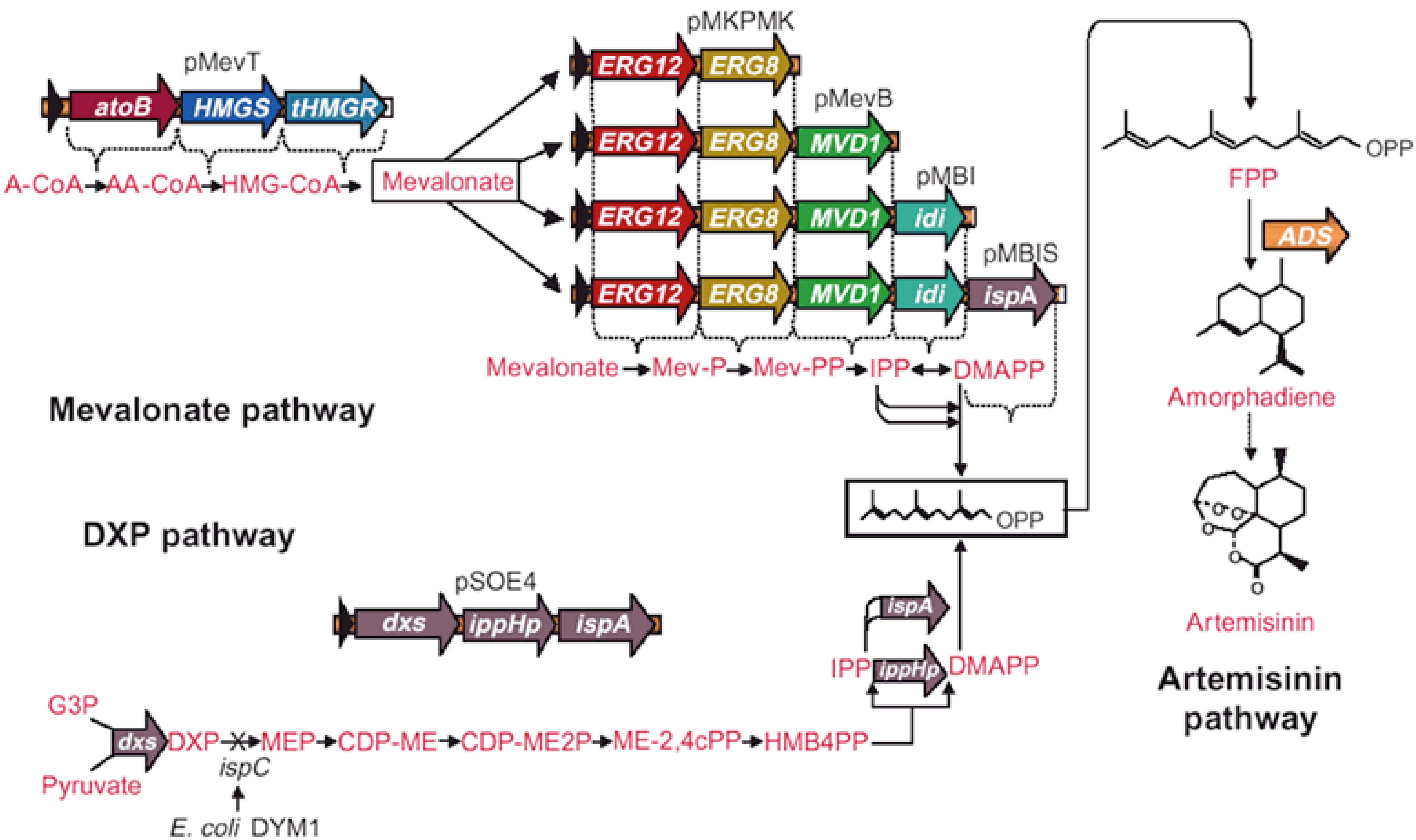
Example 2: Microbial Fuel Cell



The family Geobacteraceae can metabolize organic compounds directly at electrode surfaces, transferring electrons and producing an electrical current.

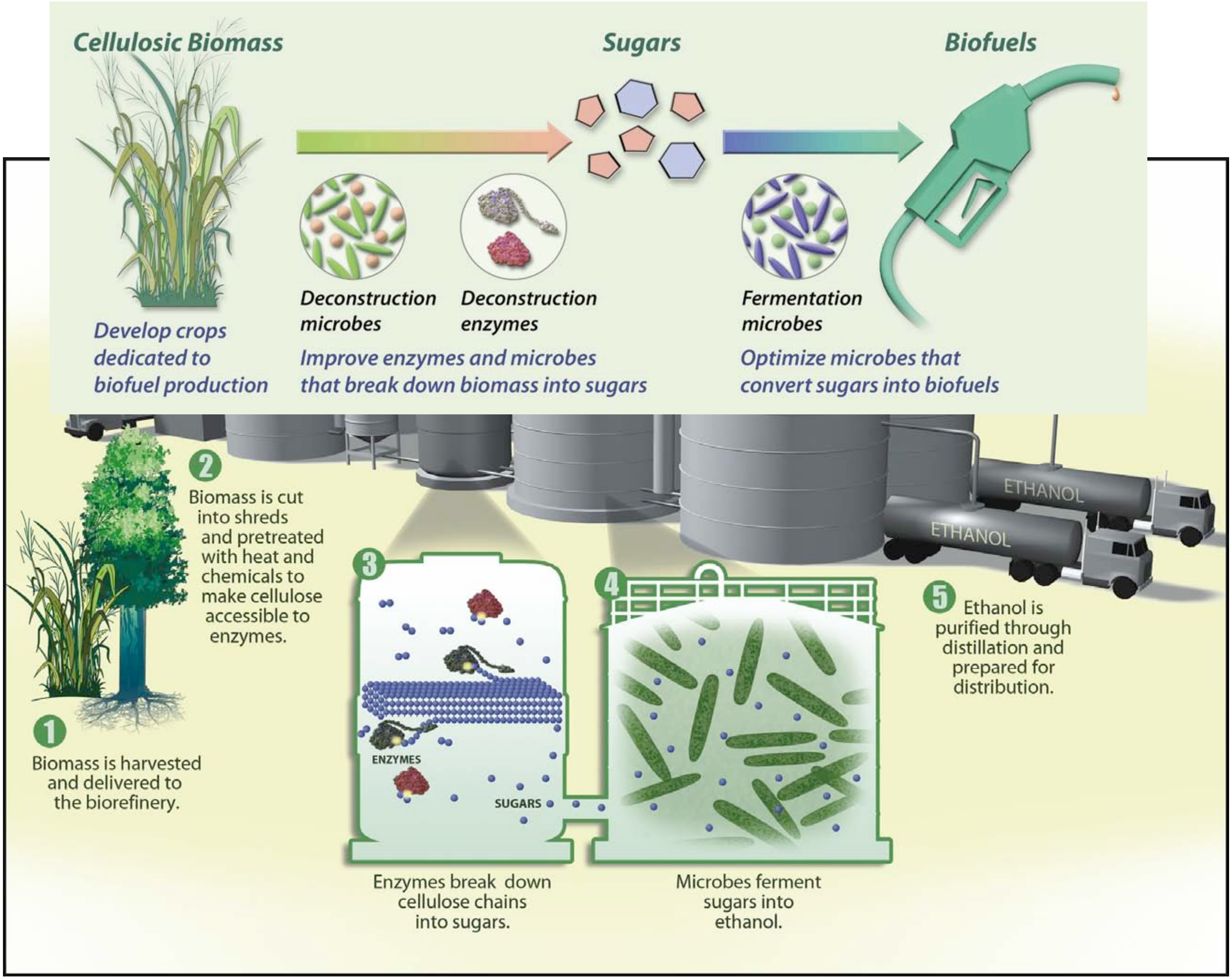
"Electrode-Reducing Microorganisms that Harvest Energy from Marine Sediments," Science 295, 483–85 (2002).

Example 3: Cellular Factories making Pharmaceuticals through metabolic engineering

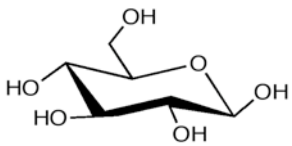


Martin VJ et al, Nat Biotech 2003

Example 4: Biofuels

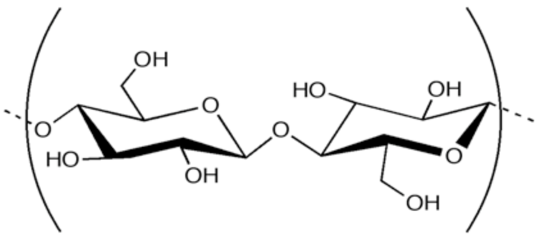


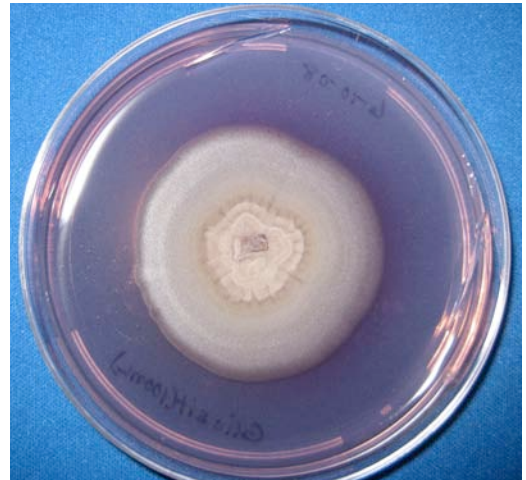
Example 5: Fuel Producer - Mycodiesel
Gliocladium roseum



Glucose

Cellulose





G. roseum
1x PDA
20 days

C_6-C_{16}

Courtesy of M. Griffin

The ultimate mash-up: identifying and revising genetic components

- Bioremediation
- Cellular factories
- Biofuels
- Testing theories