



Coupling Science to the Future of High Performance Computing

Rick Stevens

Argonne National Laboratory

The University Chicago

Stevens@anl.gov



*Argonne National Laboratory is managed by
The University of Chicago for the U.S. Department of Energy*

A Paradigm Shift

Two battleships assigned to the training squadron had been at sea on maneuvers in heavy weather for several days. . .

The visibility was poor with patchy fog, so the captain remained on the bridge keeping an eye on all activities.

Shortly after dark, the lookout on the wing of the bridge reported, "**Light, bearing on the starboard bow.**"

"**Is it steady or moving astern?**" the captain called out.

Lookout replied, "**Steady, captain**", which meant the ships were on a dangerous collision course.

The captain then called to the signalman, "**Signal that ship we are on a collision course, advise you change course 20 degrees.**"

The lookout replied "**No change in position, captain**", the captain registering increasing alarm.

The captain said, "**Send, I'm a captain, change course 20 degrees.**"

"**I'm a seaman second class**", came the reply. "**You had better change course 20 degrees.**"

By that time, the captain was furious. He spat out, "**Send, I'm a battleship. change course 20 degrees.**"

A Paradigm Shift

Two battleships assigned to the training squadron had been at sea on maneuvers in heavy weather for several days. . .

The visibility was poor with patchy fog, so the captain remained on the bridge keeping an eye on all activities.

Shortly after dark, the lookout on the wing of the bridge reported, "**Light, bearing on the starboard bow.**"

"**Is it steady or moving astern?**" the captain called out.

Lookout replied, "**Steady, captain**", which meant the ships were on a dangerous collision course.

The captain then called to the signalman, "**Signal that ship we are on a collision course, advise you change course 20 degrees.**"

The lookout replied "**No change in position, captain**", the captain registering increasing alarm.

The captain said, "**Send, I'm a captain, change course 20 degrees.**"

"**I'm a seaman second class**", came the reply. "**You had better change course 20 degrees.**"

By that time, the captain was furious. He spat out, "**Send, I'm a battleship. change course 20 degrees.**"

Back came the flashing light, "**I'm a lighthouse.**"

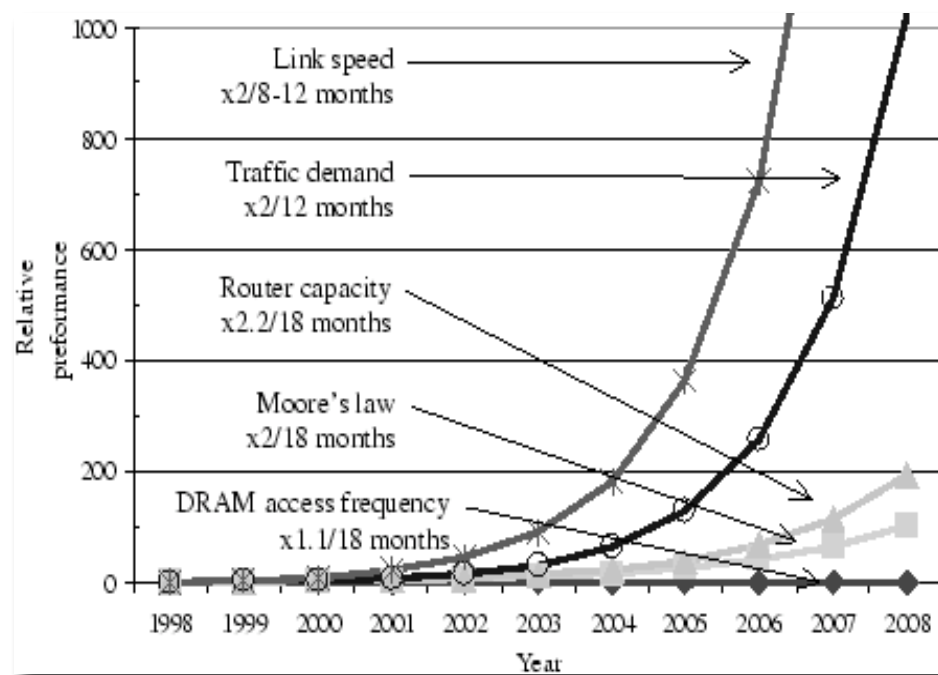
Some Future Trends



H O W T O
T H I N K
A B O U T T H E
an Arlington Institute Workshop
Future
with John Petersen

Thinking about Trends

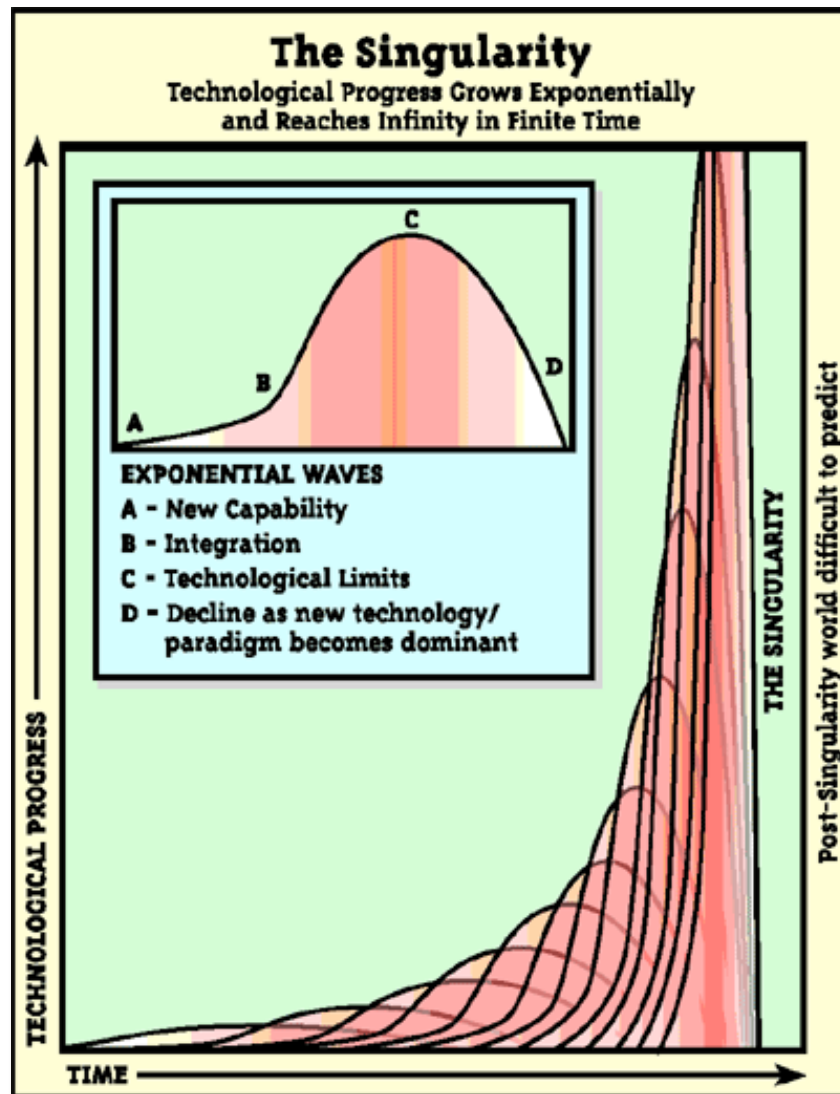
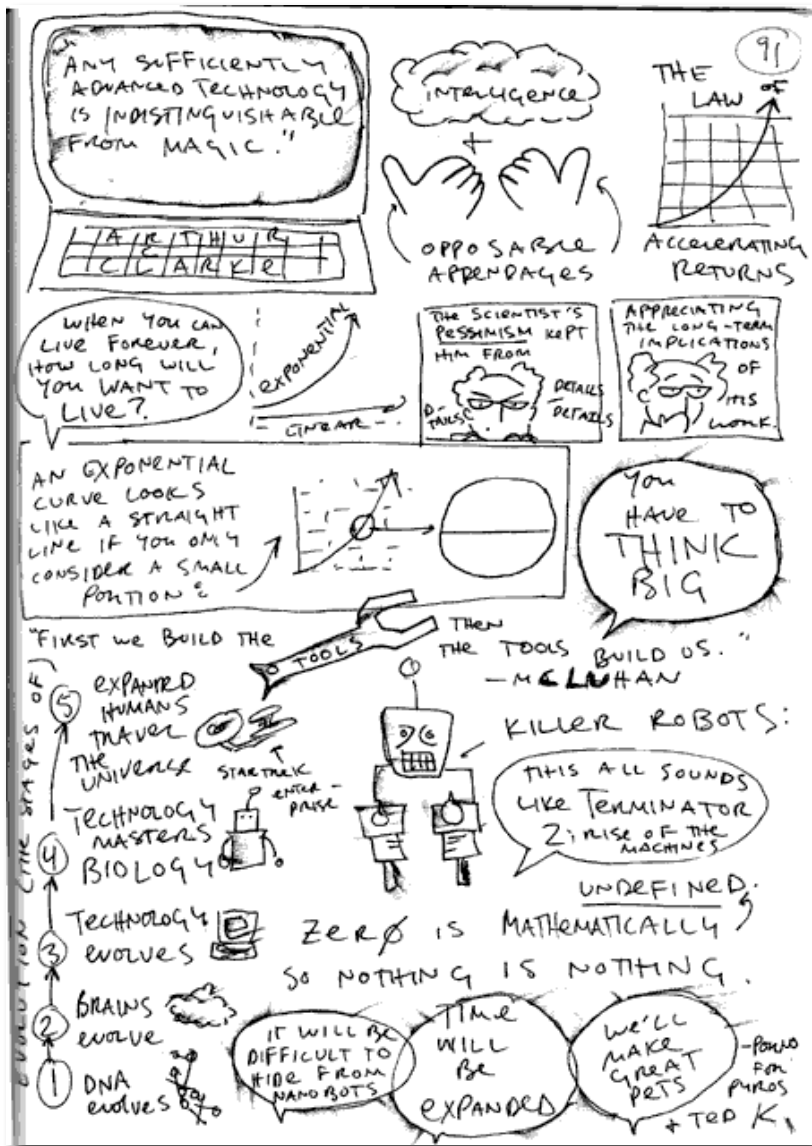
- Slow forces that build up over time and change things slowly but surely
 - Moore's Law
 - Global Warming
 - Wireless
 - Digital Imaging
- Unanticipated Rapid Impact
 - Peer-to-peer
 - Tipping points
 - Social Network Applications



Thinking about Trends

- Slow forces that build up over time and change things slowly but surely
 - Moore's Law
 - Global Warming
 - Wireless
 - Digital Imaging
- Unanticipated Rapid Impact
 - Peer-to-peer
 - Social Network Applications





Humanity's Top Ten Problems for next 50 years

1. ENERGY
2. WATER
3. FOOD
4. ENVIRONMENT
5. POVERTY
6. TERRORISM & WAR
7. DISEASE
8. EDUCATION
9. DEMOCRACY
10. POPULATION



2007	7	Billion People
2050	8-10	Billion People

Richard Smalley's Top Ten List

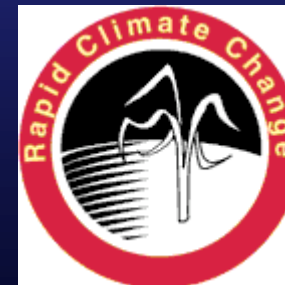


Forces of Change

Trends: continuously building changes

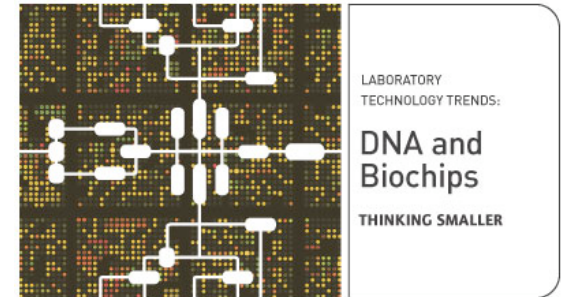


Wild Cards: low probability, extremely high impact events



The Scientific “Super Trends”

- Three Major Developments
 - Complexity of Problems \Rightarrow Interdisciplinary Science
 - Larger Teams with more International Collaboration managed as Virtual Organizations
 - Mega-capital investment in scientific facilities will be limited
 - Micro/Nano Scale Instrumentation \Rightarrow Benchtop Revolution
 - Empowering experimentalists to ask new classes of questions (e.g. looking at all proteins at once in a cell instead of just one)
 - Decreasing value of historical datasets
 - High-Throughput and Robotics \Rightarrow Data Volume Revolution
 - Experimentalists will begin to dominate data generation and will require access to increasingly capable information technology infrastructures
 - Cross cutting experiments become possible

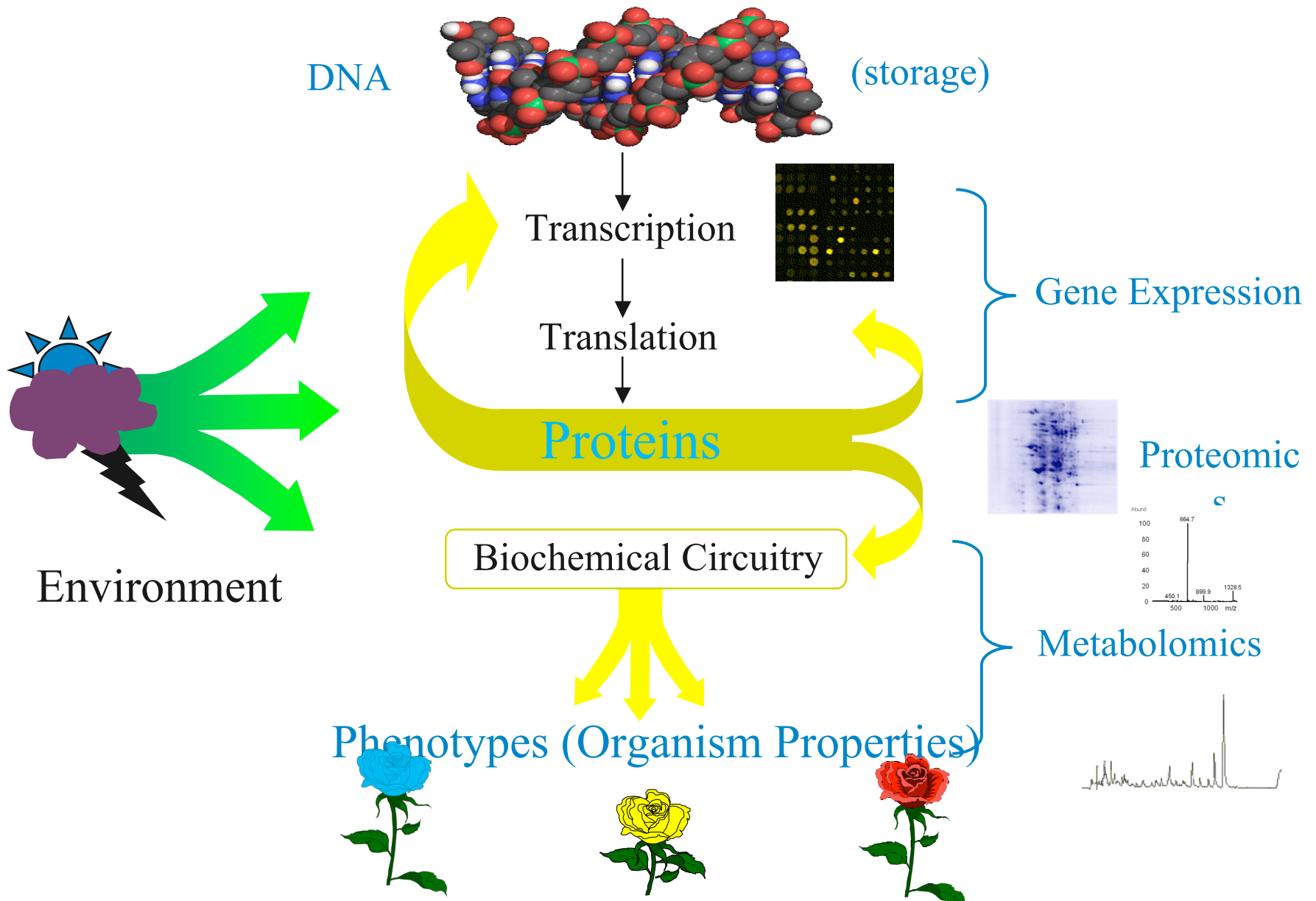


Modeling Microbiology





Predicting Phenotypes from Genotypes — the prediction of system level behavior from collections of functional components

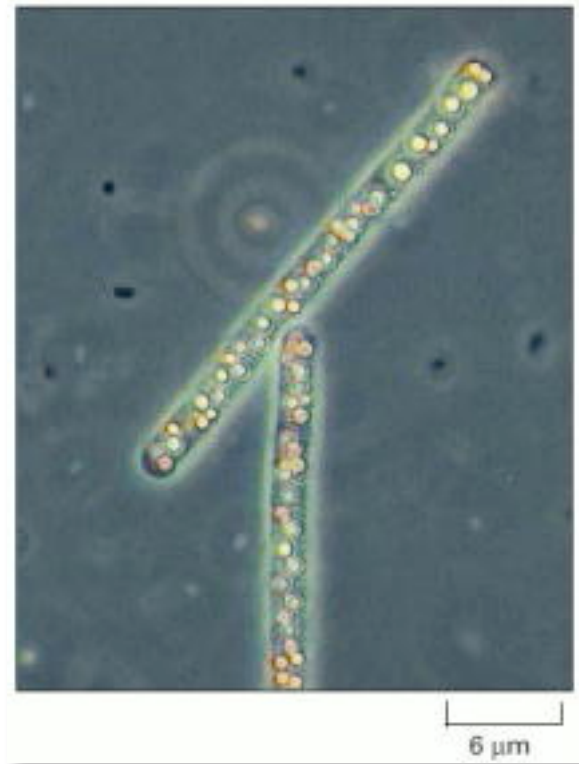


Adapted From Bruno Sobral VBI

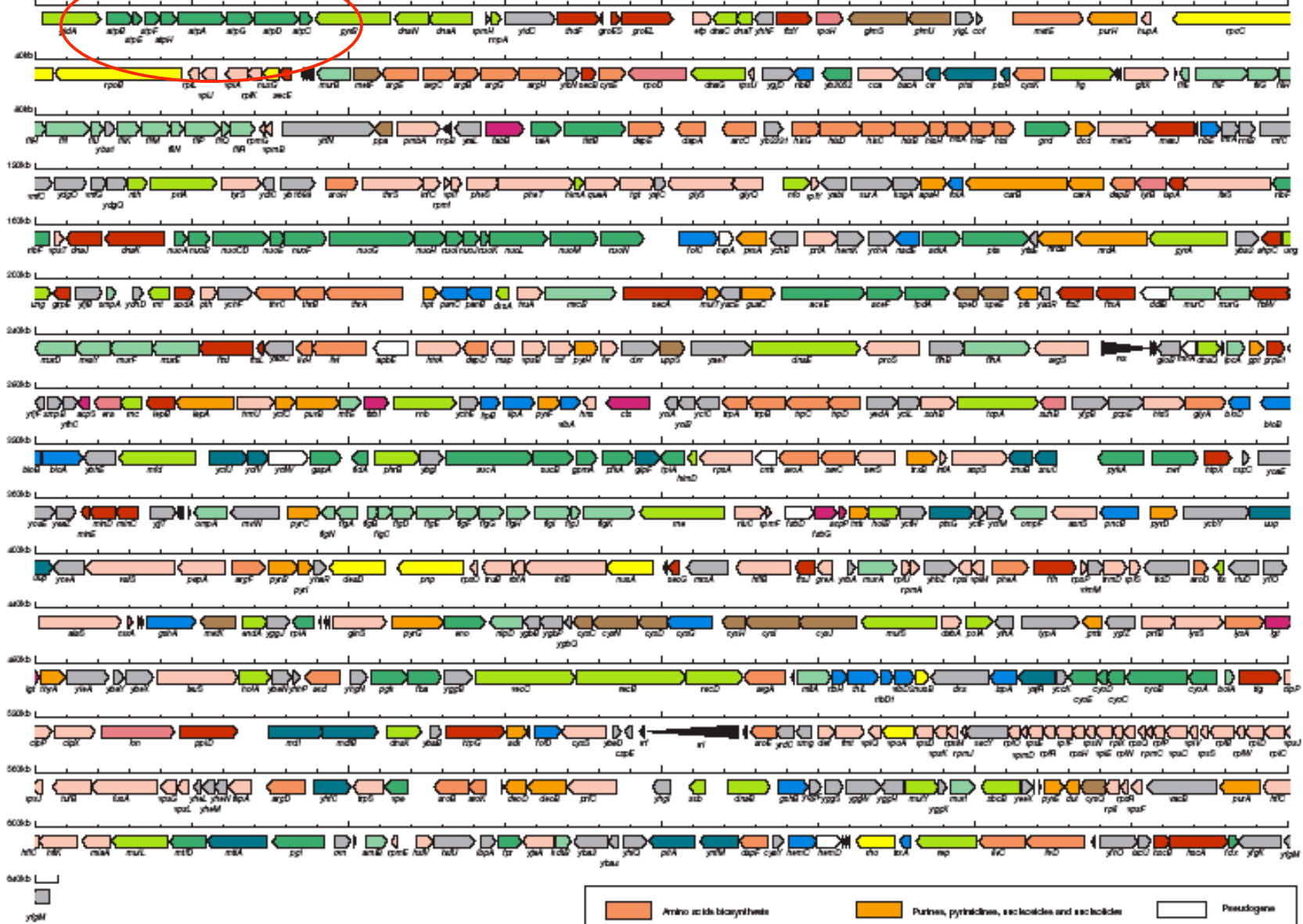
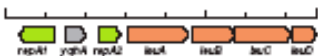
Genome + Environment = Phenotype

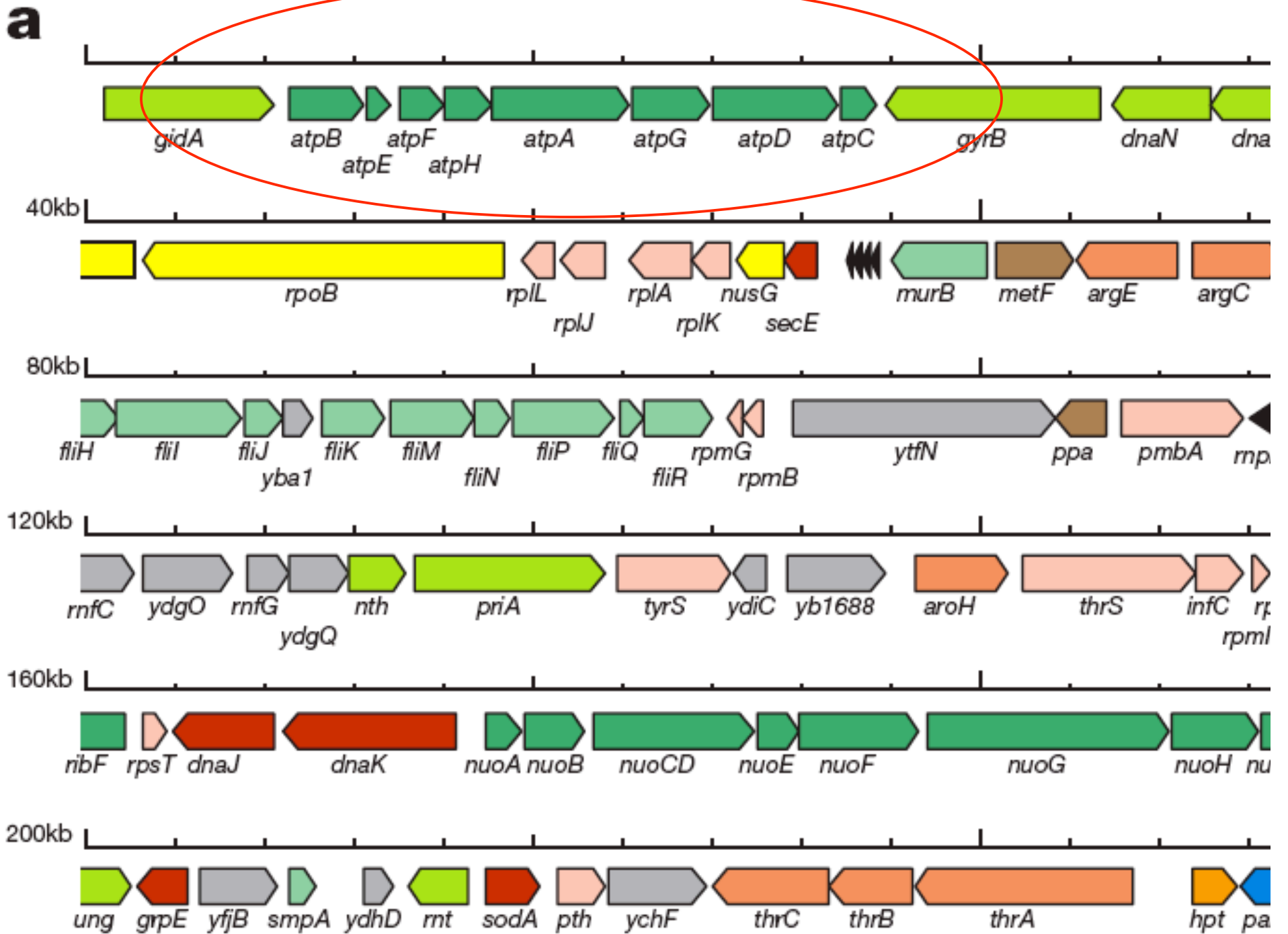
Microbial Organisms are Interesting to Study

- Extremely Diverse Metabolisms
- Window on Biodiversity
- Ancient Origins
- Foundation of the Biosphere
- Agents of Symbiogenesis
- Infectious Disease
- Human Microbiota and Metagenomes
- Complex Community Structures
- Industrial and Agricultural Applications
- Biotechnology Applications
- Biofuels and Alternative Feed stocks
- Inexpensive and Experimentally Tractable



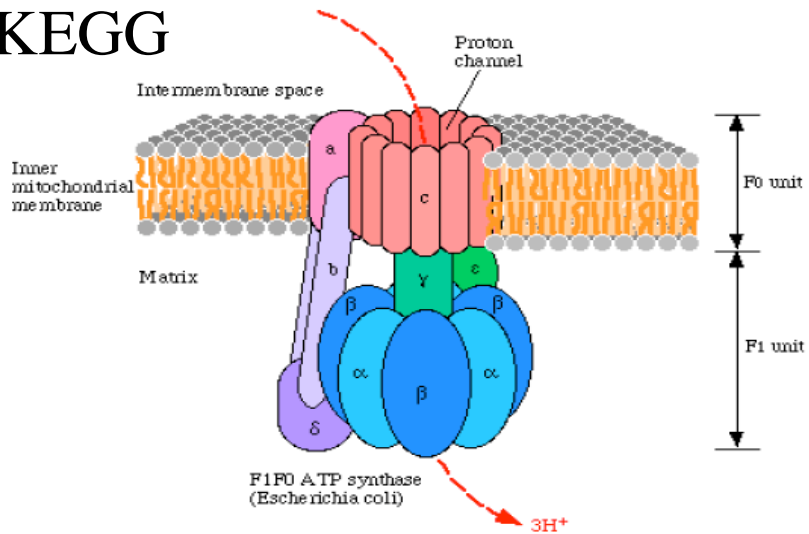
Beggiatoa, which lives in sulfurous environments, gets its energy by oxidizing H_2S and can fix carbon even in the dark. Note the yellow deposits of sulfur inside the cells. (Courtesy of Ralph W. Wolfe.)

a**b****c**



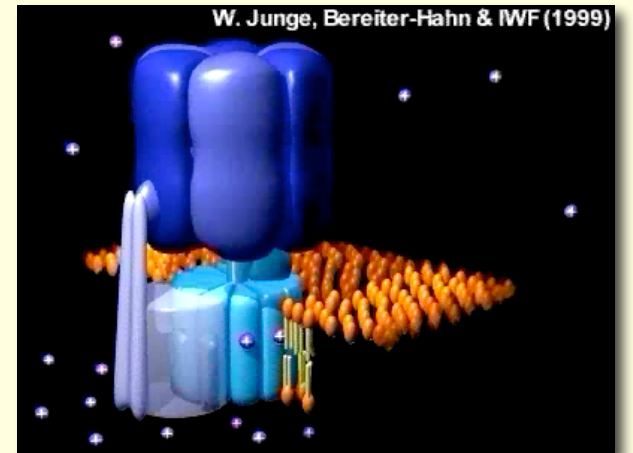
From KEGG

ATP SYNTHESIS



F-type ATPase (Bacteria)

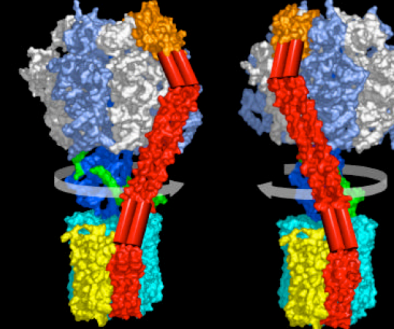
beta	alpha	gamma	delta	epsilon	c	a	b
------	-------	-------	-------	---------	---	---	---



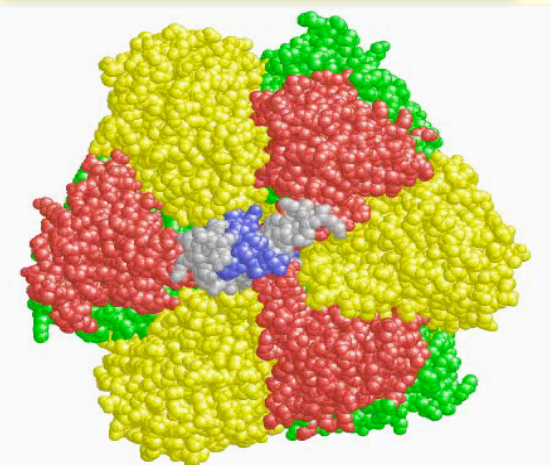
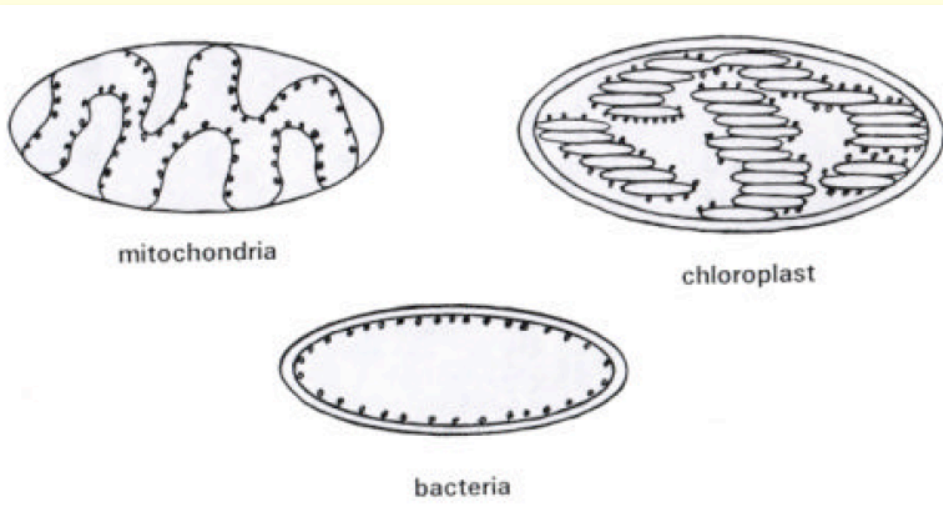
W. Junge, Bereiter-Hahn & IWF (1999)

ATP Synthesis

ATP-driven H⁺ Pumping

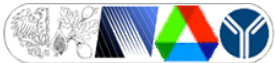
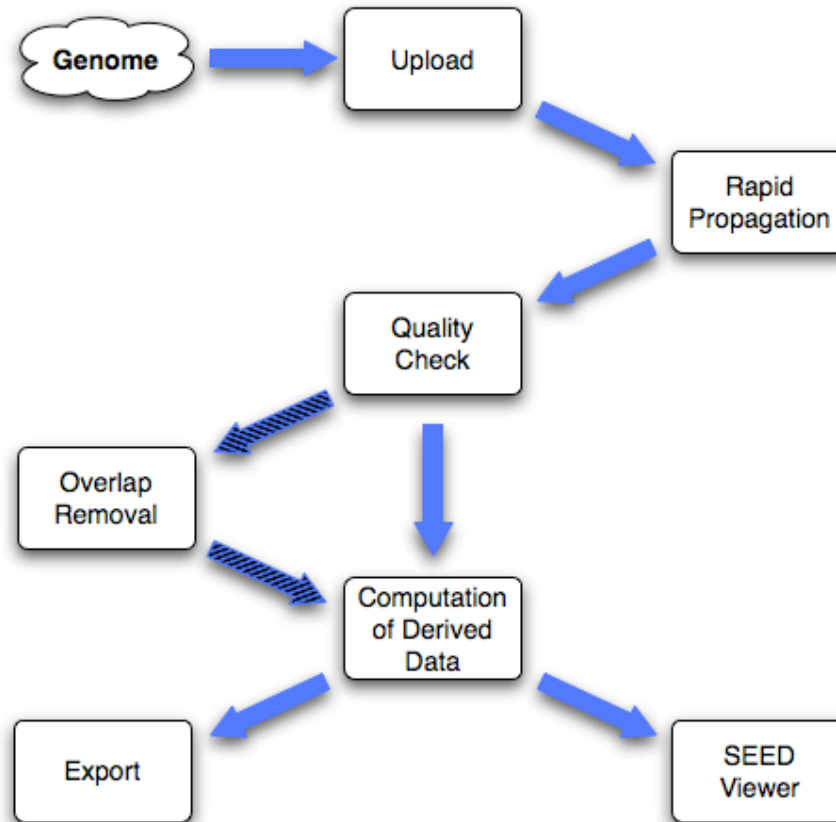


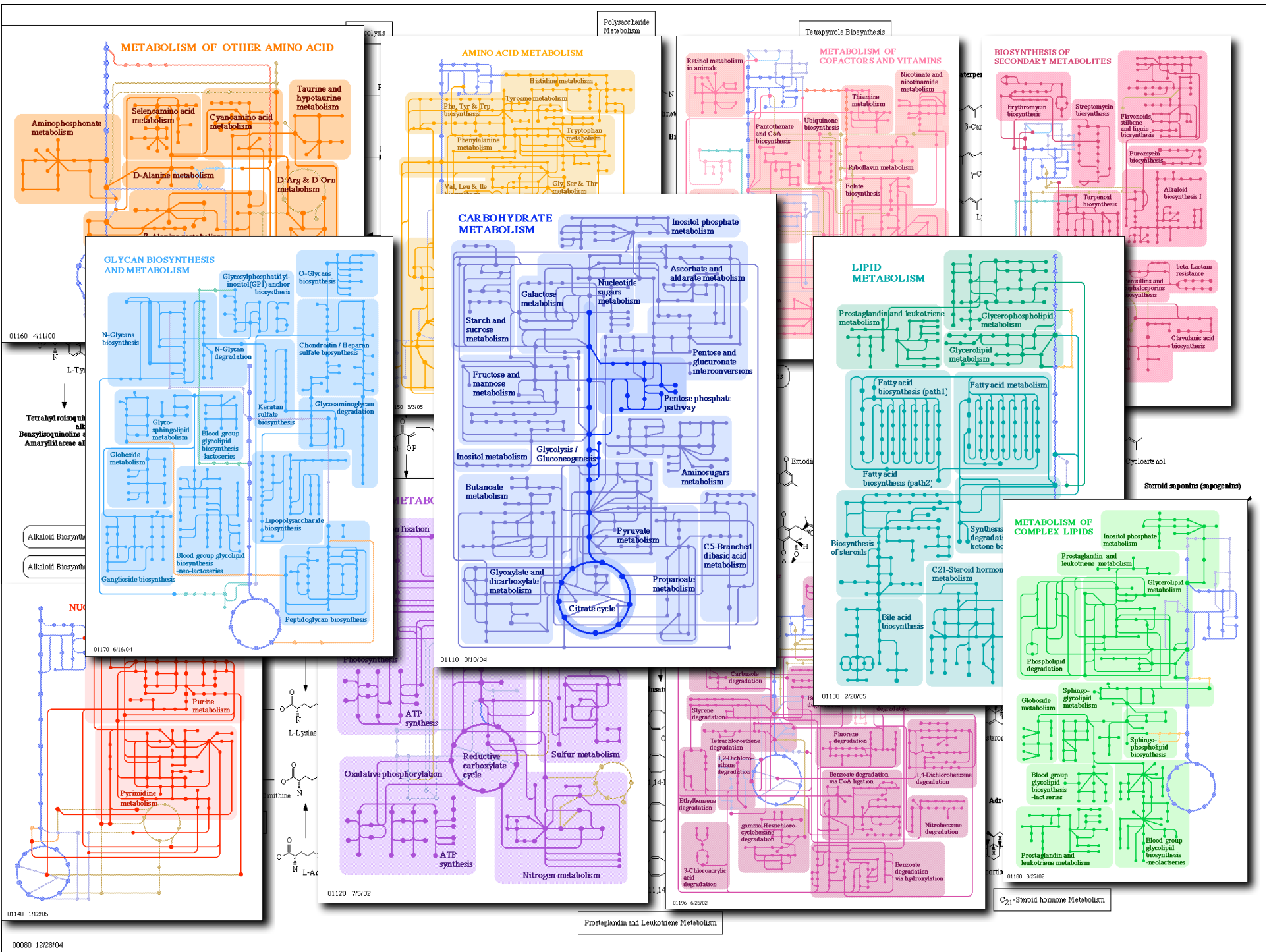
Hardy et al (2003) J. Biomembr. Bioenerg. 35, 398-397

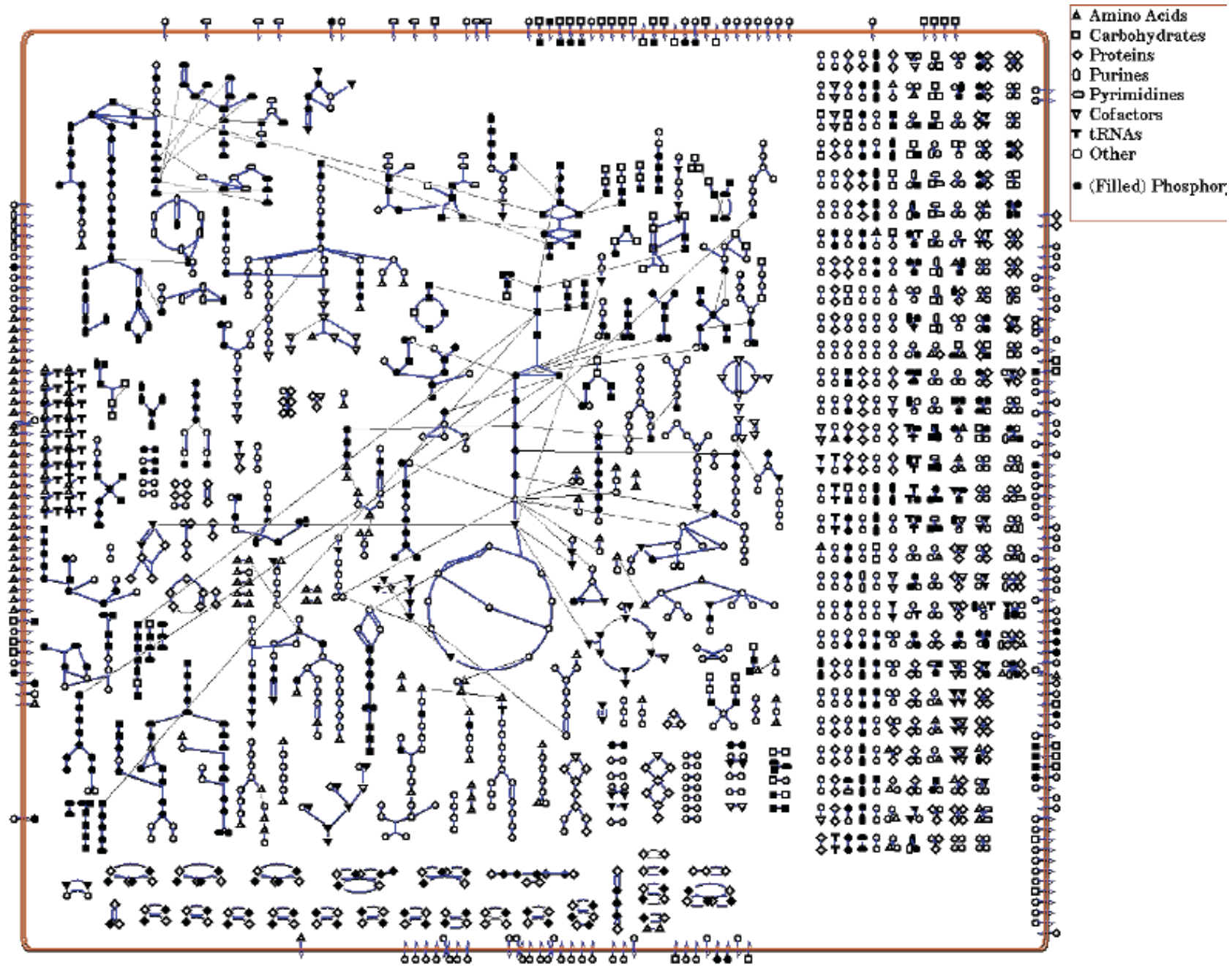


A top view of alpha3-beta3-gamma
By Hongyun Wang & George Oster, U.C. Berkeley

RAST Pipeline



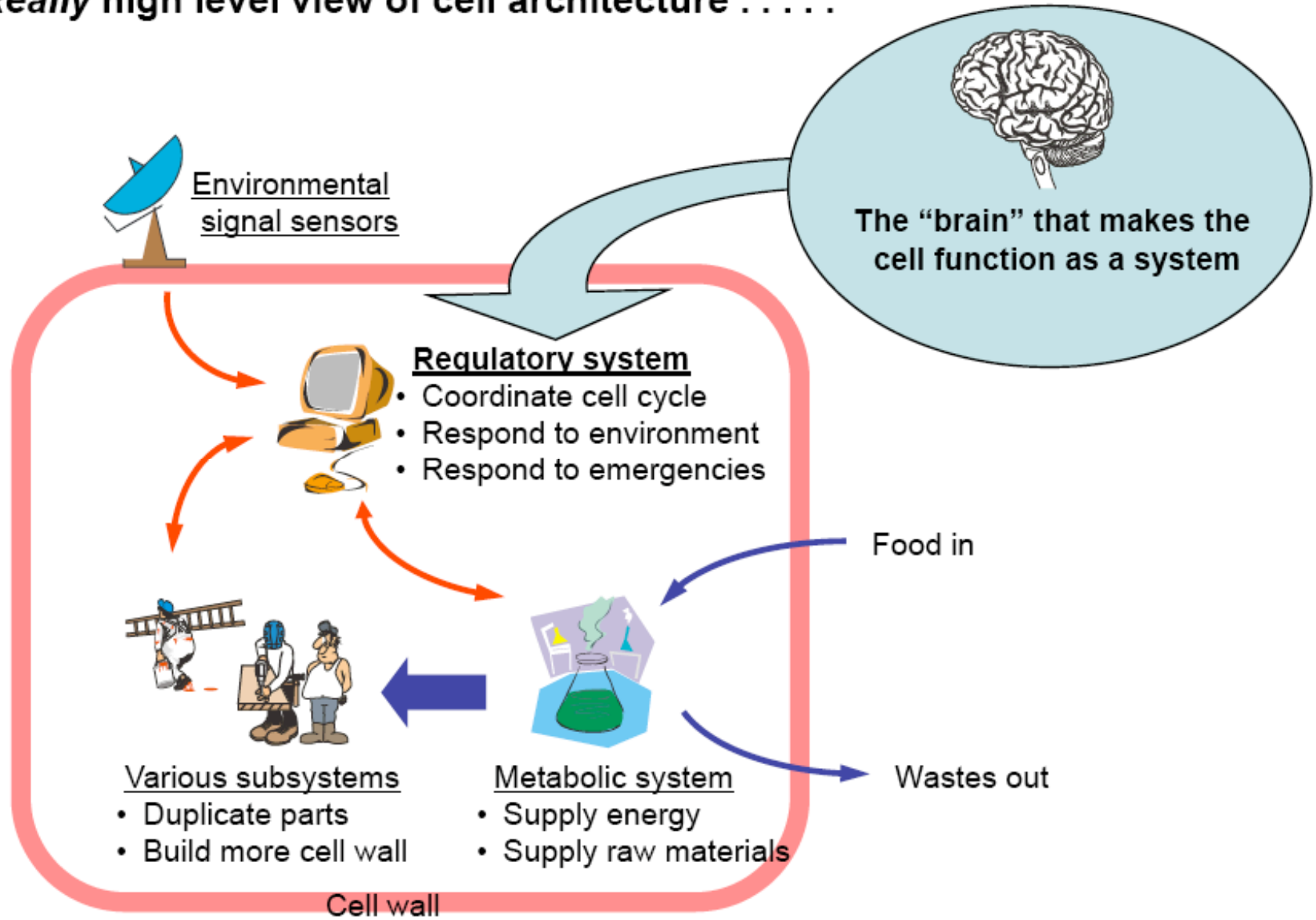




***E. coli* K-12 Metabolic Overview**

Source: EcoCyc

Really high level view of cell architecture



Regulation of transcription factors in E. coli

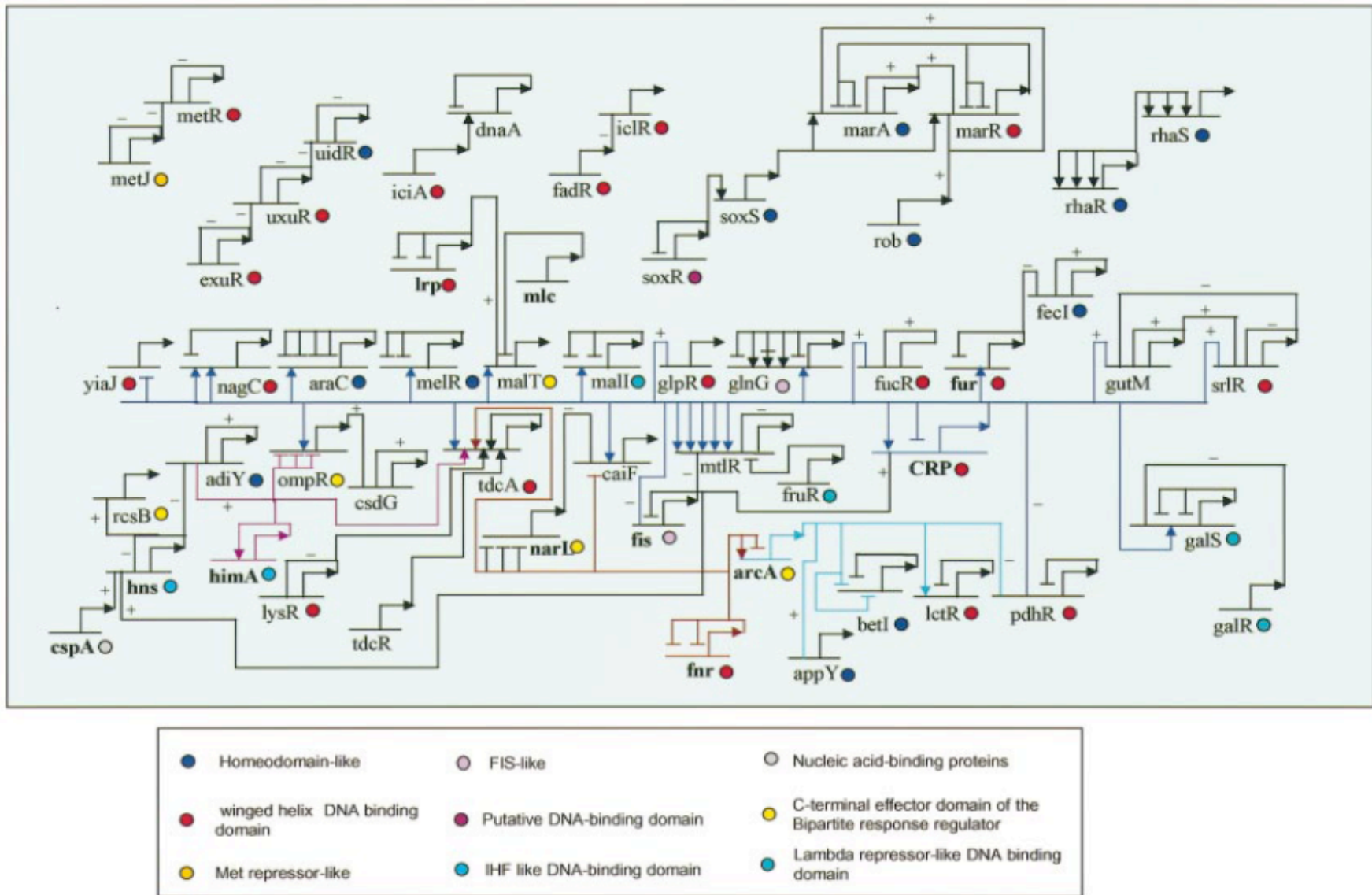
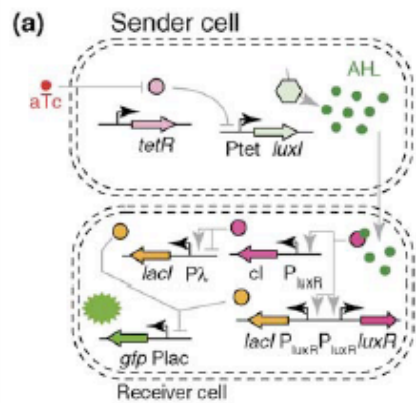
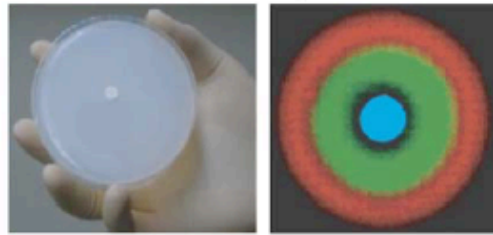


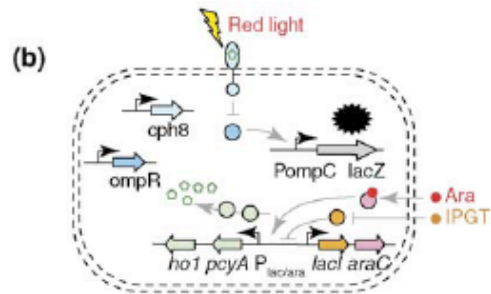
Figure 3. The transcription factor regulatory network in *E. coli*. When more than one transcription factor regulates a gene, the order of their binding sites is as given in the figure. An arrowhead is used to indicate positive regulation when the position of the binding site is known. A horizontal bar is used to indicate negative regulation when the position of the binding site is known. In cases where only the nature of regulation is known, without binding site information, + and - are used to indicate positive and negative regulation, respectively. These examples may be indirect rather than direct regulation. The DBD families are indicated by circles of different colours as given in the key. The names of global regulators are in bold.



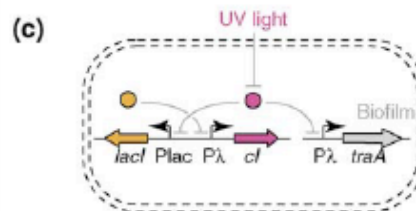
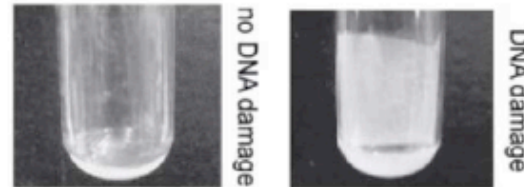
Produce two-dimensional patterns



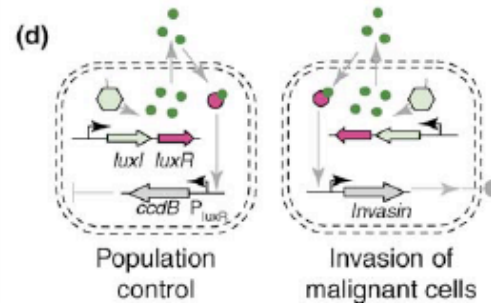
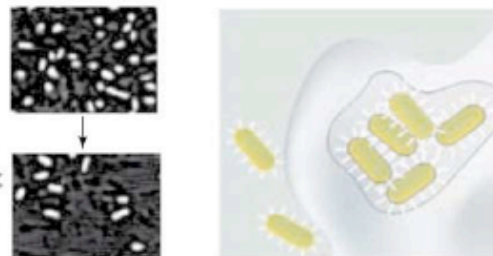
Take your picture



Form a biofilm



Commit suicide or kill malignant cells



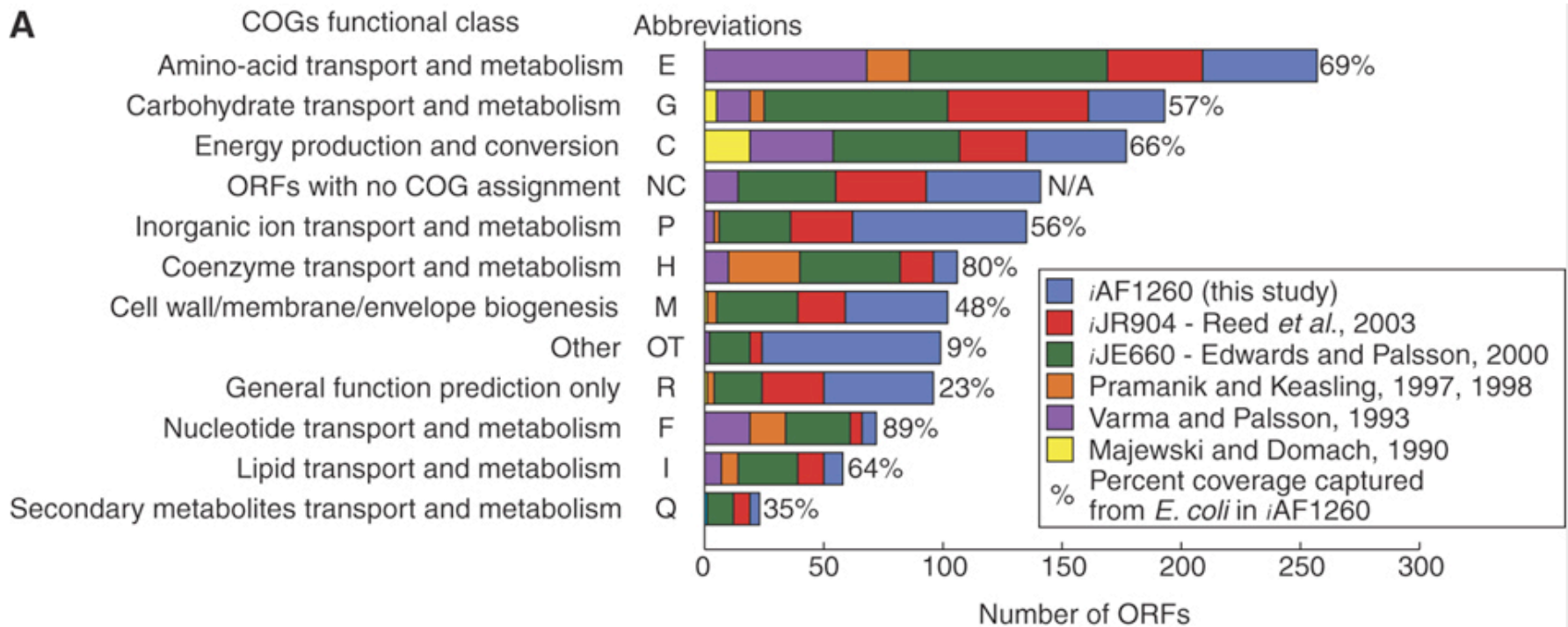
Current Opinion in Biotechnology

DNA parts for programming bacteria.

Name	Genes ^a	Performance
Sensors – small molecule inducers		
Lac		
Tet		
Ara		
Ara-lac		
Genetic circuits – switches and logic		
Inverter		
Biphasic switch		
Toggle switch		
Cell-cell communication ^d		
Genetic circuits – dynamic responses		
Pulse generator		
Actuators		
Suicide		
Biofilm		
Adhesion/invasion		

17 Years of Progress in FBA Model Development (E. coli)

Adam Feist, Jenny Reed, Chris Henry, Peter Karp, Bernard Palsson



Molecular Systems Biology 3 Article number: 121

doi:10.1038/msb4100155

Published online: 26 June 2007



	A	B	C	D	E	F	G	H
1	GenomeID	Name	NumGeneSEED	NumGenesInFIGFam	RxnsFromBsub	NumECsSEED	NumRxnsSEED	
2	36873.1	Burkholderia xenovorans LB400	8738	2386	3411	2041	1053	
3	134537.1	Burkholderia fungorum	8445	2063	2984	1603	850	
4	216591.1	Burkholderia cenocepacia J2315	7820	1841	2345	1963	857	
5	264198.3	Ralstonia eutropha JMP134	6535	1535	2343	1601	830	
6	269483.3	Burkholderia cepacia R18194	7717	1721	2249	1493	756	
7	573.2	Klebsiella pneumoniae MGH78578	9028	1524	2248	3003	1475	
8	216594.1	Mycobacterium marinum M	5769	1576	2241	1476	689	
9	224308.1	Bacillus subtilis subsp. subtilis str. 168	4112	1935	2150	1114	649	
10	272560.3	Burkholderia pseudomallei K96243	5728	1611	1950	1603	711	
11	269801.1	Bacillus cereus G9241	6147	1663	1939	987	603	
12	198094.1	Bacillus anthracis str. Ames	5664	1865	1934	1085	675	
13	226900.1	Bacillus cereus ATCC 14579	5557	1805	1929	1097	663	
14	222523.1	Bacillus cereus ATCC 10987	5921	1836	1926	1105	661	
15	281309.1	Bacillus thuringiensis serovar konkukian str. 97	5121	1779	1924	958	590	
16	288681.3	Bacillus cereus ZK	5137	1809	1904	1427	823	
17	260799.1	Bacillus anthracis str. Sterne	5287	1813	1903	960	611	
18	261594.1	Bacillus anthracis str. 'Ames Ancestor'	5618	1765	1872	1001	616	
19	216595.1	Pseudomonas fluorescens SBW25	6240	1460	1870	1636	750	
20	216596.1	Rhizobium leguminosarum bv. viciae 3841	7401	1354	1833	1565	820	
21	247156.1	Nocardia farcinica IFM 10152	5609	1287	1819	1206	662	
22	227882.1	Streptomyces avermitilis MA-4680	7792	1365	1811	1113	633	
23	272558.1	Bacillus halodurans C-125	4099	1384	1794	1085	673	
24	100226.1	Streptomyces coelicolor A3(2)	8154	1361	1785	1187	676	
25	205922.3	Pseudomonas fluorescens PFO-1	5736	1334	1769	1347	692	
26	257310.1	Bordetella bronchiseptica RB50	5023	1036	1742	1067	593	
27	199310.1	Escherichia coli CFT073	5382	1396	1716	1474	837	
28	216593.1	Escherichia coli E2348/69	5403	1370	1677	1507	846	
29	83333.1	Escherichia coli K12	4311	1375	1666	1505	792	
30	246196.1	Mycobacterium smegmatis str. MC2 155	6812	1374	1626	731	364	
31	223926.1	Vibrio parahaemolyticus RIMD 2210633	4864	1285	1618	1135	695	
32	269482.1	Burkholderia cepacia R1808	7229	1393	1618	1167	589	
33	257311.1	Bordetella parapertussis 12822	4451	1001	1610	1000	548	
34	155864.1	Escherichia coli O157:H7 EDL933	5324	1357	1604	1485	777	
35	216599.1	Shigella sonnei 53G	5851	1318	1602	1528	799	
36	216592.1	Escherichia coli 042	5715	1347	1600	1511	798	
37	262316.1	Mycobacterium avium subsp. paratuberculosis s	4505	1216	1599	731	437	
38	208964.1	Pseudomonas aeruginosa PAO1	5684	1216	1598	1286	679	
39	83334.1	Escherichia coli O157:H7	5343	1353	1594	1462	763	
40	594.1	Salmonella enterica subsp. enterica serovar Ga	5240	1341	1589	1587	864	
41	83332.1	Mycobacterium tuberculosis H37Rv	3928	1200	1579	857	448	
42	83331.1	Mycobacterium tuberculosis CDC1551	4355	1183	1578	777	446	
43	220664.3	Pseudomonas fluorescens Pf-5	6137	1299	1577	1563	802	
44	224911.1	Bradyrhizobium japonicum USDA 110	8593	1191	1564	1454	749	
45	233413.1	Mycobacterium bovis AF2122/97	4008	1202	1559	802	458	
46	196600.1	Vibrio vulnificus YJ016	4922	1198	1542	1123	683	
47	891.1	Desulfuromonas acetoxidans	6567	1400	1536	1847	864	
48	216895.1	Vibrio vulnificus CMCP6	4512	1194	1529	1100	680	
49	1806.1	Mycobacterium microti OV254	4232	1114	1519	1045	567	
50	320372.3	Burkholderia pseudomallei 1710b	6347	1293	1518	0	0	
51	160488.1	Pseudomonas putida KT2440	5533	1400	1487	1155	633	
52	99287.1	Salmonella typhimurium LT2	4528	1325	1482	1416	789	
53	216598.1	Shigella dysenteriae M131649	6284	1240	1480	1482	771	
54	54388.1	Salmonella paratyphi	7185	1055	1469	2525	1315	
55	235.1	Brucella abortus	3342	1357	1464	1197	573	

statistics_all.txt

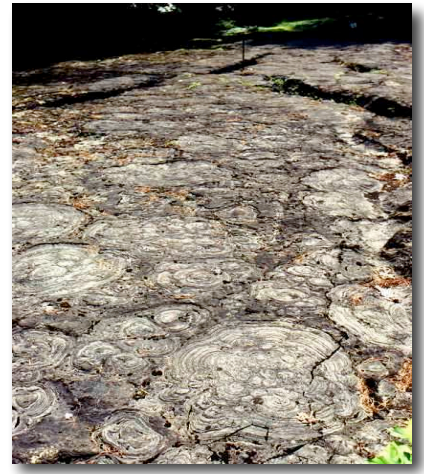
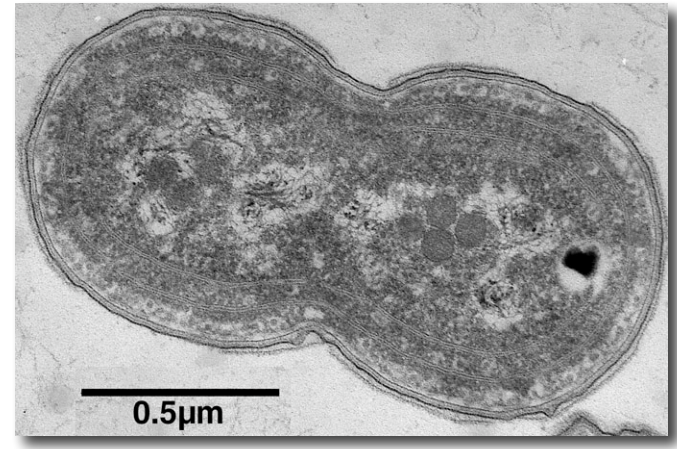
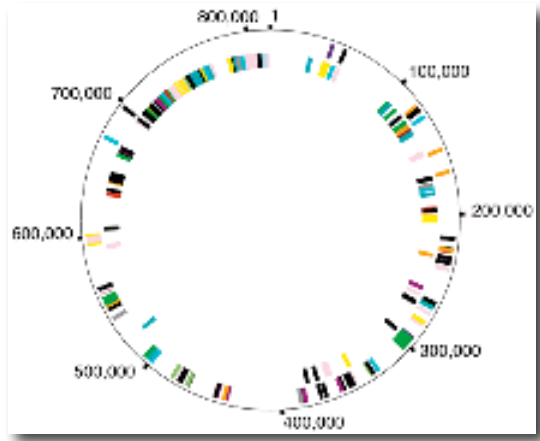
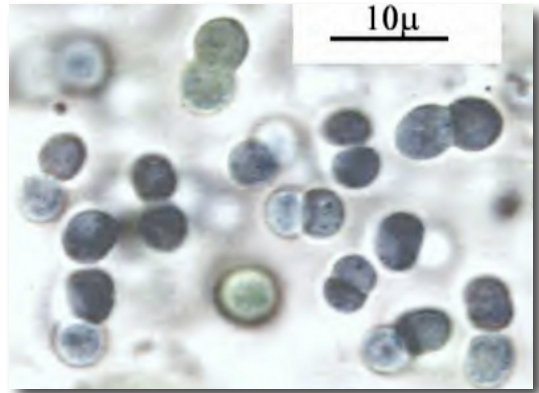
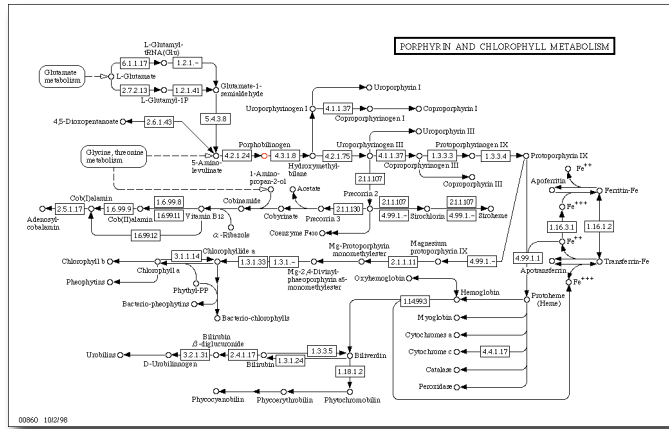
	A	B	C	D	E	F	G	H
1	GenomeID	Name	NumGeneSEED	NumGenesInFIGFam	RxnsFromBsub	NumECsSEED	NumRxnsSEED	
2	36873.1	Burkholderia xenovorans LB400	8738	2386	3411	2041	1053	
3	134537.1	Burkholderia fungorum	8445	2063	2984	1603	850	
4	216591.1	Burkholderia cenocepacia J2315	7820	1841	2345	1963	857	
5	264198.3	Ralstonia eutropha JMP134	6535	1535	2343	1601	830	
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21	247156.1	Nocardia farcinica IFM 10152	5609	1287	1819	1206	662	
22	227882.1	Streptomyces avermitilis MA-4680	7792	1365	1811	1113	633	
23	272558.1	Bacillus halodurans C-125	4099	1384	1794	1085	673	
24	100226.1	Streptomyces coelicolor A3(2)	8154	1361	1785	1187	676	
25	205922.3	Pseudomonas fluorescens PFO-1	5736	1334	1769	1347	692	
26	257810.1	Bordetella bronchiseptica RD58	5829	1896	1712	1867	598	
27	199310.1	Escherichia coli CFT073	5382	1396	1716	1474	837	
28	216593.1	Escherichia coli E2348/69	5403	1370	1677	1507	846	
29	83333.1	Escherichia coli K12	4311	1375	1666	1505	792	
30	216186.1	Mycobacterium smegmatis str. MC2 155	6812	1371	1626	731	361	
31	223926.1	Vibrio parahaemolyticus RIMD 2210633	4864	1285	1618	1135	695	
32	269482.1	Burkholderia cepacia R1808	7229	1393	1618	1167	589	
33	257311.1	Bordetella parapertussis 12822	4451	1001	1610	1000	548	
34	155864.1	Escherichia coli O157:H7 EDL933	5324	1357	1604	1485	777	
35	216599.1	Shigella sonnei 53G	5851	1318	1602	1528	799	
36	216592.1	Escherichia coli 042	5715	1347	1600	1511	798	
37	262316.1	Mycobacterium avium subsp. paratuberculosis s	4505	1216	1599	731	437	
38	208964.1	Pseudomonas aeruginosa PAO1	5684	1216	1598	1286	679	
39	83334.1	Escherichia coli O157:H7	5343	1353	1594	1462	763	
40	594.1	Salmonella enterica subsp. enterica serovar Ga	5240	1341	1589	1587	864	
41	83332.1	Mycobacterium tuberculosis H37Rv	3928	1200	1579	857	448	
42	83331.1	Mycobacterium tuberculosis CDC1551	4355	1183	1578	777	446	
43	220664.3	Pseudomonas fluorescens Pf-5	6137	1299	1577	1563	802	
44	224911.1	Bradyrhizobium japonicum USDA 110	8593	1191	1564	1454	749	
45	233413.1	Mycobacterium bovis AF2122/97	4008	1202	1559	802	458	
46	196600.1	Vibrio vulnificus YJ016	4922	1198	1542	1123	683	
47	891.1	Desulfuromonas acetoxidans	6567	1400	1536	1847	864	
48	216895.1	Vibrio vulnificus CMCP6	4512	1194	1529	1100	680	
49	1806.1	Mycobacterium microti OV254	4232	1114	1519	1045	567	
50	320372.3	Burkholderia pseudomallei 1710b	6347	1293	1518	0	0	
51	160488.1	Pseudomonas putida KT2440	5533	1400	1487	1155	633	
52	99287.1	Salmonella typhimurium LT2	4528	1325	1482	1416	789	
53	216598.1	Shigella dysenteriae M131649	6284	1240	1480	1482	771	
54	54388.1	Salmonella paratyphi	7185	1055	1469	2525	1315	
55	235.1	Brucella abortus	3342	1357	1464	1197	573	

statistics_all.txt

Semi-Automated Generation of FBA Models

automated model generation is advancing dramatically
 approaching manual construction in some cases

Genes → Proteins → Cell Networks → Cells → Populations → Communities → Ecosystems



Community Structure and Metabolism

Gene W. Tyson¹, Jarrod Chapman^{3,4}, Philip Hugenholtz¹, Eric E. Allen¹, Rachna J. Ram¹, Paul M. Richardson⁴, Victor V. Solovyev⁴, Edward M. Rubin⁴, Daniel S. Rokhsar^{3,4} & Jillian F. Banfield^{1,2}

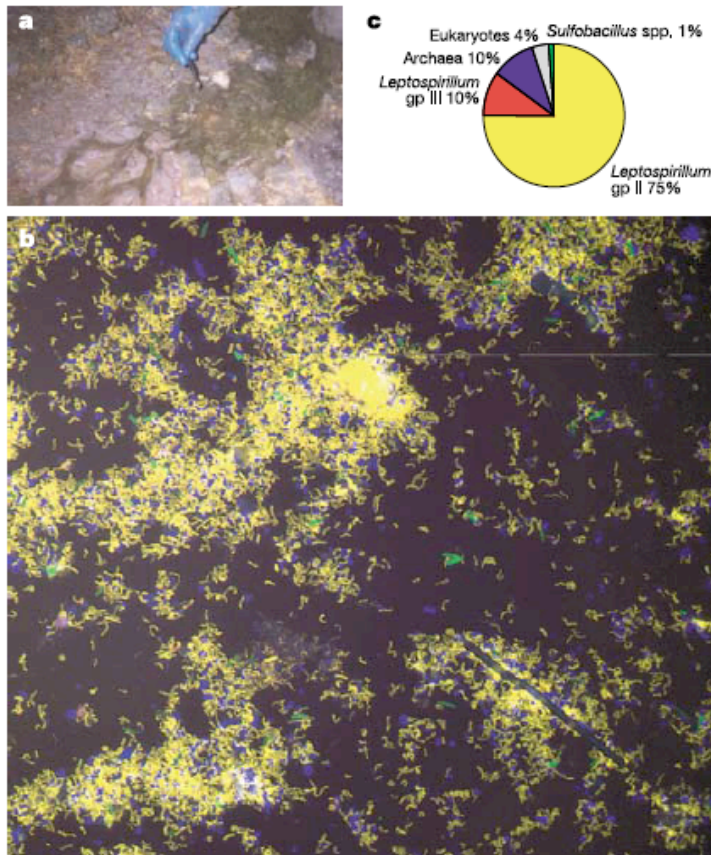


Figure 1 The pink biofilm. **a**, Photograph of the biofilm in the Richmond mine (hand included for scale). **b**, FISH image of **a**. Probes targeting bacteria (EUBmix; fluorescein isothiocyanate (green)) and archaea (ARC915; Cy5 (blue)) were used in combination with a probe targeting the *Leptospirillum* genus (LF655; Cy3 (red)). Overlap of red and green (yellow) indicates *Leptospirillum* cells and shows the dominance of *Leptospirillum*. **c**, Relative microbial abundances determined using quantitative FISH counts.

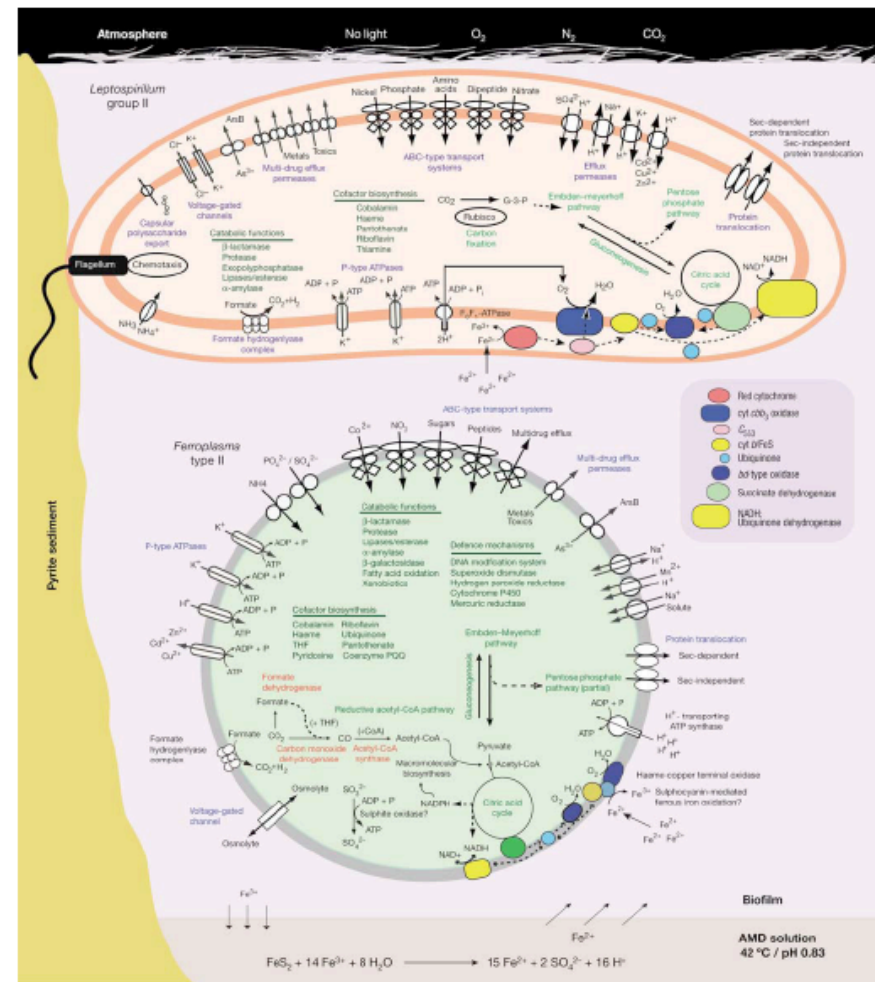
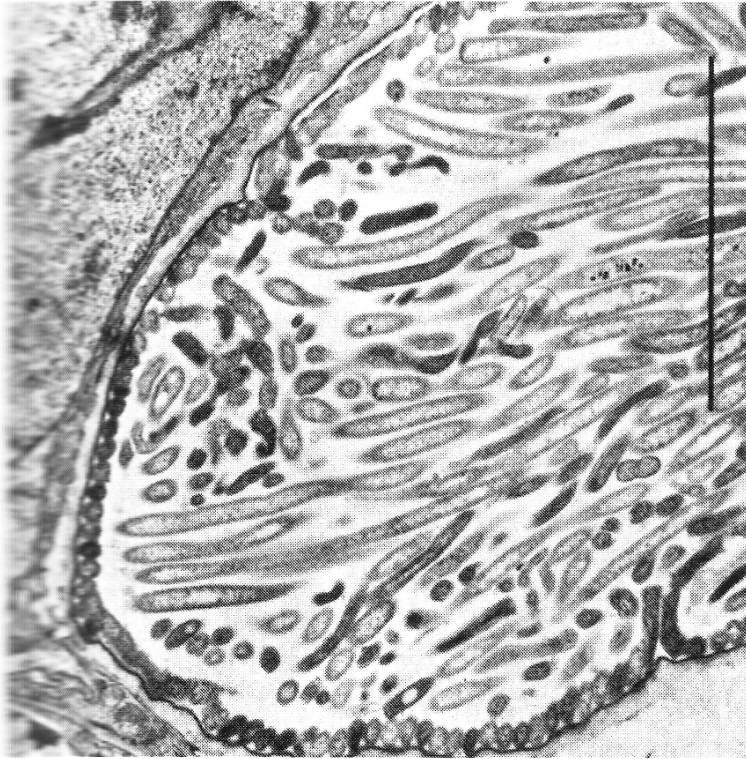


Figure 4 Cell metabolic cartoons constructed from the annotation of 2,180 ORFs identified in the *Leptospirillum* group II genome (53% with putative assigned function) and 1,531 ORFs in the *Ferroplesium* type II genome (58% with assigned function). The cell cartoons are shown within a biofilm that is 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.



A Diverse Bacterial Community



- Pocket in the hindgut wall of the Sonoran desert termite *Pterotermes occidentis*
- 10 billion bacteria per milliliter
- Anoxic environment
- ~30 strains are facultative aerobes
- Many/most are unknown

From Five Kingdoms

Lynn Margulis and Karlene Schwartz



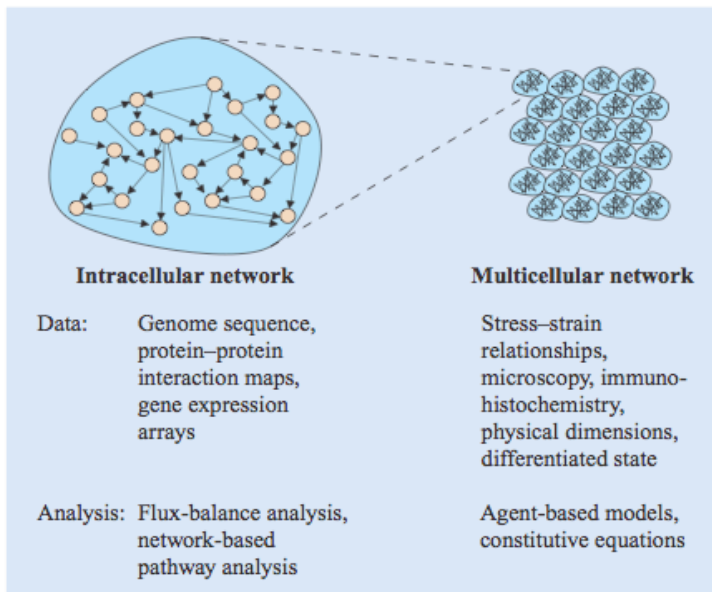


Figure 1

Coupling intracellular networks with tissue-level physiology. The phenotype of a given cell is a function of many proteins, metabolites, and their associated interactions. Tissue-level physiology arises from a collection of the phenotypes of the individual cells. Each scale of biological investigation has at its disposal a unique set of experimental techniques and analysis methods. The challenge remains to integrate the molecular network detail with the multicellular network that gives rise to human pathologies (e.g., cancer and cardiovascular disease).

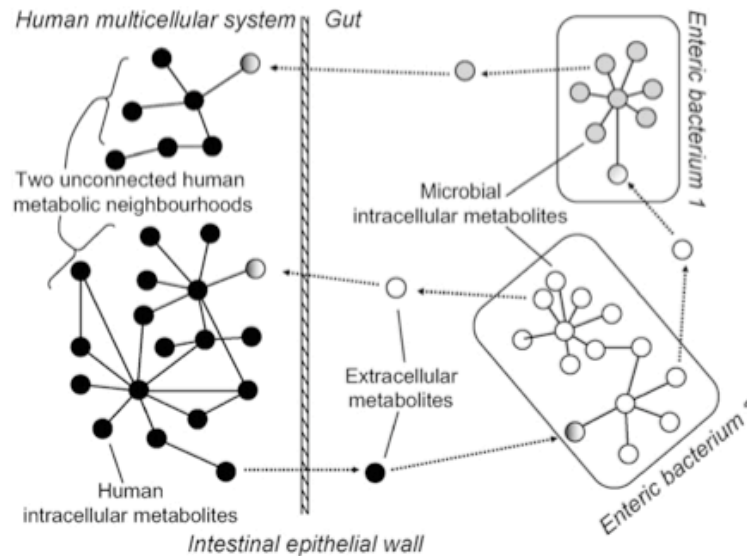


Figure 2 A complex network of metabolites derived from the human genome (*black circles*) enters the gut, where it is processed by *enteric bacteria* into metabolites (*white circles*). Some of these products are secreted into the human system, while others are secreted into the gut. Note that the metabolites are not connected to the human system.

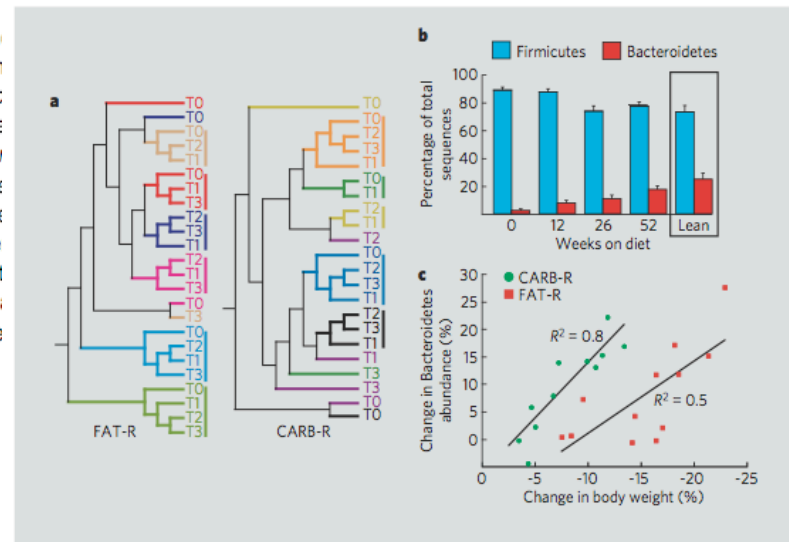


Figure 1 | Correlation between body-weight loss and gut microbial ecology. **a**, Clustering of 16S ribosomal RNA gene sequence libraries of faecal microbiota for each person (in different colours) and time point in diet therapy (T0, baseline; T1, 12 weeks; T2, 26 weeks; T3, 52 weeks) in the two diet-treatment groups (fat restricted, FAT-R; carbohydrate restricted, CARB-R), based on UniFrac analysis of the 18,348-sequence phylogenetic tree. **b**, Relative abundance of Bacteroidetes and Firmicutes. For each time point, values from all available samples were averaged (n was 11 or 12 per time point). Lean-subject controls include four stool samples from two people taken 1 year apart, plus three other stool samples⁶. Mean values \pm s.e. are plotted. **c**, Change in relative abundance of Bacteroidetes in subjects with weight loss above a threshold of 2% weight loss for the CARB-R diet and 6% for the FAT-R diet.

Modeling the Human Gut Microbial Environment could Lead to new therapy for obesity

Simulation Systems Under Development

TABLE 5.1 Sample Simulation Programs

Name	Descriptors ^a	Web Site
Gepasi/Copasi	fkFW	http://gepasi.dbs.aber.ac.uk/softw/gepasi.html
BioSim	qWMU	http://www.molgen.mpg.de/~biosim/BioSim/BioSimHome.html
Jarnac	krfbFWS	http://members.tripod.co.uk/sauro/Jarnac.htm
MCELL	rsU	http://www.mcell.cnl.salk.edu/
Virtual Cell	ksDFWMU	http://www.nrcam.uchc.edu/
E-Cell	kWUS	http://www.e-cell.org/
Neuron	ksFWMUS	http://neuron.duke.edu/
Genesis	ksUS	http://www.bbb.caltech.edu/GENESIS/genesis.html
Plas	kfbFW	http://correio.cc.fc.ul.pt/~aenf/plas.html
Ingeneue	qkFMWUS	http://www.ingeneue.org/
DynaFit	kfW	http://www.biokin.com/dynafit/
Stochsim	rS	http://www.zoo.cam.ac.uk/comp-cell/StochSim.html
T7 Simulator	kUS	http://virus.molsci.org/t7/
Molecularizer/Stochastirator	krUS	http://opnsrctbio.molsci.org/alpha/comps/sim.html

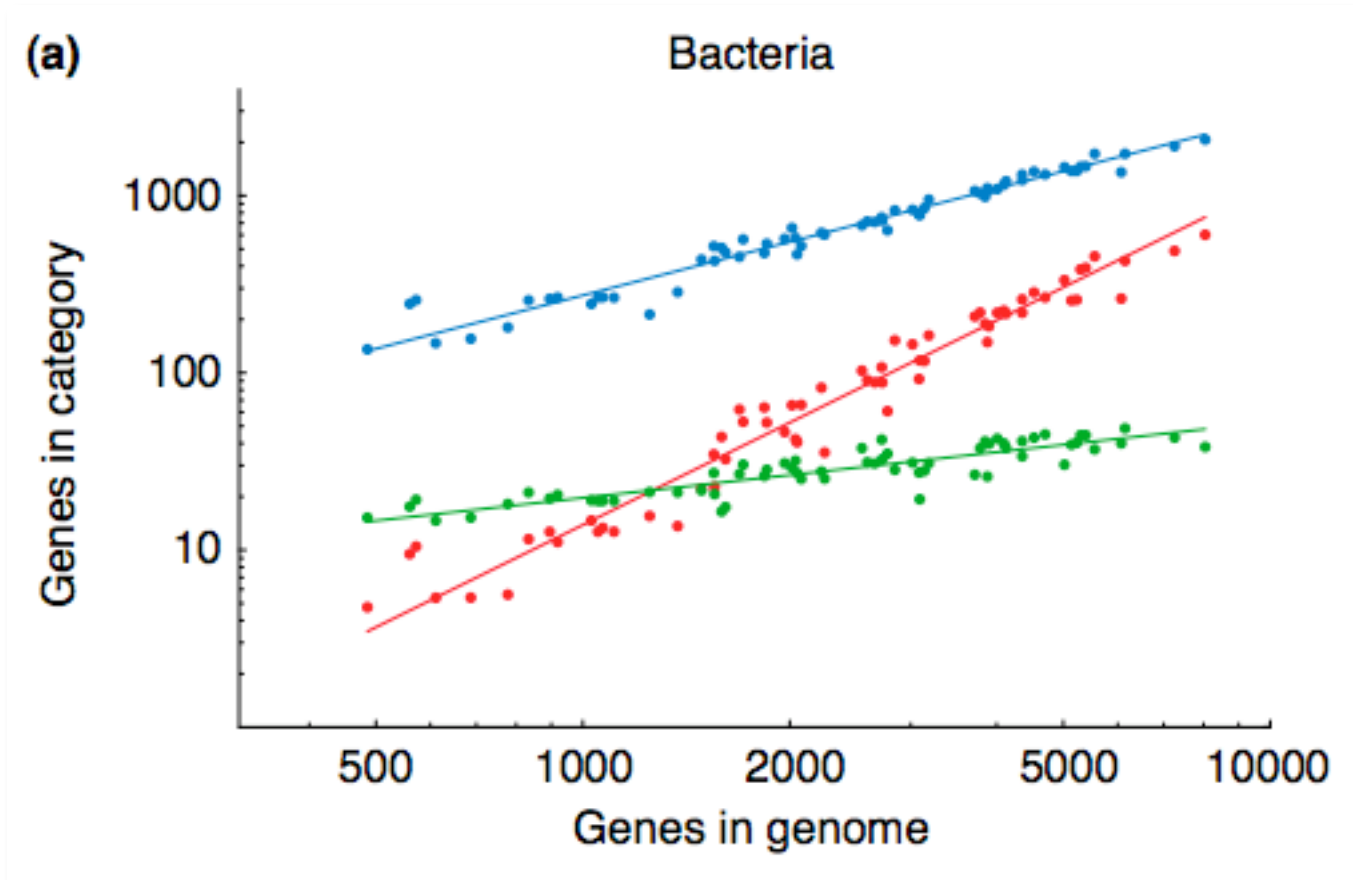
NOTE: All packages have facilities for chemical kinetic simulation of one sort or another. Some are better designed for metabolic systems, others for electrochemical systems, and still others for genetic systems.

^aThe descriptors are as follows: b, bifurcation analyses and steady-state calculation; f, flux balance or metabolic control and related analyses; k, deterministic kinetic simulation; q, qualitative simulation; r, stochastic process models; s, spatial processes; D, database connectivity; F, fitting, sensitivity, and optimization code; M, runs on Macintosh; S, source code available; U, runs on Linux or Unix; W, runs on windows.



Genome Size v. Protein Family (Function)

Metabolism
Transcription Regulation
Cell Cycle

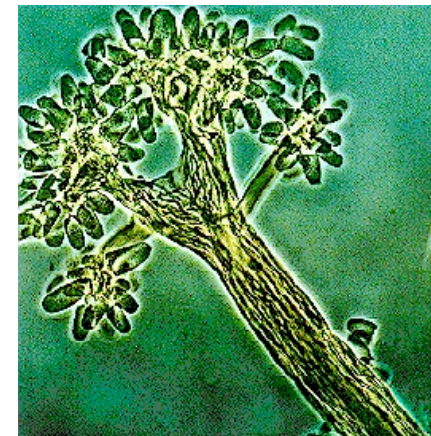
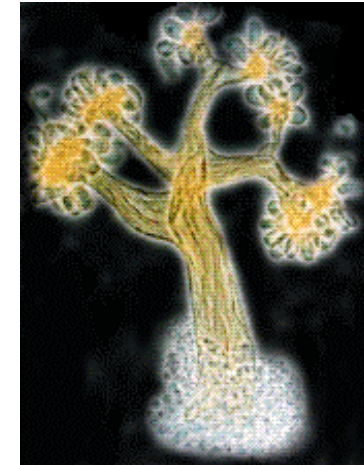
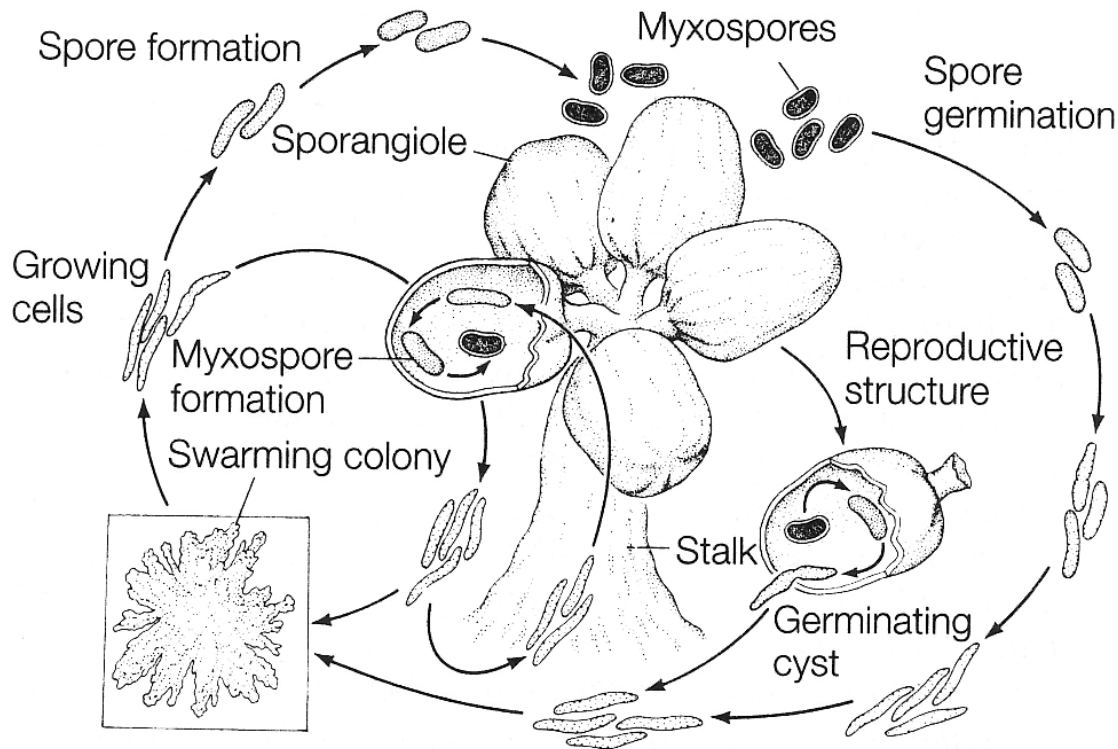


Erik van Nimwegen

Center for Studies in Physics and Biology, the Rockefeller University, 1230 York Avenue, New York, NY 12001, USA



Understanding Bacterial Life Cycles

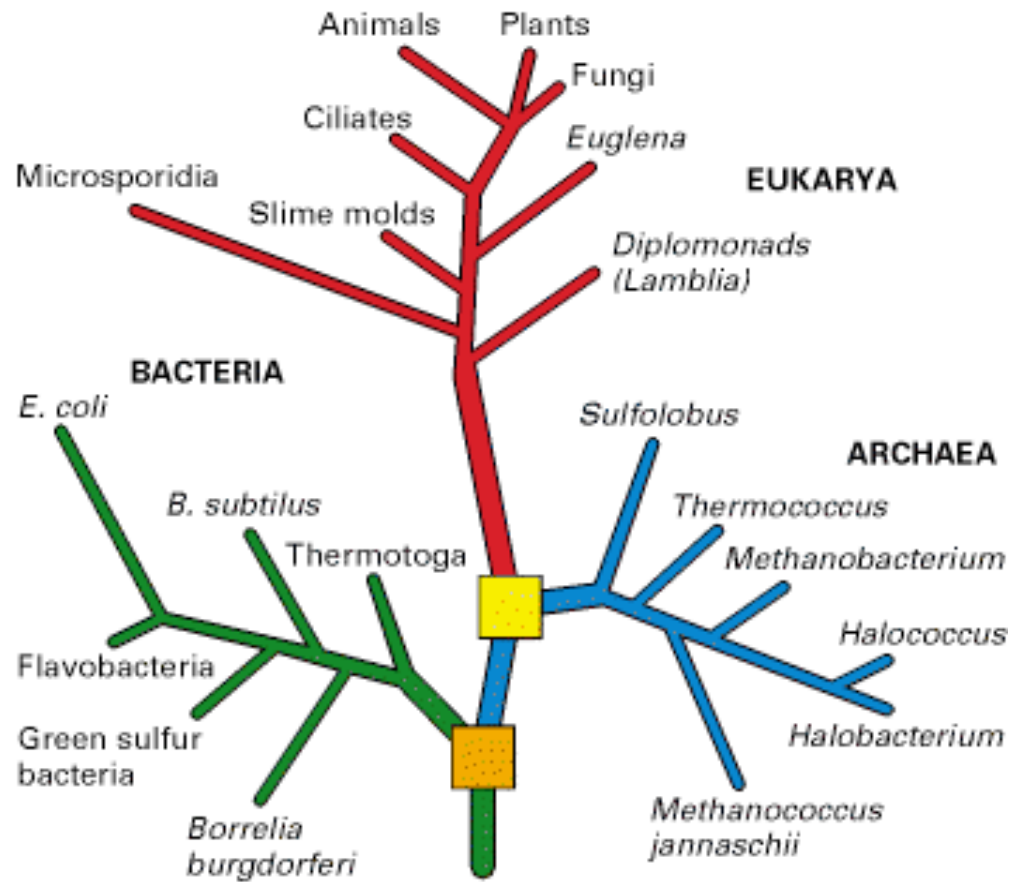



F Life cycle of *Stigmatella aurantiaca*. [Drawing by L. Meszoly; labeled by M. Dworkin.]

From Lynn Margulis and Karlene Schwartz



Looking for LUCA



 Presumed common progenitor of all extant organisms


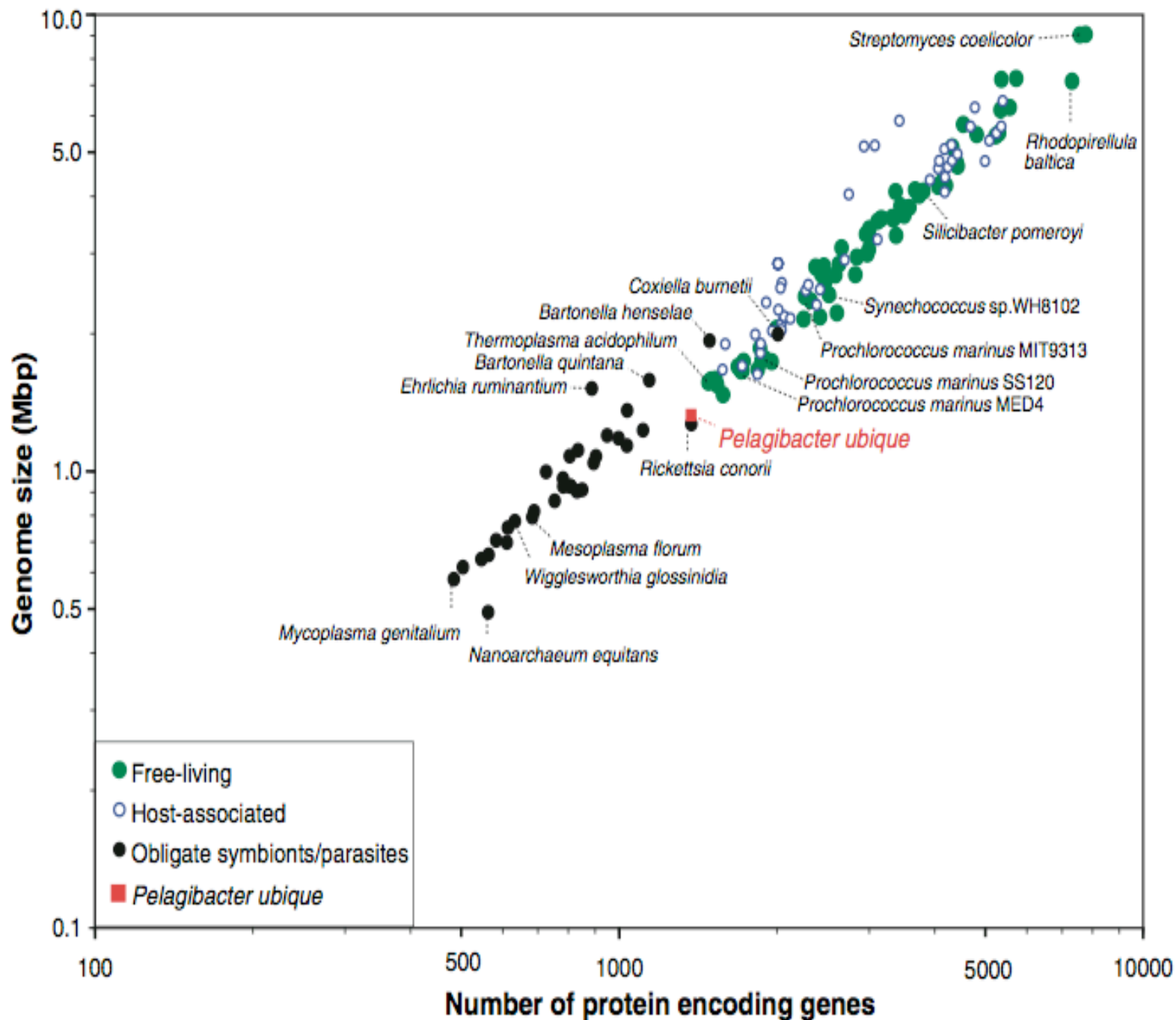
 Presumed common progenitor of archaeobacteria and eukaryotes

Fig. 1. Number of predicted protein-encoding genes versus genome size for 244 complete published genomes from bacteria and archaea. *P. ubiquus* has the smallest number of genes (1354 open reading frames) for any free-living organism.



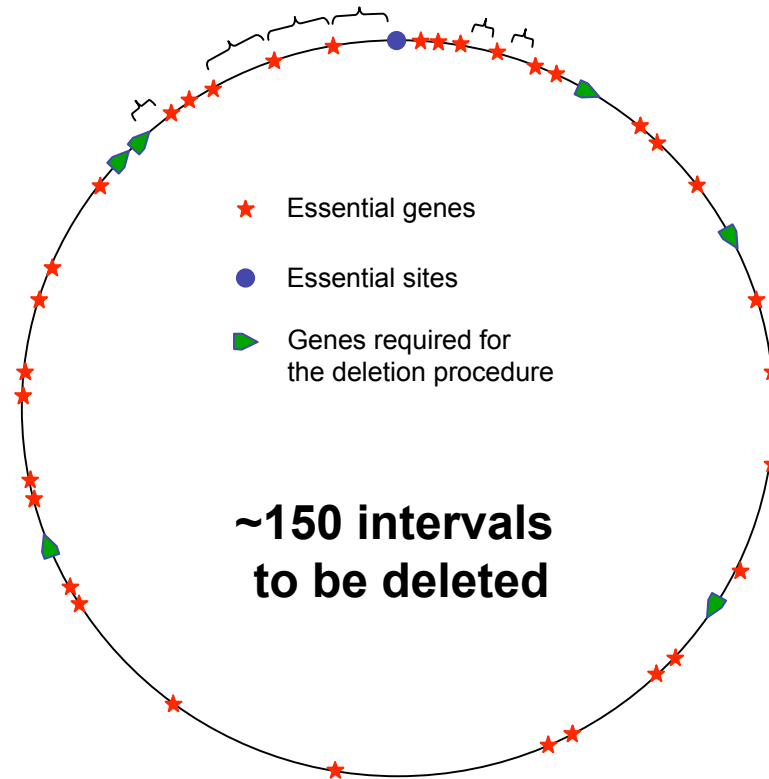
Argonne/INRA Minimum Organism Project

- Systematically reduce the *Bacillus subtilis* genome via interval deletion
 - Interval deletion methods
 - Reduced strain consolidation
 - Conduct phenotype analysis of reduced organisms
- Improve the methods for chromosomal reduction (wet lab)
- Develop whole genome flux balance model for *Bacillus subtilis*
 - Predict essentiality, growth conditions and metabolic phenotypes
 - Produce a model for each reduced strain (mass production)
- Reconcile essential genes/intervals from experiment and in silico predictions for each mutant
- Extend modeling and simulation beyond core metabolic functions to incorporate information processing, DNA and RNA processing, cell walls and replication processes (“logistical” models)
- Produce within the three year demonstration period significant progress towards an integrated model and experimentally driven system
- Develop potential approaches for engineering based on mini-bsub



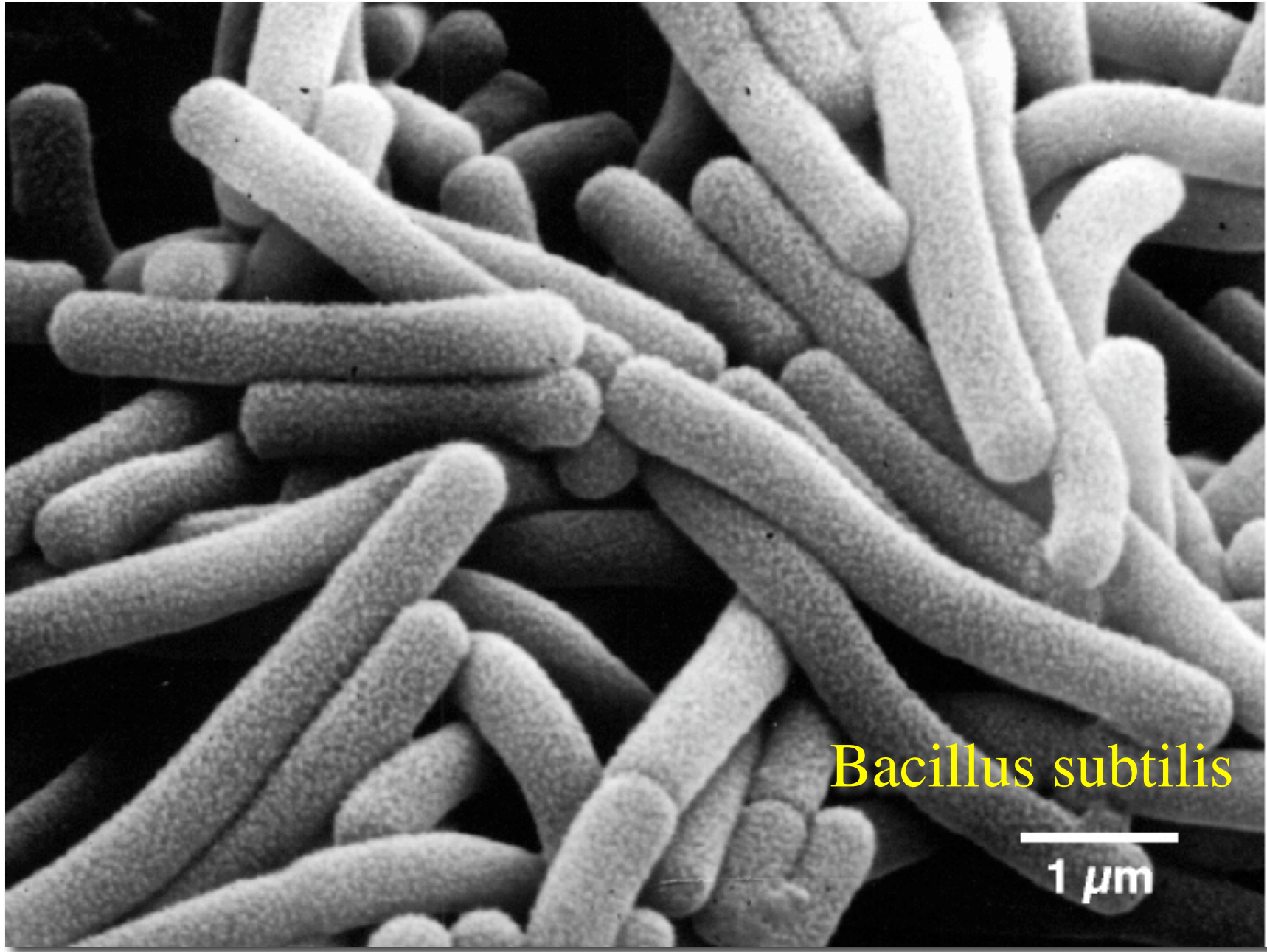
Minimizing the *Bacillus subtilis* chromosome

29/02/2006



Strategy

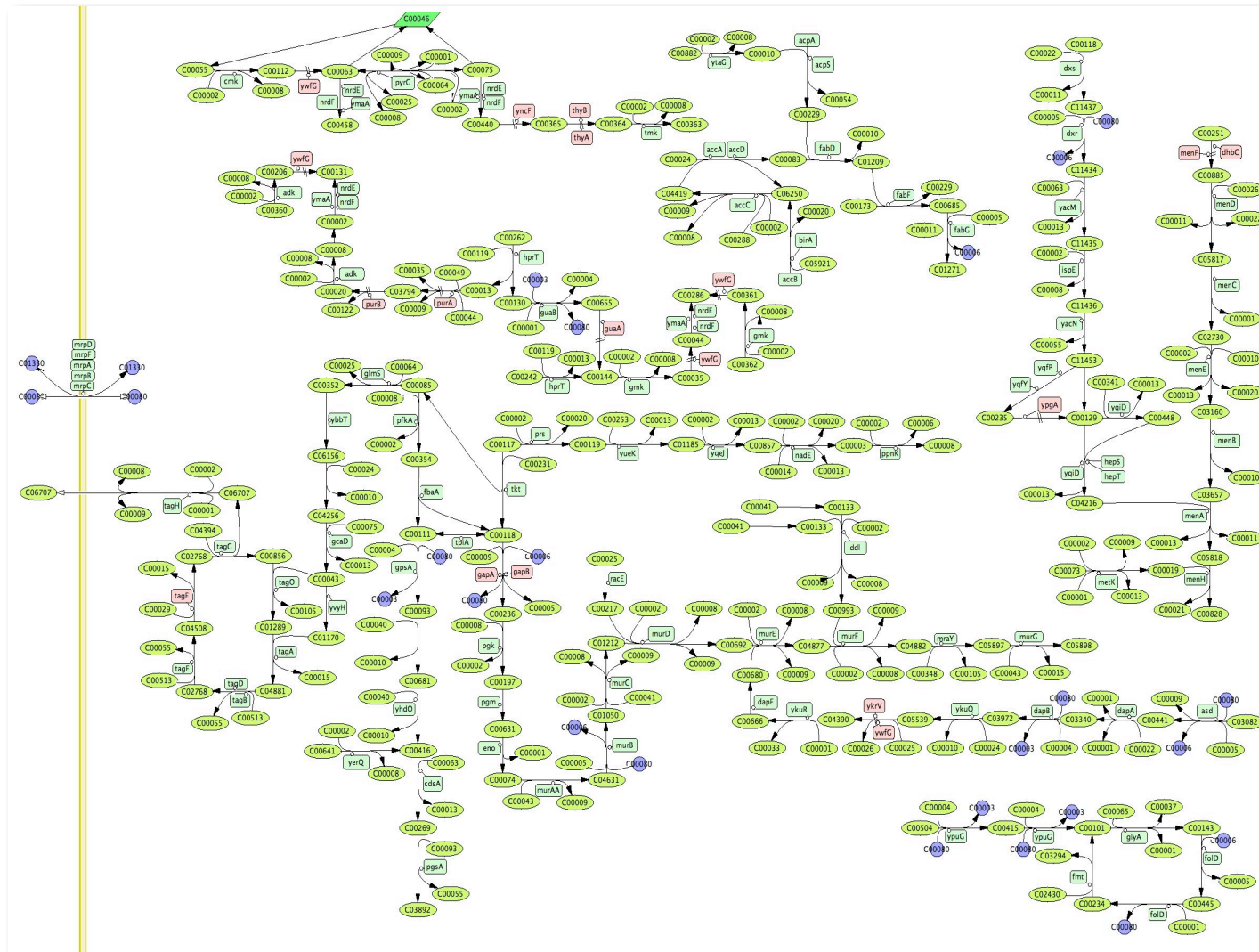
- 1 - Generate single deletion of each interval (marked)
- 2 - Identify essential and dispensable intervals
- 3 - Combine deletions in dispensable intervals



Bacillus subtilis

1 μm

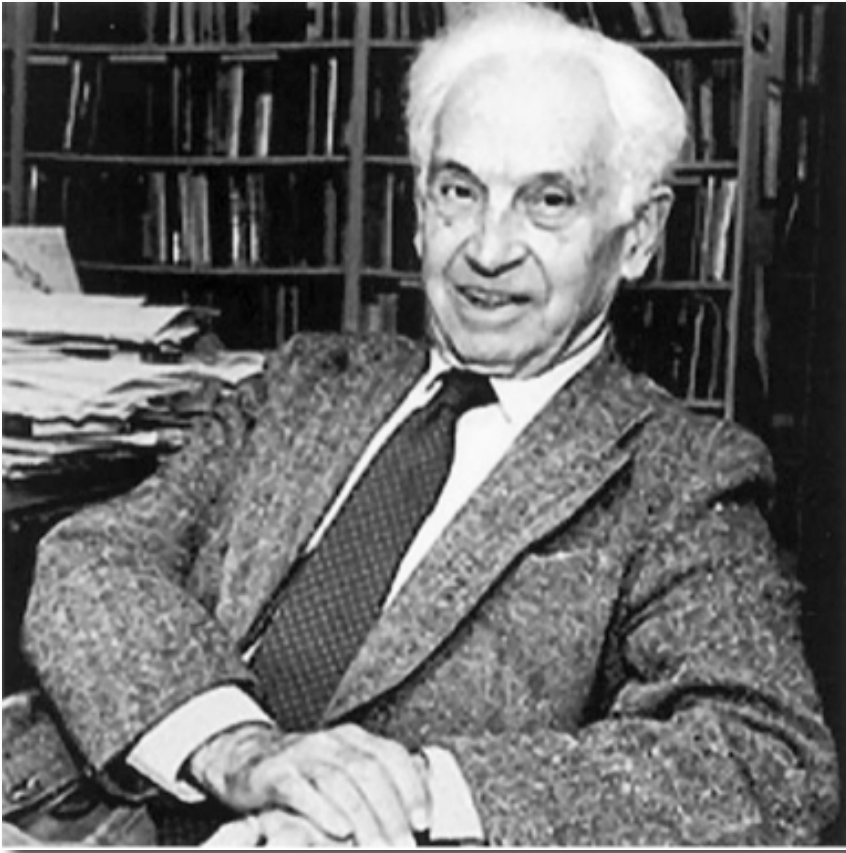
Model of *Bacillus subtilis* Essential Core



Connecting to the Computing

- Goal: Reconstructing from genomes a predictive model of an organisms function (phenotype).
- Software: $O(100)$ tools currently in the analysis toolchain and $O(10)$ in the modeling toolchain
- Curation of data and databases major factors in progress
- Tools are loosely coupled but use integrated databases
- Perl, Python, C, C++, Libraries, Matlab, R, etc.
- Human productivity is the goal (computing > thinking time)
- Systems: workstations, small clusters, large-clusters (database providers), grids, HPC
- Data management is a major problem due to complexity and update cycle not volume
- Future Architectures: Interactive analysis servers, special purpose, general purpose HPC
- **The Big Opportunity: Near real-time genome analysis to enable personalized medicine**

Biological Science is Fundamentally Different from the Physical Sciences

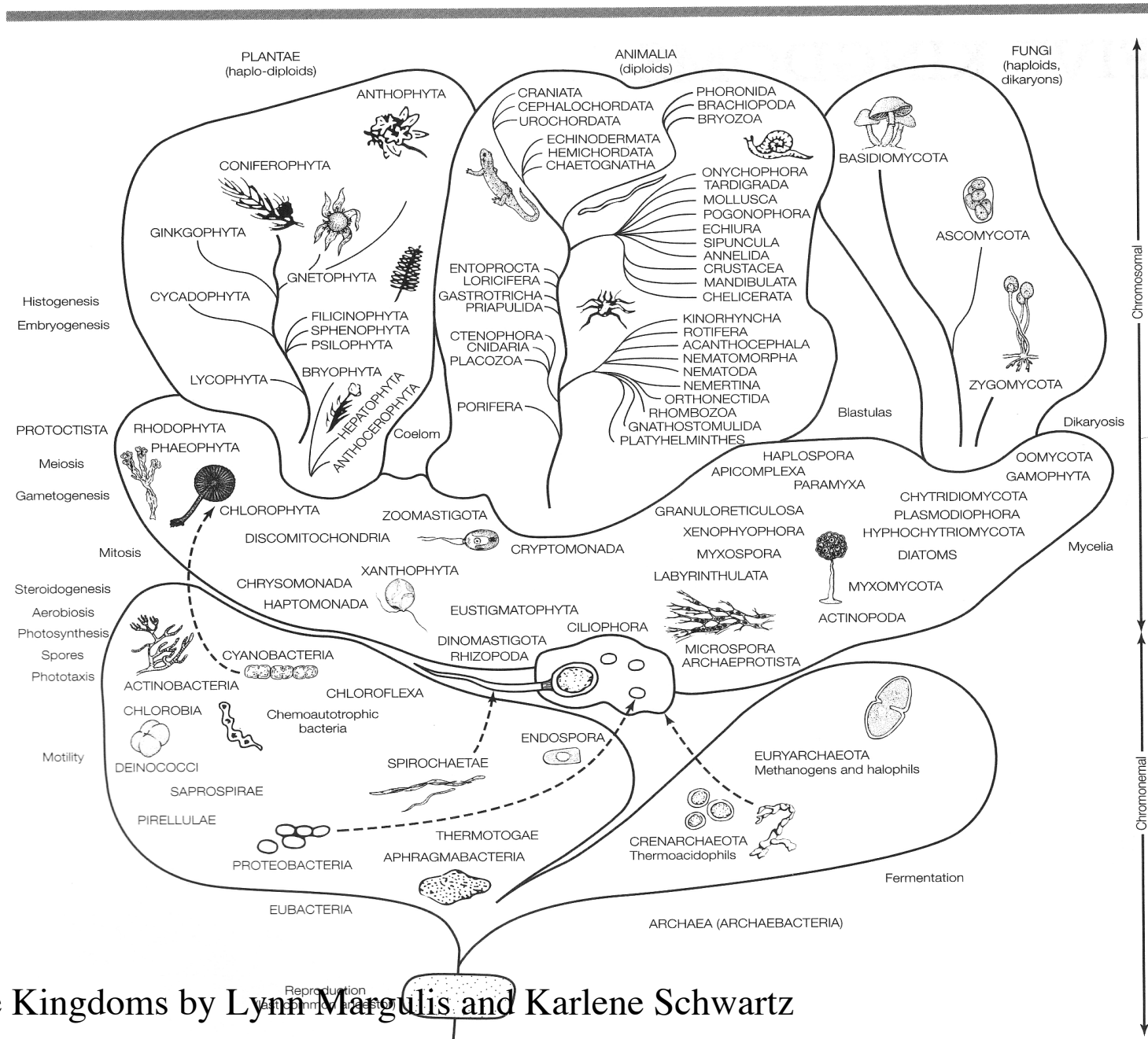


Ernst Mayr (1904-2005)

- Biology is more than physics
 - Differentiate inanimate and living processes
- Dual Causality (Mayr)
 - Physiochemical laws
 - *Data light, reversible, time invariant*
 - Genetic programs
 - *Data rich, irreversible, time variant*
- Ecological, economic and social laws
 - Higher-order principles that govern the behavior of collections of autonomous entities

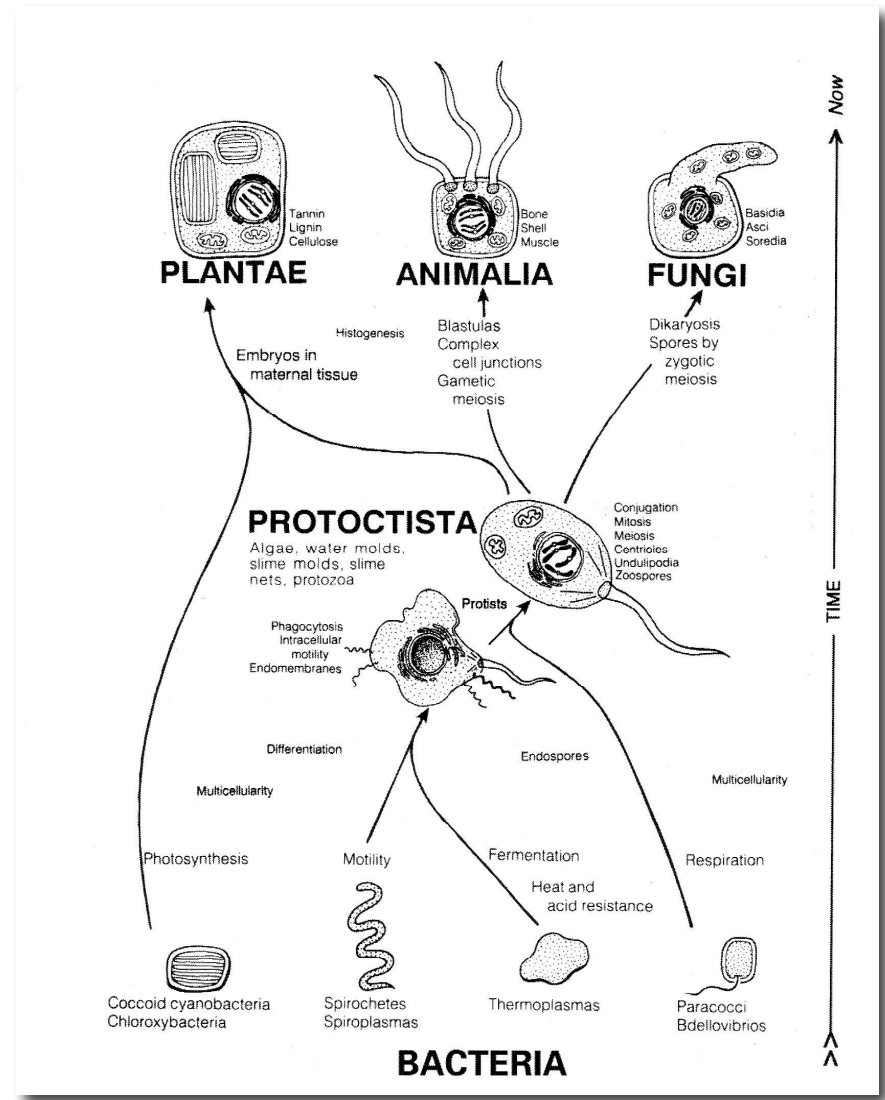
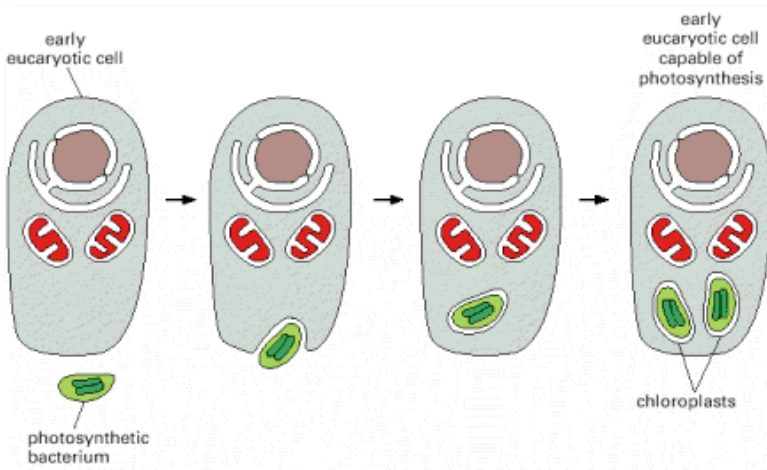
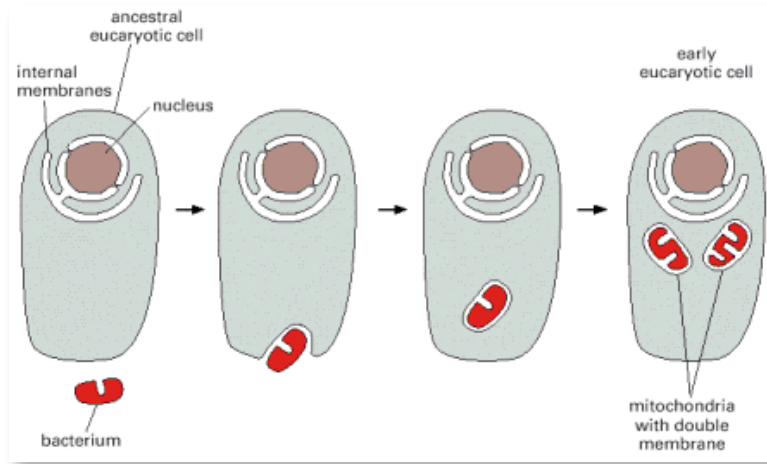


The phyla of life on Earth based on our modification of the Whittaker five-kingdom system and the symbiotic theory of the origin of eukaryotic cells.



From Five Kingdoms by Lynn Margulis and Karlene Schwartz

Understanding the Evolution of Cellular Functions and the Role of Symbiogenesis



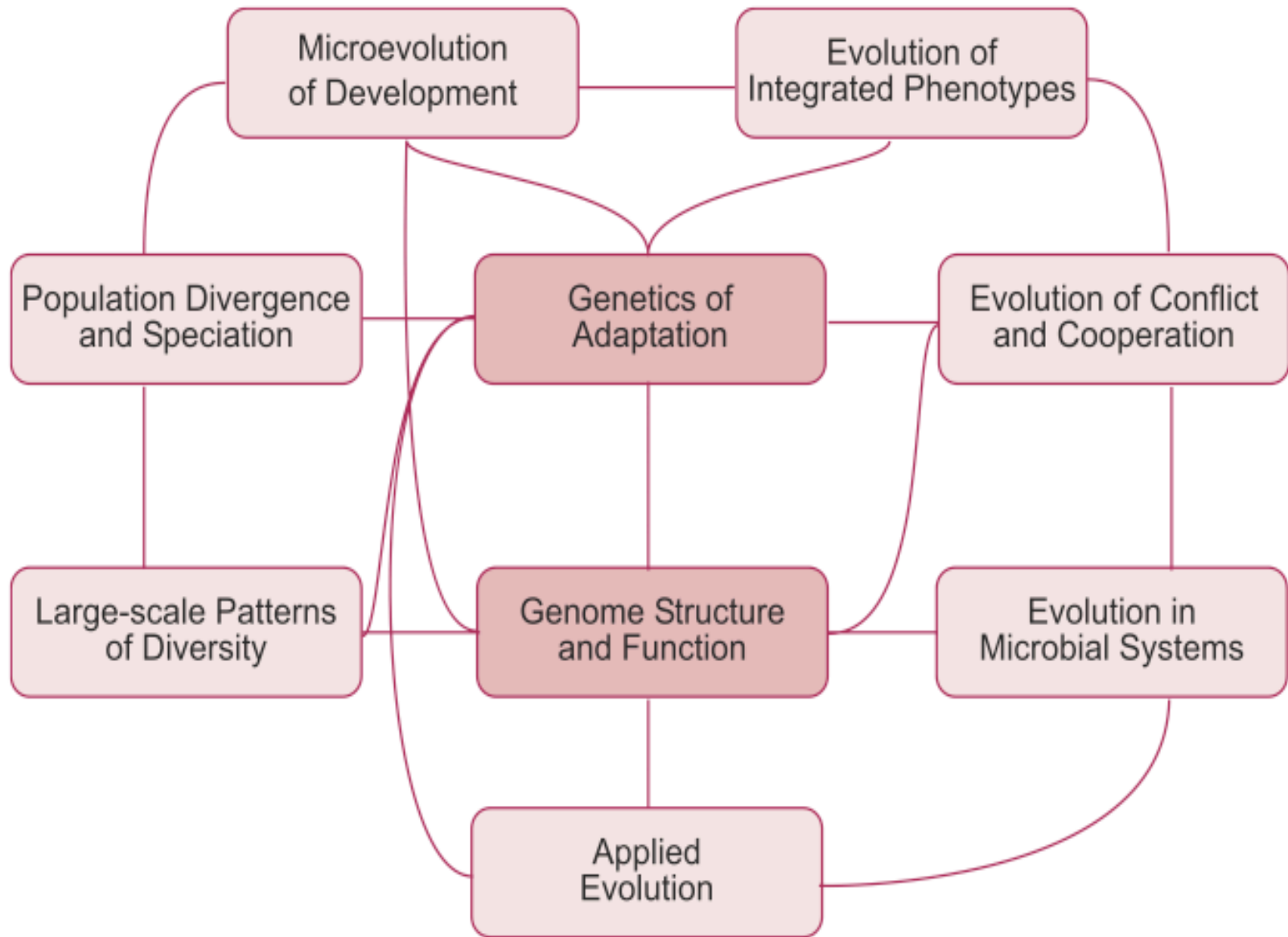


Figure 1: Connections among research frontiers identified in this report.

Evolutionary Timeline

Time (Myr ago)	Event
4600	Formation of the approximately homogeneous solid Earth by planetesimal accretion
4300	Melting of the Earth due to radioactive and gravitational heating which leads to its differentiated interior structure as well as outgassing of molecules such as water, methane, ammonia, hydrogen, nitrogen, and carbon dioxide
4300	Atmospheric water is photodissociated by ultraviolet light to give oxygen atoms which are incorporated into an ozone layer and hydrogen molecules which escape into space
4000	Bombardment of the Earth by planetesimals stops
3800	The Earth's crust solidifies--formation of the oldest rocks found on Earth
3800	Condensation of atmospheric water into oceans
3500-2800	Prokaryotic cell organisms develop
3500-2800	Beginning of photosynthesis by blue-green algae which releases oxygen molecules into the atmosphere and steadily works to strengthen the ozone layer and change the Earth's chemically reducing atmosphere into a chemically oxidizing one
2400	Rise in the concentration of oxygen molecules stops the deposition of uraninites (since they are soluble when combined with oxygen) and starts the deposition of banded iron formations
2000	The Oklo natural fission reactor in Gabon goes into operation
1600	The last reserves of reduced iron are used up by the increasing atmospheric oxygen--last banded iron formations
1500	Eukaryotic cell organisms develop
1500-600	Rise of multicellular organisms
580-545	Fossils of Ediacaran organisms are made
545	Cambrian explosion of hard-bodied organisms
528-526	Fossilization of the Chengjiang site
517-515	Fossilization of the Burgess Shale
500-450	Rise of the fish--first vertebrates



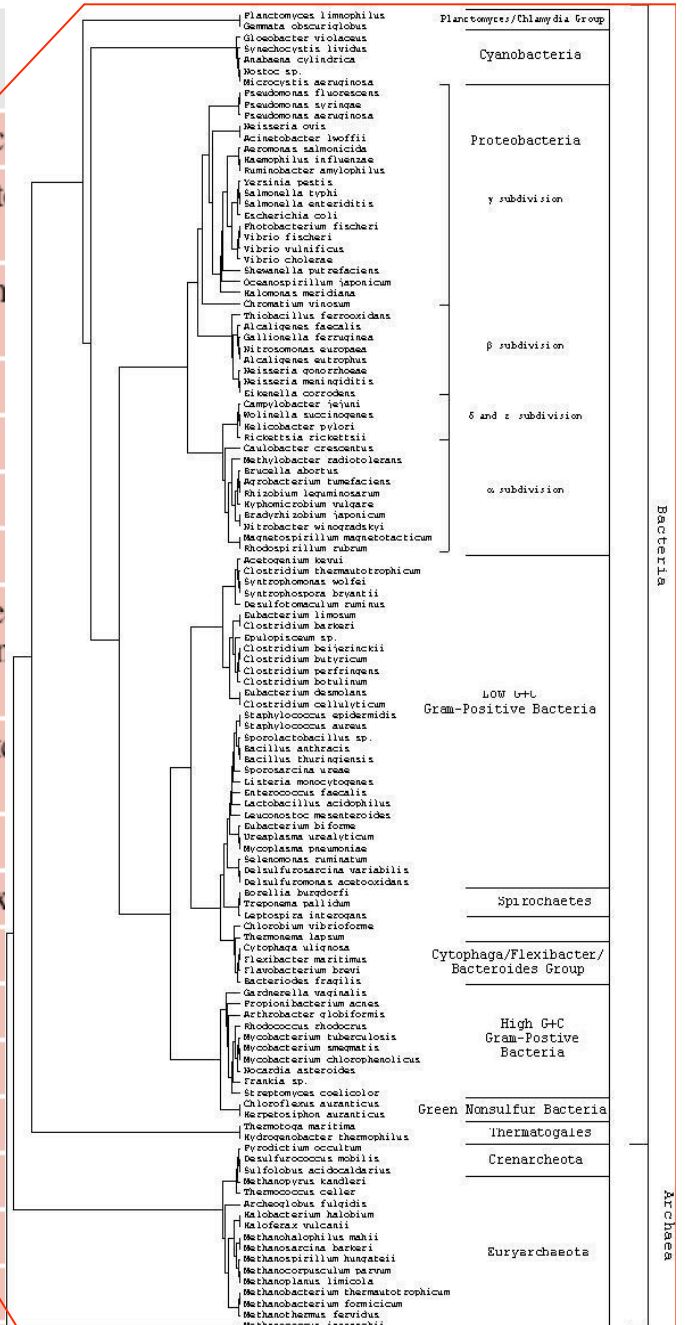
Evolutionary Timeline

Time (Myr ago)	Event
4600	Formation of the approximately homogeneous solid Earth by planetesimal accretion
4300	Melting of the Earth due to radioactive and gravitational heating which leads to its differentiated interior structure as well as outgassing of molecules such as water, methane, ammonia, hydrogen, nitrogen, and carbon dioxide
4300	Atmospheric water is photodissociated by ultraviolet light to give oxygen atoms layer and hydrogen molecules which escape into space
4000	Bombardment of the Earth by planetesimals stops
3800	The Earth's crust solidifies--formation of the oldest rocks found on Earth
3800	Condensation of atmospheric water into oceans
3500-2800	Prokaryotic cell organisms develop
3500-2800	Beginning of photosynthesis by blue-green algae which releases oxygen molecule works to strengthen the ozone layer and change the Earth's chemically reducing oxidizing one
2400	Rise in the concentration of oxygen molecules stops the deposition of uraninites (since they are soluble when combined with oxygen) and starts the deposition of banded iron formations
2000	The Oklo natural fission reactor in Gabon goes into operation
1600	The last reserves of reduced iron are used up by the increasing atmospheric oxygen--last banded iron formations
1500	Eukaryotic cell organisms develop
1500-600	Rise of multicellular organisms
580-545	Fossils of Ediacaran organisms are made
545	Cambrian explosion of hard-bodied organisms
528-526	Fossilization of the Chengjiang site
517-515	Fossilization of the Burgess Shale
500-450	Rise of the fish--first vertebrates



Evolutionary Timeline

Time (Myr ago)	Event
4600	Formation of the approximately homogeneous solid Earth by planetesimal accretion
4300	Melting of the Earth due to radioactive and gravitational heating which leads to magma oceans as well as outgassing of molecules such as water, methane, ammonia, hydrogen, and carbon dioxide
4300	Atmospheric water is photodissociated by ultraviolet light to give oxygen atoms and hydrogen molecules which escape into space
4000	Bombardment of the Earth by planetesimals stops
3800	The Earth's crust solidifies--formation of the oldest rocks found on Earth
3800	Condensation of atmospheric water into oceans
3500-2800	Prokaryotic cell organisms develop
3500-2800	Beginning of photosynthesis by blue-green algae which releases oxygen molecules (which works to strengthen the ozone layer and change the Earth's chemically reducing atmosphere to an oxidizing one)
2400	Rise in the concentration of oxygen molecules stops the deposition of uraninite (combined with oxygen) and starts the deposition of banded iron formations
2000	The Oklo natural fission reactor in Gabon goes into operation
1600	The last reserves of reduced iron are used up by the increasing atmospheric oxygen
1500	Eukaryotic cell organisms develop
1500-600	Rise of multicellular organisms
580-545	Fossils of Ediacaran organisms are made
545	Cambrian explosion of hard-bodied organisms
528-526	Fossilization of the Chengjiang site
517-515	Fossilization of the Burgess Shale
500-450	Rise of the fish--first vertebrates



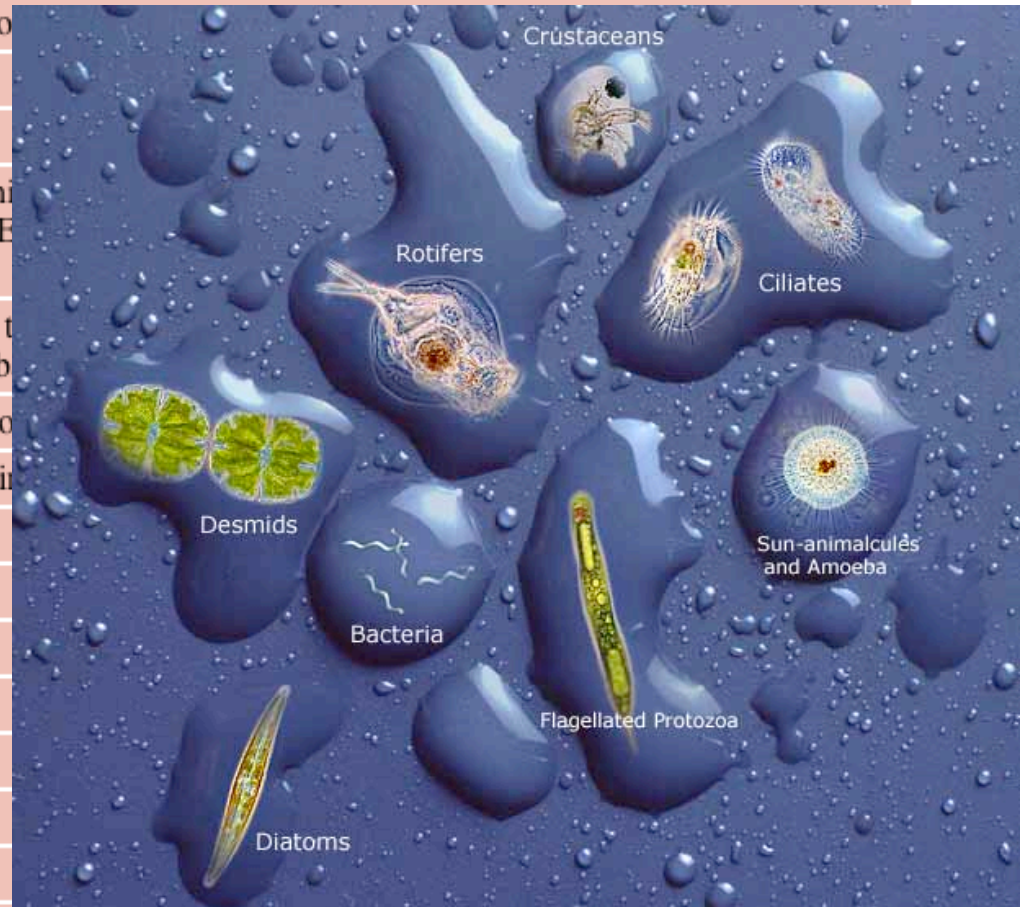
Evolutionary Timeline

Time (Myr ago)	Event
4600	Formation of the approximately homogeneous solid Earth by planetesimal accretion
4300	Melting of the Earth due to radioactive and gravitational heating which leads to its differentiated interior structure as well as outgassing of molecules such as water, methane, ammonia, hydrogen, nitrogen, and carbon dioxide
4300	Atmospheric water is photodissociated by ultraviolet light to give oxygen atoms which are incorporated into an ozone layer and hydrogen molecules which escape into space
4000	Bombardment of the Earth by planetesimals stops
3800	The Earth's crust solidifies--formation of the oldest rocks found on Earth
3800	Condensation of atmospheric water into oceans
3500-2800	Prokaryotic cell organisms develop
3500-2800	Beginning of photosynthesis by blue-green algae which releases oxygen molecules into the atmosphere and steadily works to strengthen the ozone layer and change the Earth's chemically reducing atmosphere into a chemically oxidizing one
2400	Rise in the concentration of oxygen molecules stops the deposition of uraninites (since they are soluble when combined with oxygen) and starts the deposition of banded iron formations
2000	The Oklo natural fission reactor in Gabon goes into operation
1600	The last reserves of reduced iron are used up by the increasing atmospheric oxygen--last banded iron formations
1500	Eukaryotic cell organisms develop
1500-600	Rise of multicellular organisms
580-545	Fossils of Ediacaran organisms are made
545	Cambrian explosion of hard-bodied organisms
528-526	Fossilization of the Chengjiang site
517-515	Fossilization of the Burgess Shale
500-450	Rise of the fish--first vertebrates



Evolutionary Timeline

Time (Myr ago)	Event
4600	Formation of the approximately homogeneous solid Earth by planetesimal accretion
4300	Melting of the Earth due to radioactive and gravitational heating which leads to its differentiated interior structure as well as outgassing of molecules such as water, methane, ammonia, hydrogen, nitrogen, and carbon dioxide
4300	Atmospheric water is photodissociated by ultraviolet light to give oxygen atoms which are incorporated into an ozone layer and hydrogen molecules which escape into space
4000	Bombardment of the Earth by planetesimals stops
3800	The Earth's crust solidifies--formation of the oldest rocks
3800	Condensation of atmospheric water into oceans
3500-2800	Prokaryotic cell organisms develop
3500-2800	Beginning of photosynthesis by blue-green algae which works to strengthen the ozone layer and change the Earth from a reducing to an oxidizing one
2400	Rise in the concentration of oxygen molecules stops the formation of iron pyrites (iron combined with oxygen) and starts the deposition of banded iron formations
2000	The Oklo natural fission reactor in Gabon goes into operation
1600	The last reserves of reduced iron are used up by the iron pyrites
1500	Eukaryotic cell organisms develop
1500-600	Rise of multicellular organisms
580-545	Fossils of Ediacaran organisms are made
545	Cambrian explosion of hard-bodied organisms
528-526	Fossilization of the Chengjiang site
517-515	Fossilization of the Burgess Shale
500-450	Rise of the fish--first vertebrates

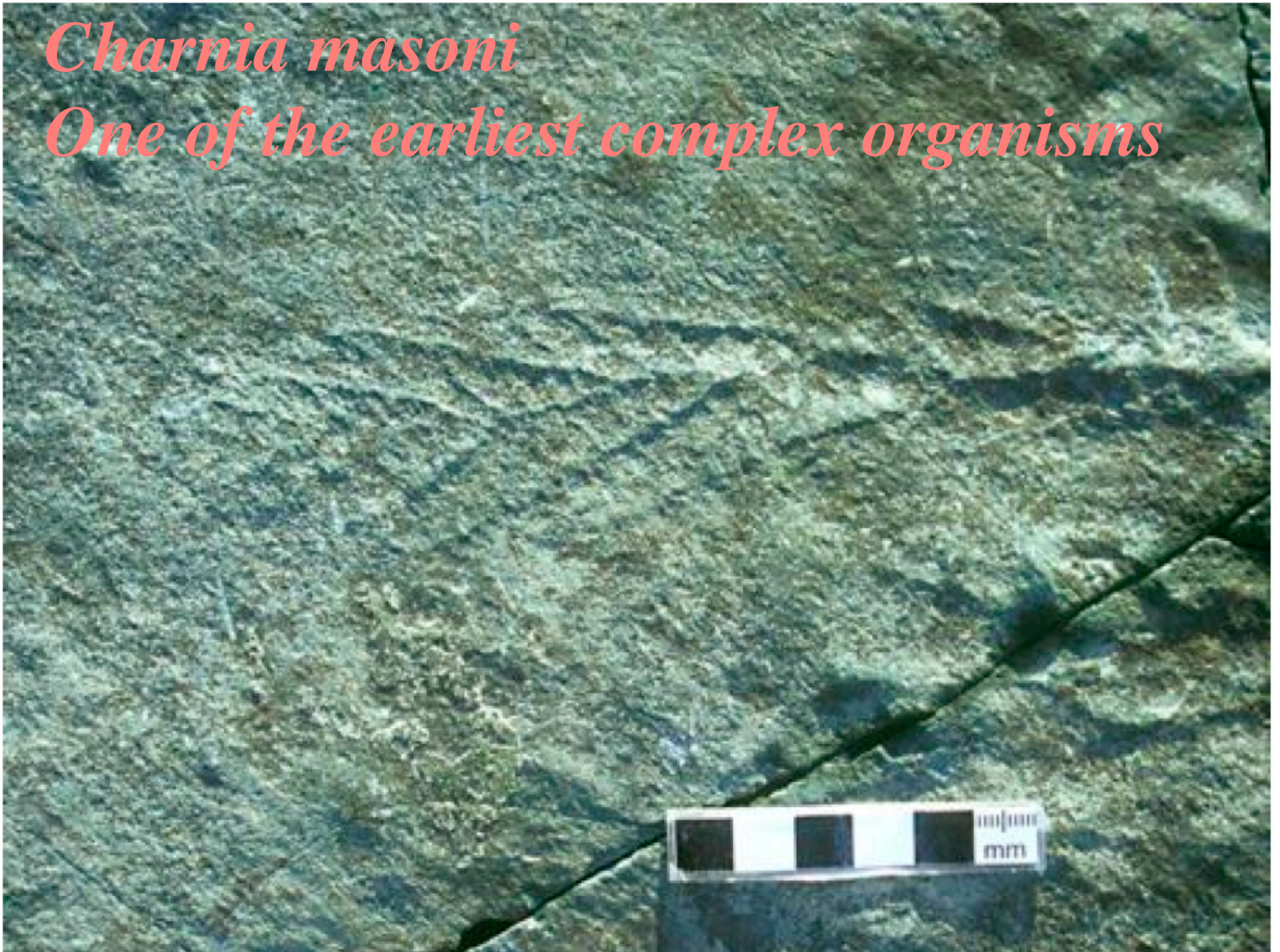


**Drook Formation rocks near Portugal Cove,
Southeastern Newfoundland**



Charnia masoni

One of the earliest complex organisms

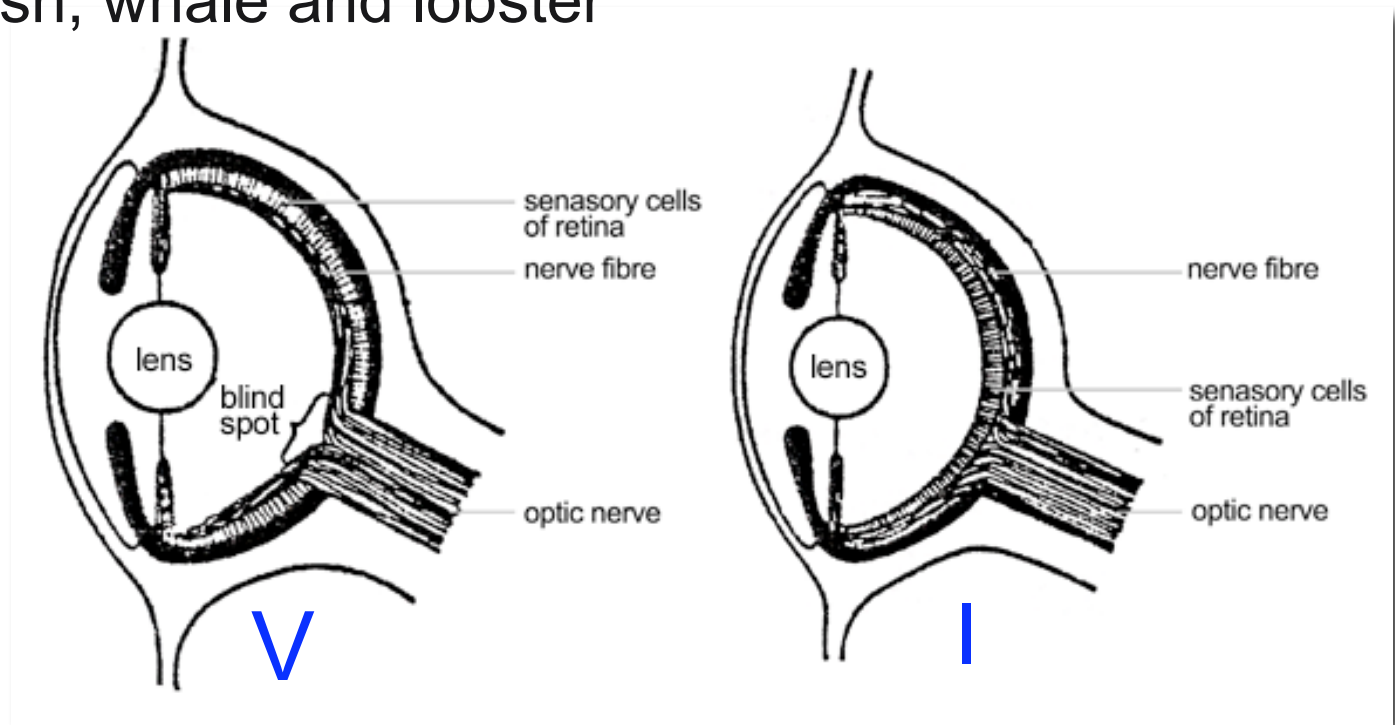


430	Waxy coated algae begin to live on land
420	Millipedes have evolved--first land animals
375	The Appalachian mountains are formed via a plate tectonic collision between North America, Africa, and Europe
375	Appearance of primitive sharks
350-300	Rise of the amphibians
350	Primitive insects have evolved
350	Primitive ferns evolve--first plants with roots
300-200	Rise of the reptiles
300	Winged insects have evolved
280	Beetles and weevils have evolved
250	Permian period mass extinction
230	Roaches and termites have evolved
225	Modern ferns have evolved
225	Bees have evolved
200	Pangaea starts to break apart
200	Primitive crocodiles have evolved
200	Appearance of mammals
145	<i>Archaeopteryx</i> walks the Earth
136	Primitive kangaroos have evolved
100	Primitive cranes have evolved
90	Modern sharks have evolved
65	K-T Boundary--extinction of the dinosaurs and beginning of the reign of mammals
60	Rats, mice, and squirrels have evolved
60	Hérons and storks have evolved
55	Rabbits and hares have evolved
50	Primitive monkeys have evolved

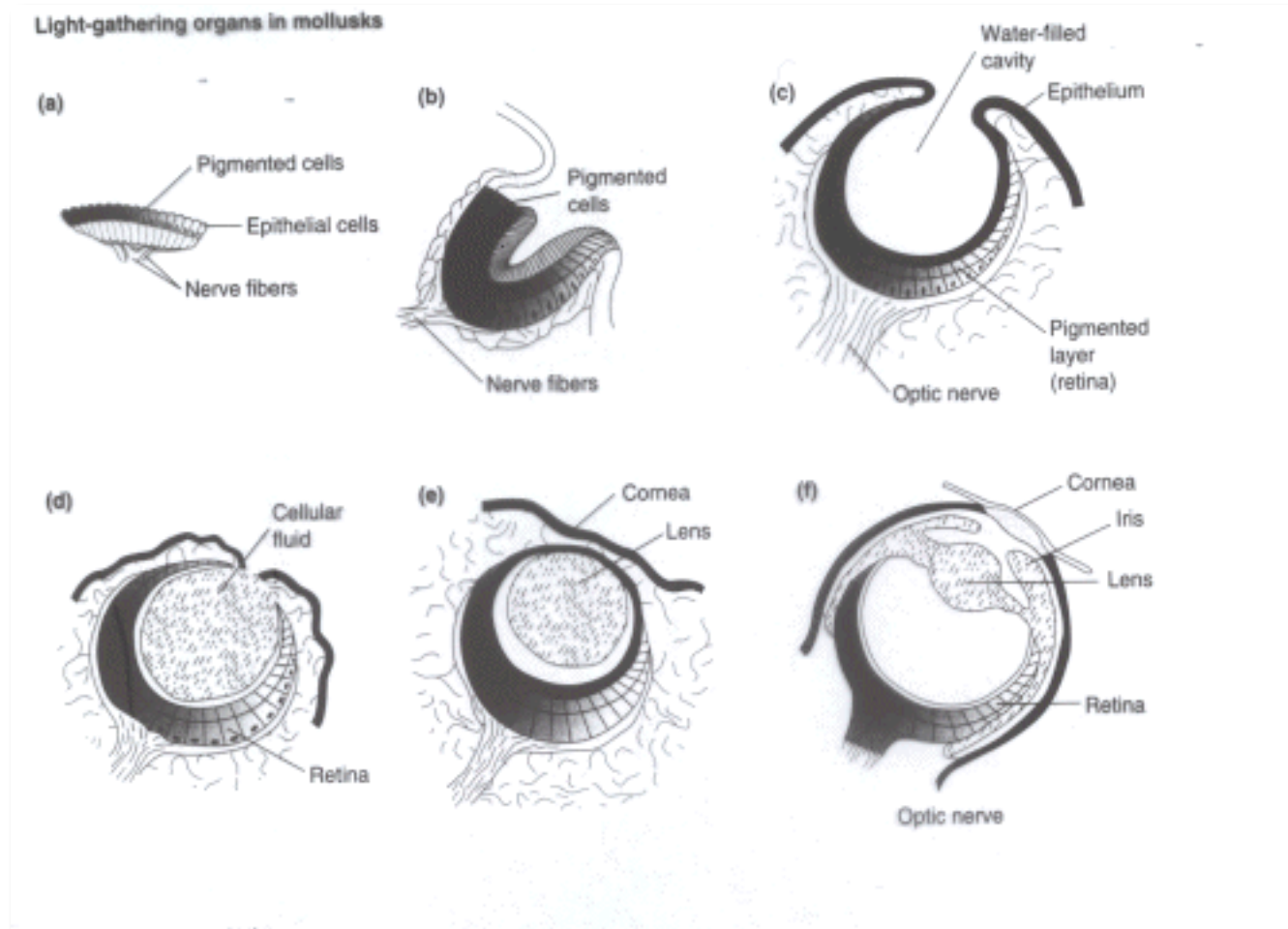
Nervous systems developed soon after multicellularity but then a large delay until intelligence arose

Understanding Convergent and Divergent Forces in Evolution — the repertoire of form and function, the independent evolution of similar structures or functions in similar or different environments

- Wings in Insects, Birds, Bats and Reptiles
- Jointed legs in insects and vertebrates
- Tail fin of fish, whale and lobster

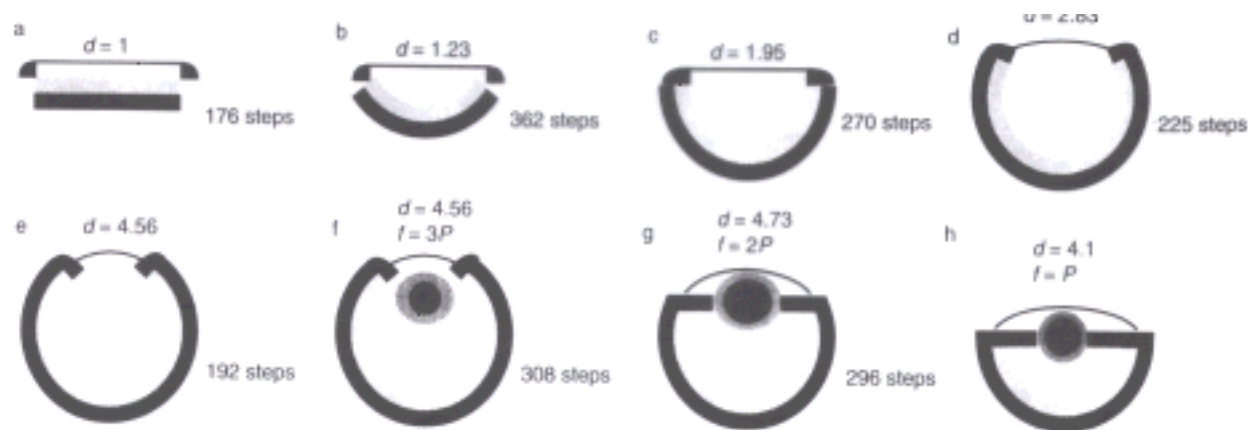


How to Evolve an Eye

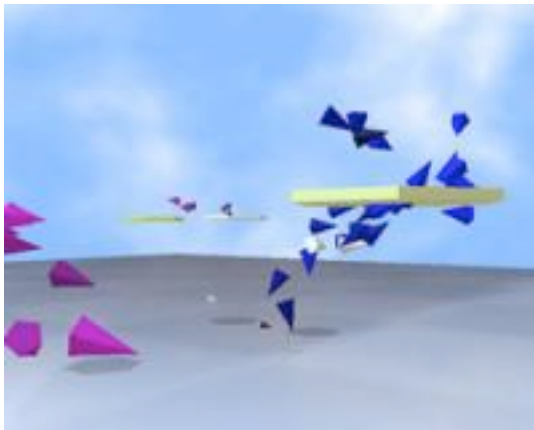
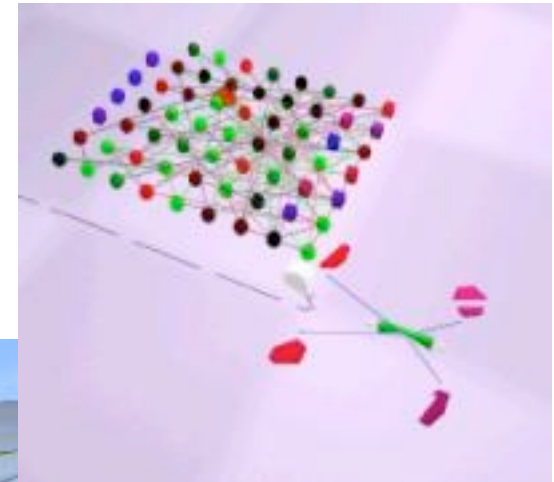
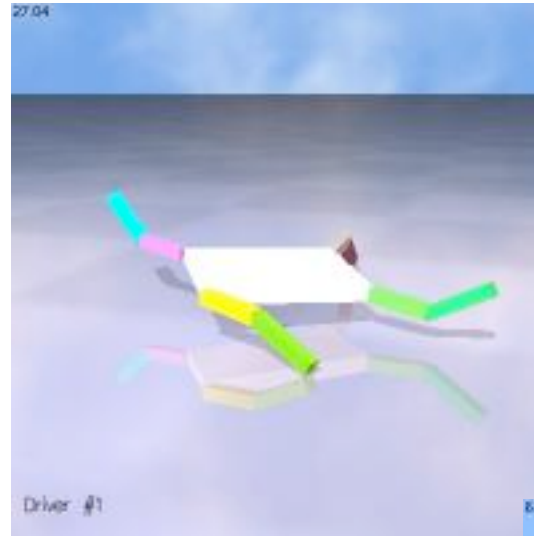
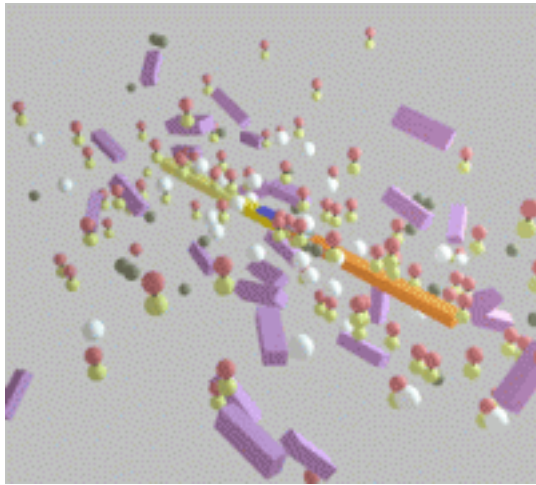


How to Evolve an Eye in 2000 Easy Steps!

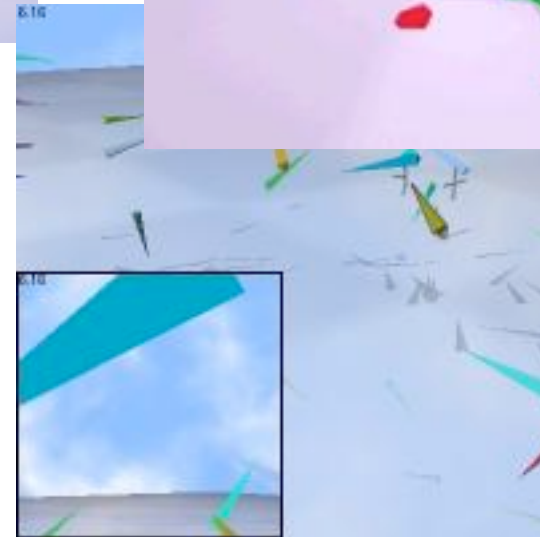
- Shape of the eye changes at random by no more than 1%
- Selection retains only those changes that improved the optical performance of the eye (ability to resolve objects)
- 2000 steps would generate a vertebrate eye.
- For realistic values of heritability and strength of selection, this would take 400,000 generations
- If one generation = one year then an eye takes less than half a million years to evolve



<http://www.spiderland.org/>



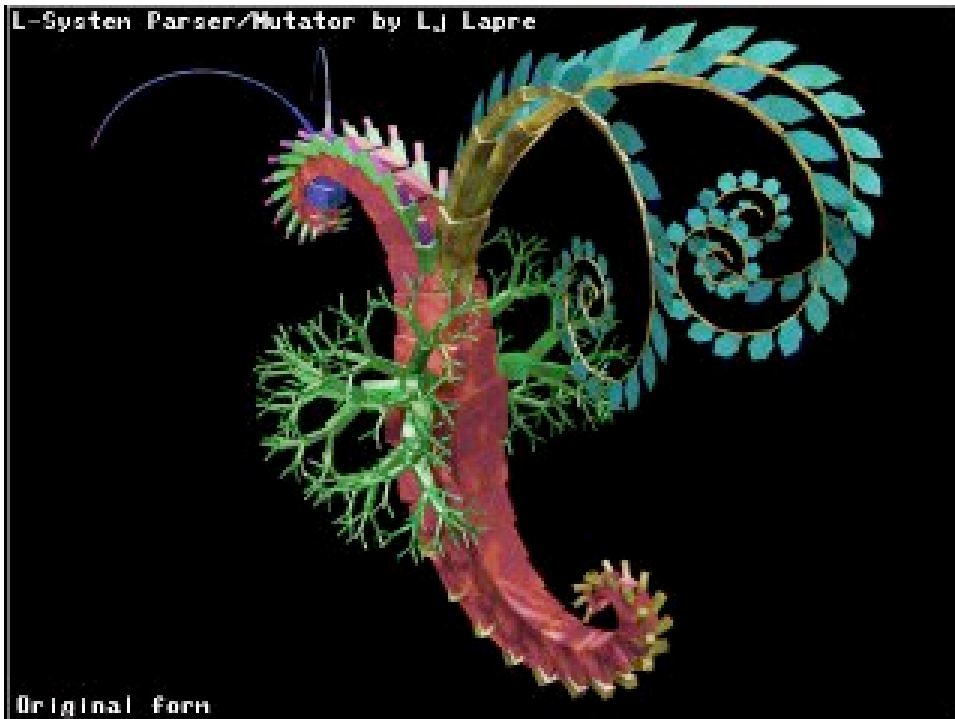
Breve



Lee Spector and Jon Klein (Hampshire College)



L-System Parser/Mutator by L.J. Lapre

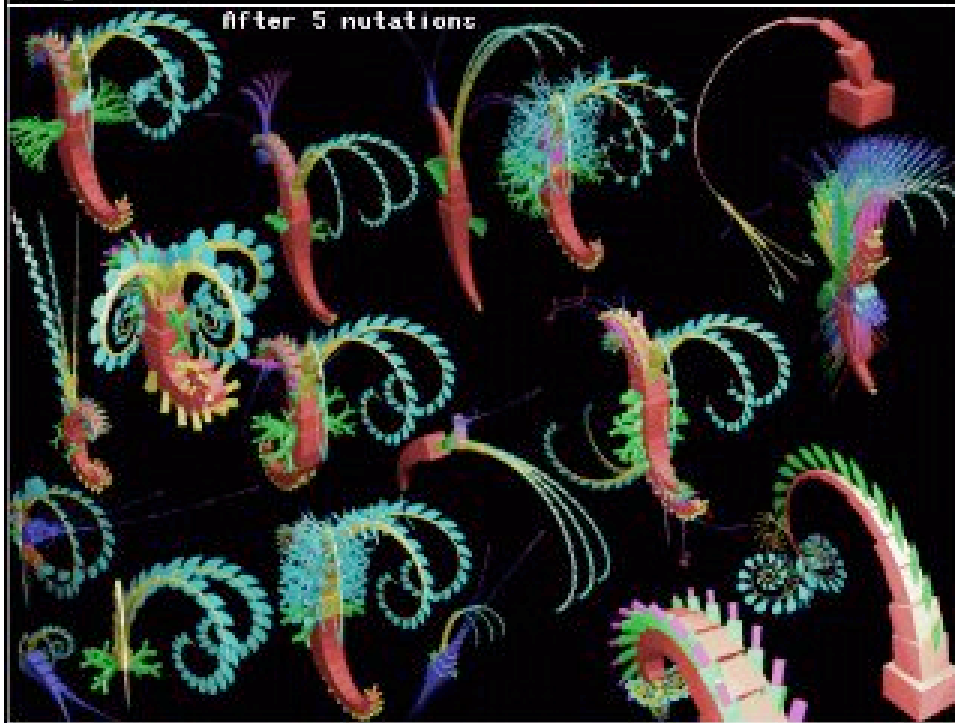


Original form

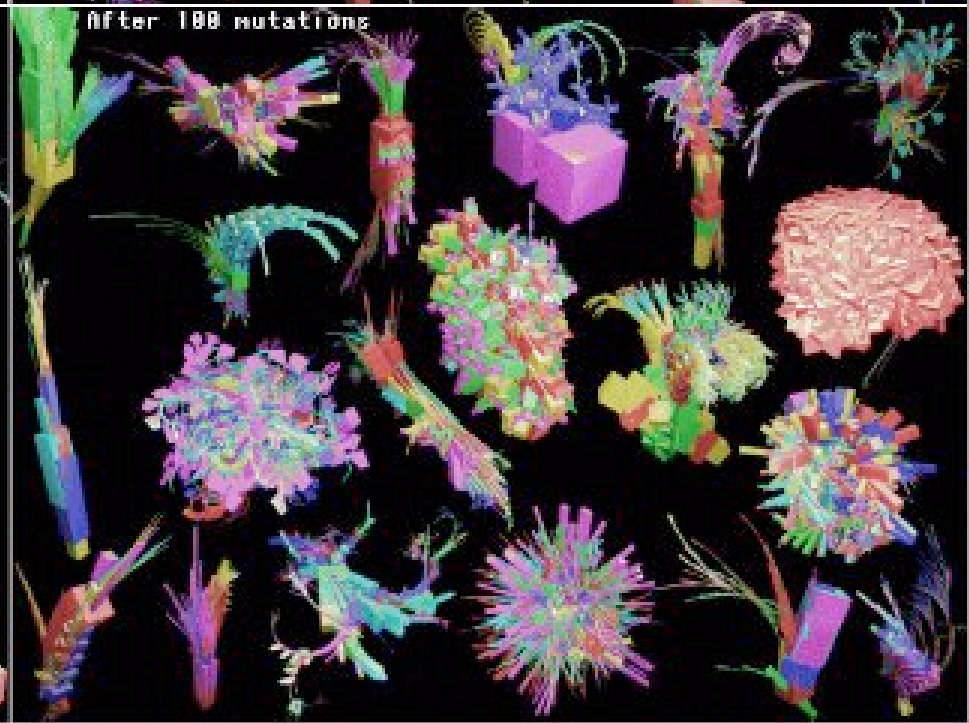
After 25 mutations



After 5 mutations



After 100 mutations

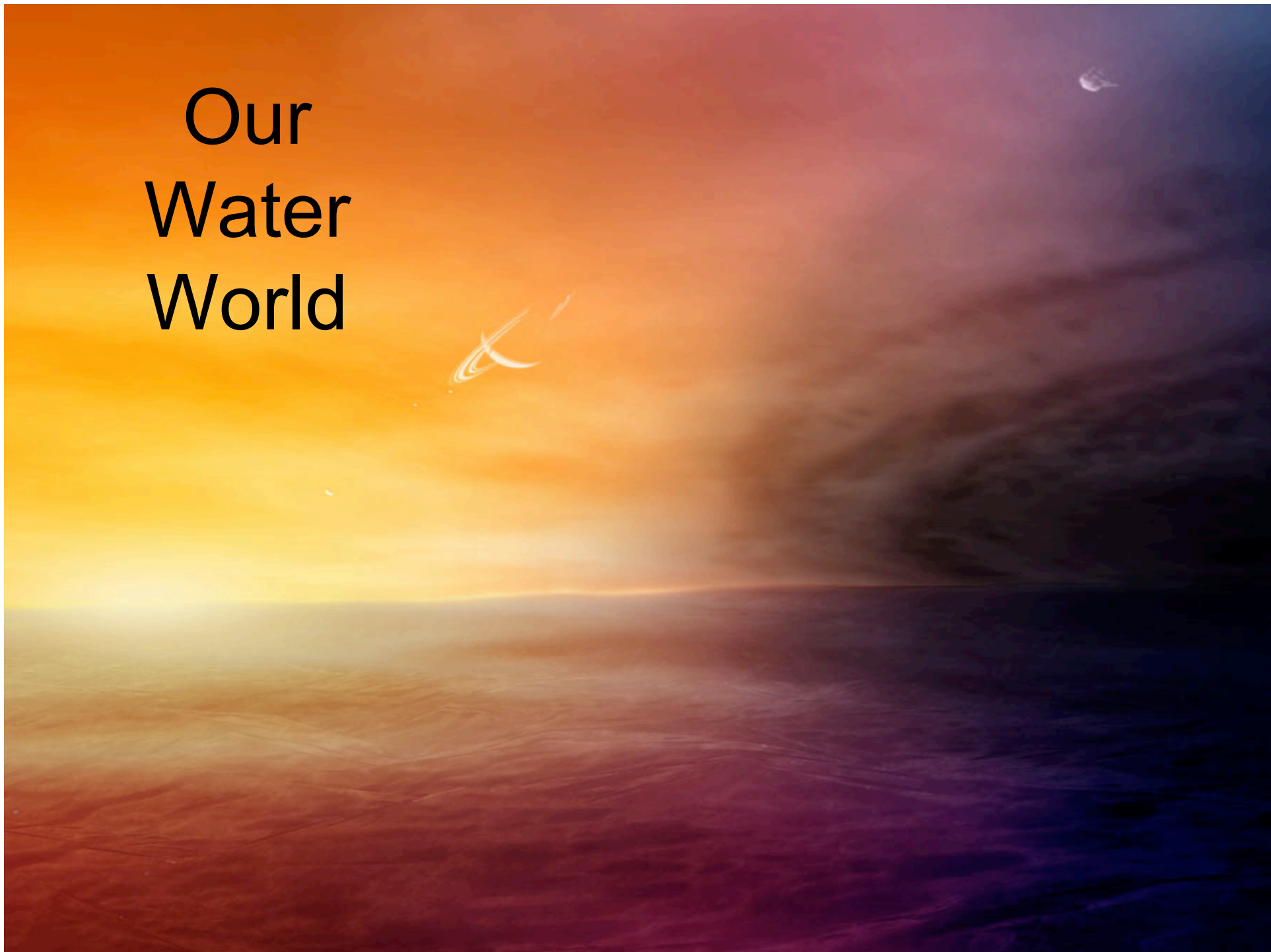


Connecting to the Computing

- Goal: Reconstruction of the mechanisms, tempo and dynamics of major evolutionary transitions, towards explanatory models and predictive models for novel systems design (constructive evolution)
- Software: Wide range of tools from phylogenetic algorithms, sequence evolution simulators, agent based tools and artificial life simulators to computer graphics O(50-100)
- Dominated by heuristics, stochastic and evolutionary algorithms (large-sampling spaces) most existing tools have limited biological context
- Data analysis leveraged genomics and comparative analysis
- Fortran, C, C++, Lisp, Objective-C, scripting languages, pushGP, etc.
- Systems: Workstations, SMP, clusters, little use of grid or HPC
- Architecture: Wide open possibilities to exploit compact cores for agents or monte carlo modeling, can scale to millions, but need problems worth scaling
- **The BIG Opportunity: from a data light theory derive the competitive exclusion principle and structure of observed ecosystems**

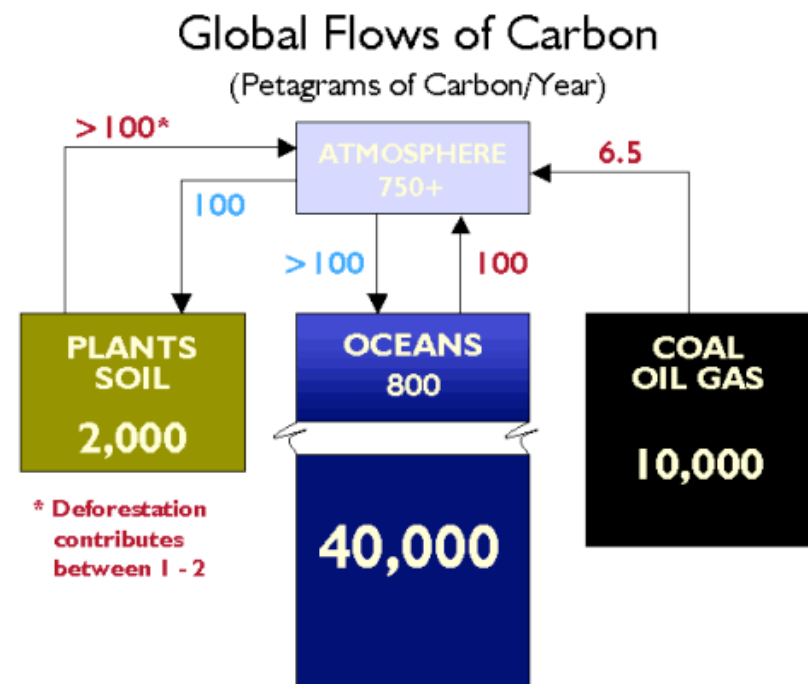
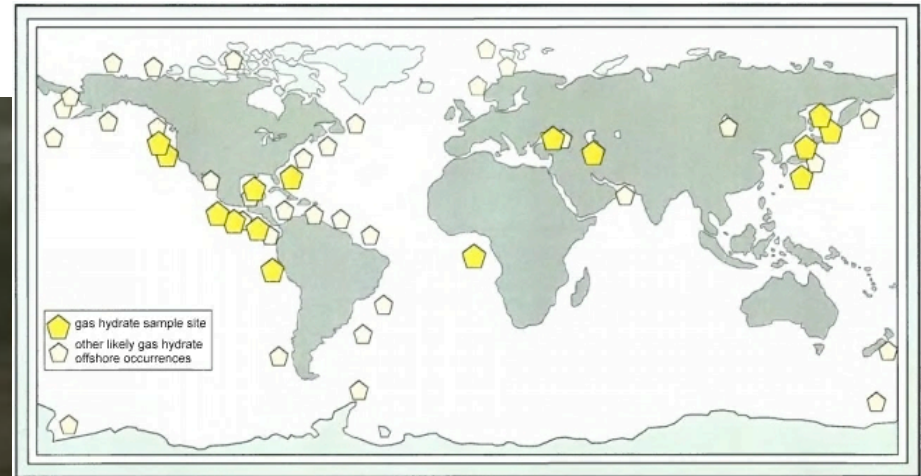


Our Water World

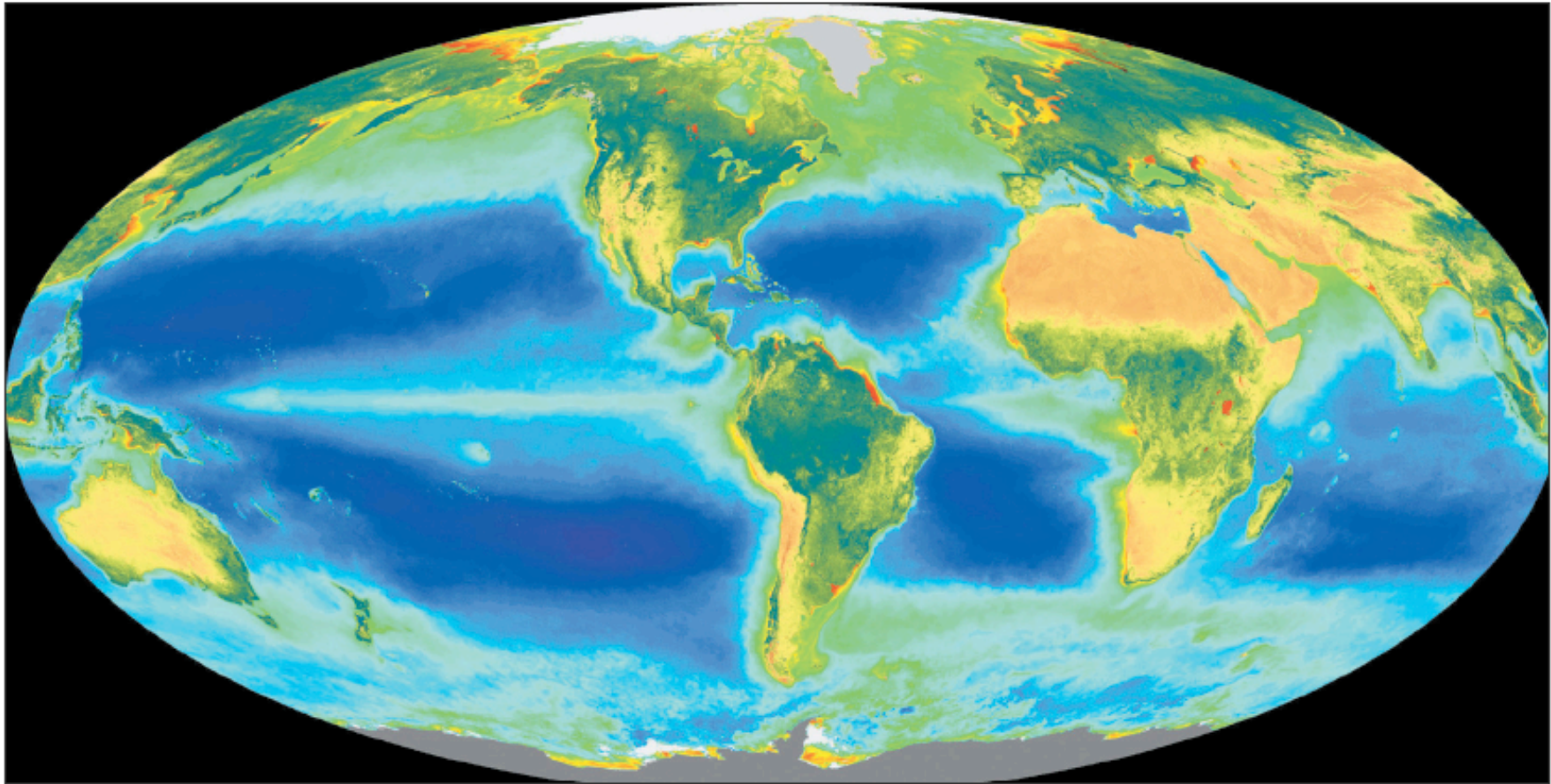


Methane Hydrates -1liter of ice contains 186 liters of methane

500-2500 Gtons of Carbon



Global Surface Chlorophyll



SeaWiFS Project, NASA/Goddard Space Flight Center and ORBIMAGE

Figure 3. Global map of annual average surface ocean chlorophyll, a measure of photosynthetic (autotrophic) biomass, derived from the SeaWiFS satellite ocean color sensor. The satellite data clearly illustrate the large-scale spatial patterns of ocean biomes driven by ocean mixing, light limitation, subsurface nutrient and iron fluxes, and atmospheric iron inputs. The ocean color scale is approximately logarithmic, with more than two orders of magnitude of change from the low biomass/low nutrient subtropical gyres (blue) to coastal upwelling regimes (yellow/orange).

Phytoplankton Bloom Bering Sea

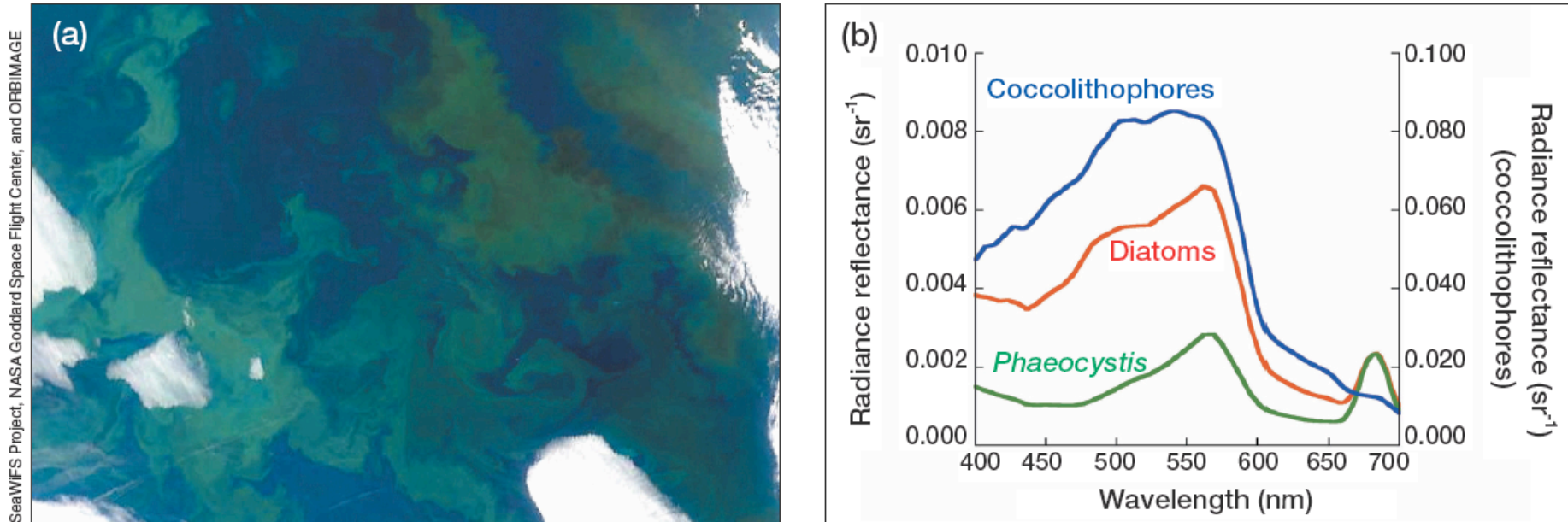


Figure 6. Under some conditions, measurements of ocean color can reveal much more than the concentration of chlorophyll. (a) True color satellite image of a phytoplankton bloom in the Bering Sea on June 7, 2001 (292 km x 200 km centered near 58.7°N, 177°W). During this period, shipboard sampling indicated blooms dominated by diatoms and the prymnesiophyte *Phaeocystis* in close proximity, probably corresponding to the lighter and darker green features in the image. Coccolithophore blooms are highly reflective and may be responsible for the brighter features in the SW corner of the image. (b) In-water measurements of hyperspectral ocean color (reflectance: the ratio of upwelling radiance to downwelling solar irradiance at the surface) from the Bering Sea reveal striking differences between blooms (note the scale change for the coccolithophores); not only does the brightness of the water (average reflectance) vary significantly, shapes of the spectra differ because pigmentation, cell size, and the quantum yield of sun-induced chlorophyll fluorescence (the peaks near 680 nm) influence the measurements, providing a key for remote sensing of species composition and perhaps physiological condition of phytoplankton. (Data from JJ Cullen and RF Davis.)

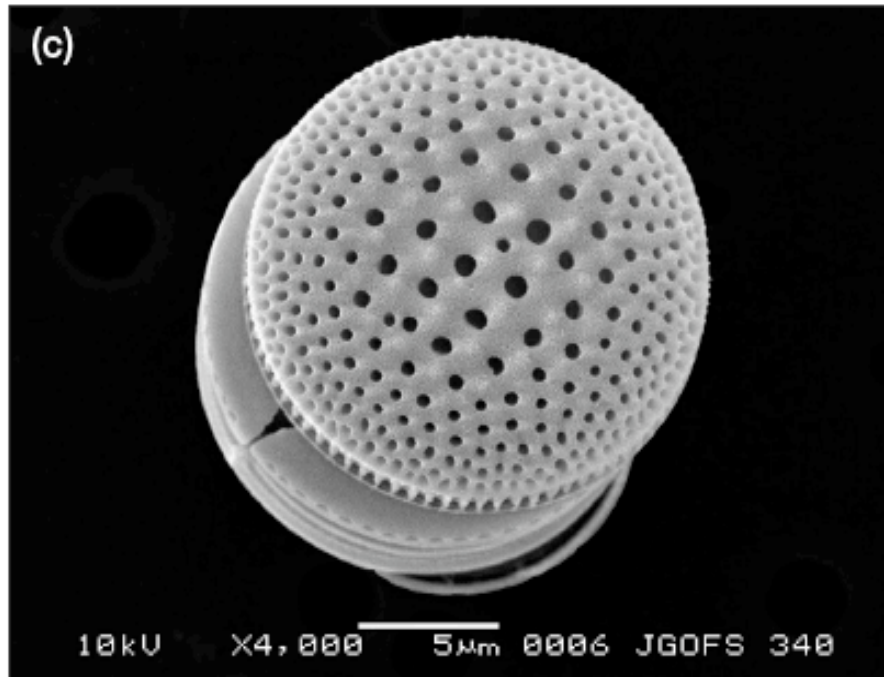
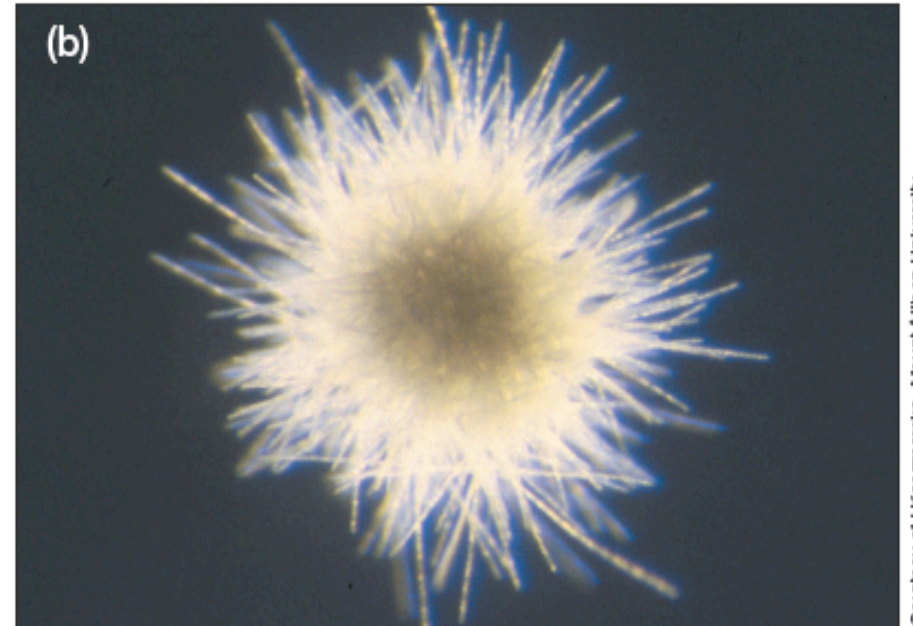
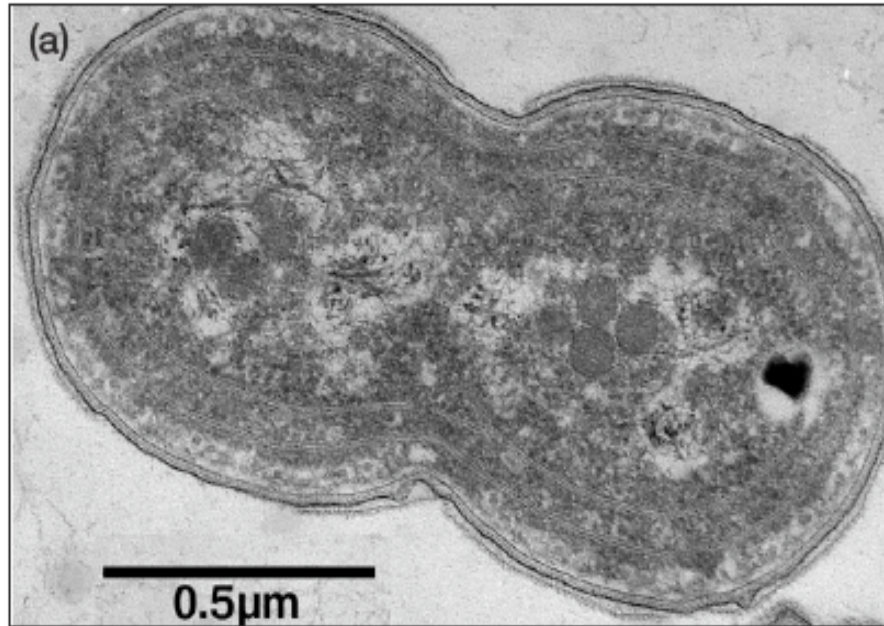


Figure 1. Oceanic photosynthetic microorganisms: (a) prokaryotic *Synechococcus* cell, a key oceanic picoplankton species, especially in nutrient poor subtropical gyres, and one of the most abundant organisms on the planet; (b) colony of cyanobacteria *Trichodesmium* (scale of image ~4mm), a nitrogen fixing species common in warm, well-stratified subtropical environments; (c) eukaryotic, open-ocean centric diatom *Thalassiosira*, an organism that forms silica shells and a contributor to the vertical export of organic carbon from the surface ocean.

(Table 2). The genomic data helping to elucidate key biogeochemical cycles also indicate that microorganisms are often able to conduct only a single specific step in a pathway. The overall transformations therefore require close coordination of microbial assem

Simple Marine Ecosystem

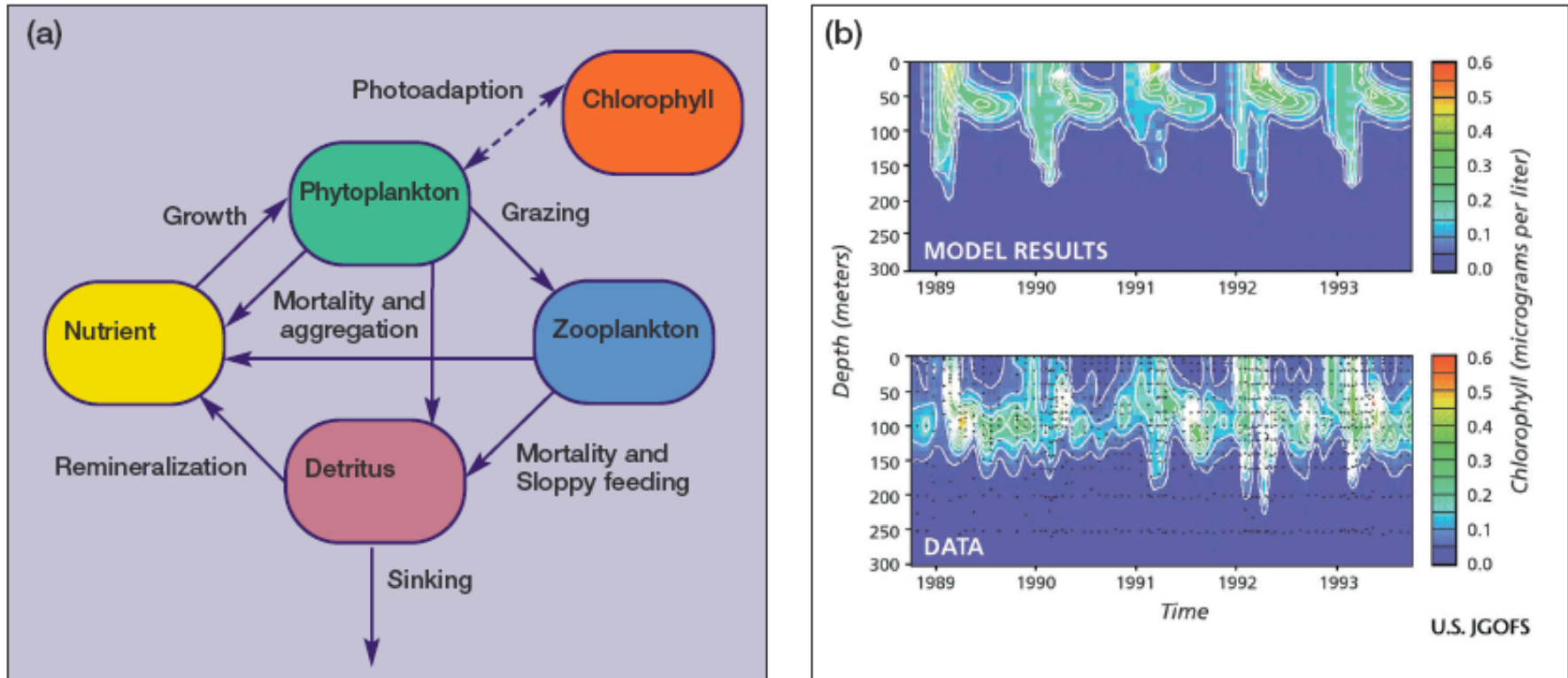


Figure 4. (a) Schematic of a simple marine ecosystem model with separate compartments for nutrients, phytoplankton, zooplankton, detritus (collectively NPZD), and chlorophyll. (b) Depth-versus-time comparison of observed and simulated chlorophyll concentrations from a 1-D version of the model applied to a multi-year record from the Bermuda Atlantic Time-series Study (BATS) site in the Sargasso Sea. Deep convection in the winter brings nutrients to the surface, generating a winter/spring phytoplankton bloom. During the summer, biological export drives surface nutrients and chlorophyll to very low levels. A subsurface, deep chlorophyll maximum forms where light from above and nutrients from below are both available.

Emergent Biogeography of Microbial Communities in a Model Ocean

Michael J. Follows,^{1*} Stephanie Dutkiewicz,¹ Scott Grant,^{1,2} Sallie W. Chisholm³

Fig. 1. Annual mean biomass and biogeography from single integration. (A) Total phytoplankton biomass ($\mu\text{M P}$, 0 to 50 m average). (B) Emergent biogeography: Modeled photo-autotrophs were categorized into four functional groups; color coding is according to group locally dominating annual mean biomass. Green, analogs of *Prochlorococcus*; orange, other small photo-autotrophs; red, diatoms; and yellow, other large phytoplankton. (C) Total biomass of *Prochlorococcus* analogs ($\mu\text{M P}$, 0 to 50 m average). Black line indicates the track of AMT13.

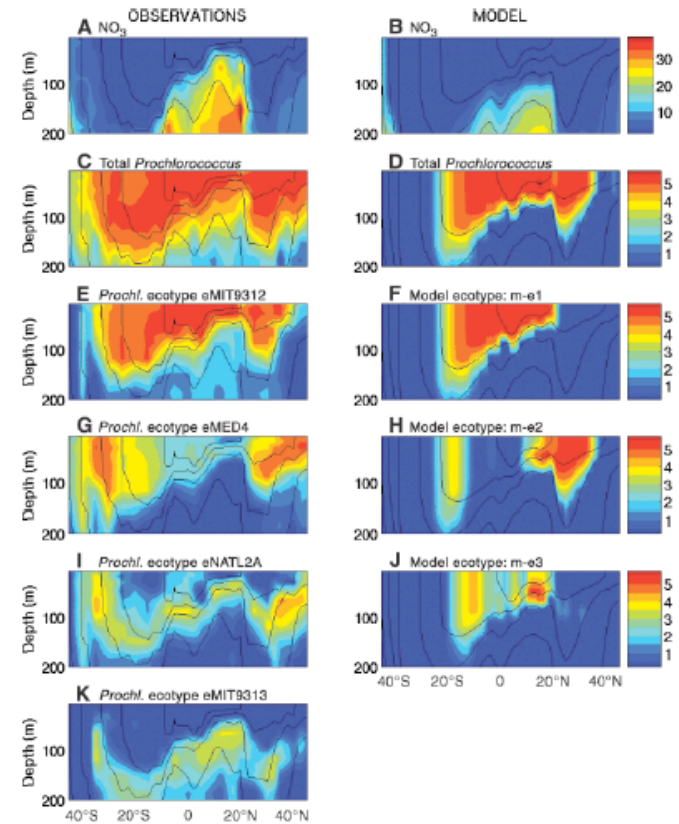
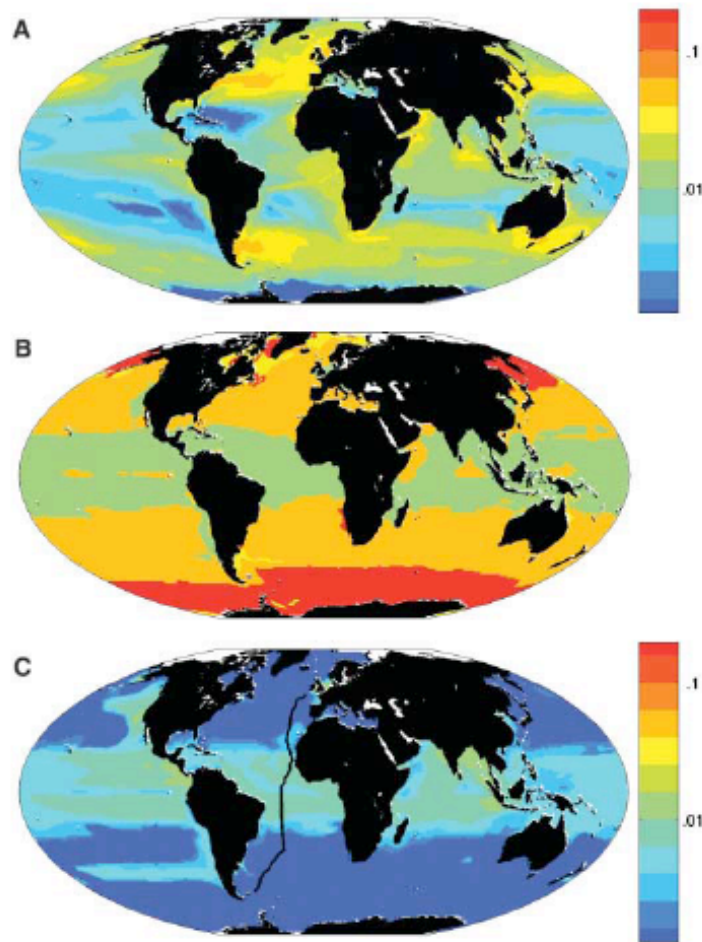


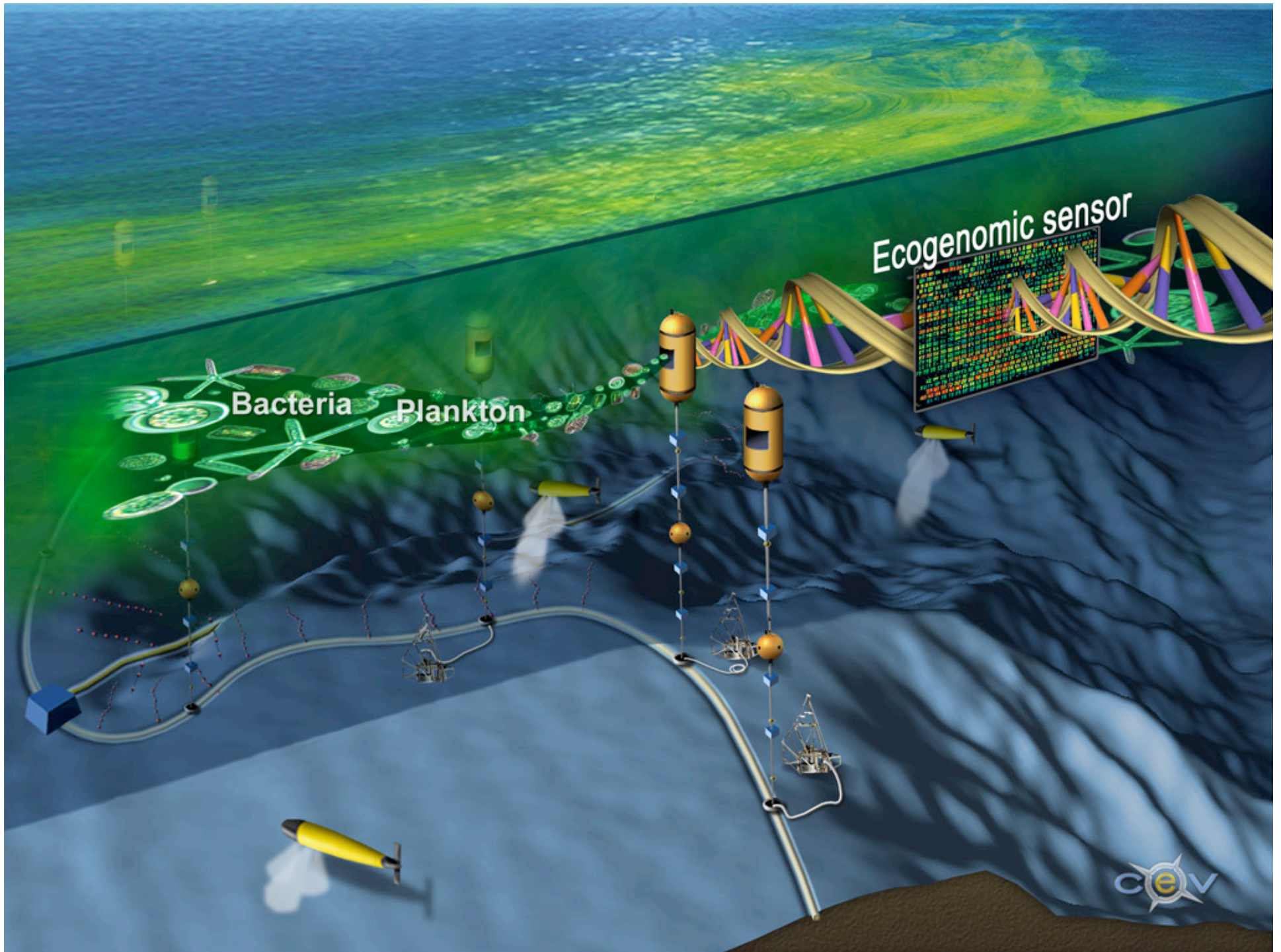
Fig. 2. Observed and modeled properties along the AMT13 cruise track. Left column shows observations (17), right column shows results from a single model integration. (A and B) Nitrate ($\mu\text{mol kg}^{-1}$); (C and D) total *Prochlorococcus* abundance [$\log(\text{cells ml}^{-3})$]. (E, G, I, and K) Distributions of the four most abundant *Prochlorococcus* ecotypes [$\log(\text{cells ml}^{-3})$] ranked vertically. (F, H, and J) The three emergent model ecotypes ranked vertically by abundance. Model *Prochlorococcus* biomass was converted to cell density assuming a quota of 1 fg P cell^{-1} (27). Black lines indicate isotherms.





NLR ———
Internet 2 ———
CA*net 4 ———

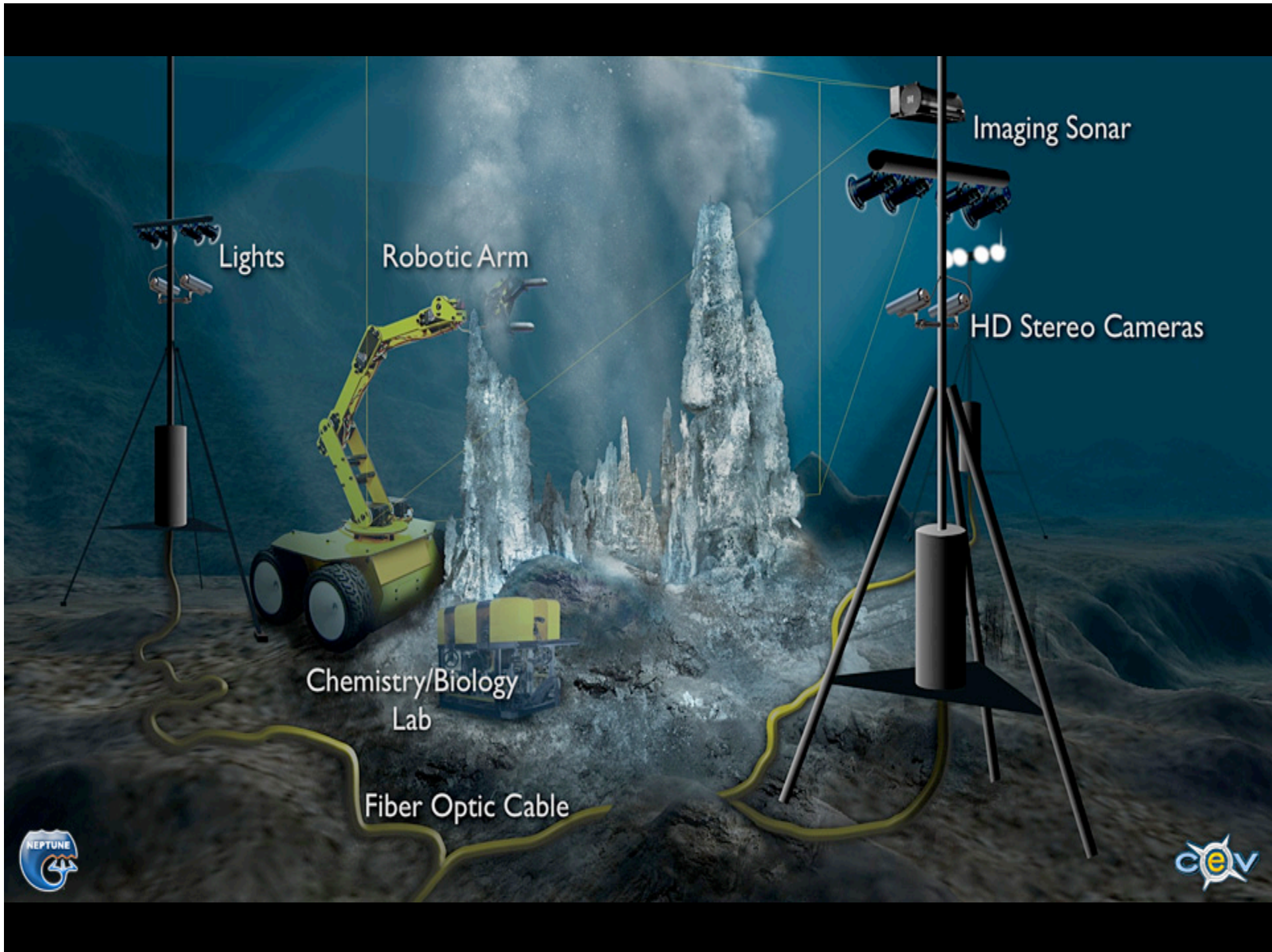




Ecogenomic sensor

Bacteria Plankton

CEV



Lights

Robotic Arm

Imaging Sonar

HD Stereo Cameras

Chemistry/Biology
Lab

Fiber Optic Cable



Connecting to the Computing

- Goal: predictive models of the physical and biological basis for the ocean's biosphere, the carbon cycle and connections to geophysical processes
- Software: Modeling and simulation and real-time sensor networks
- Ocean circulation models driving ocean ecosystems models (coupled to GCMs) wide range of software and data systems needed for integration of worldwide ocean sensor networks, wide range of levels of abstraction are used in the models
- Data analysis includes remote operations, imaging, planning, databases, autonomous probes, ecogenomics capture and analysis systems, sensor management and operations software, data assimilation, etc.
- Systems: Sensor networks with embedded computing, databases and clusters for analysis and imaging, special purpose computing to support sensor systems and data acquisition, HPC for coupling to ocean circulation and GCMs and high-performance networking
- Architecture: data analysis and sharing grid coupled to live sensors and HPC, database coupling to genomics and global climate processes
- **The BIG Opportunity: convergence of technology basis, systems integration and prototype for the emerging world-wide sensor environment and prototypes for planetary probes and remote sensing systems**



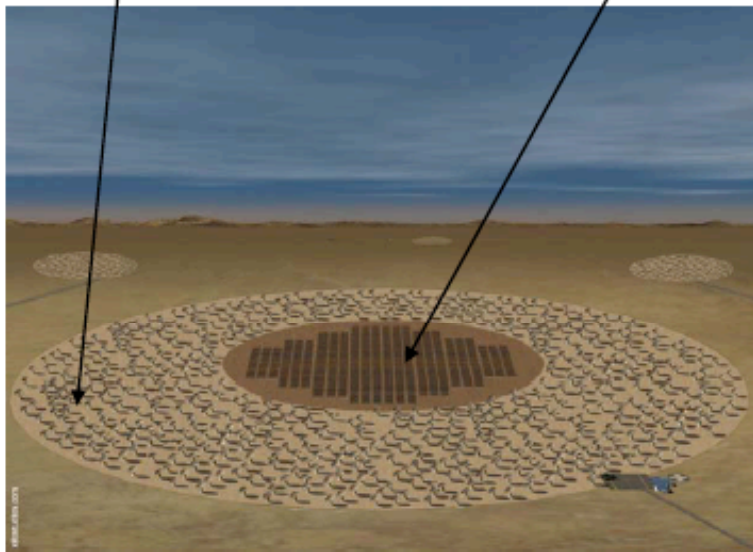
What is Out There?



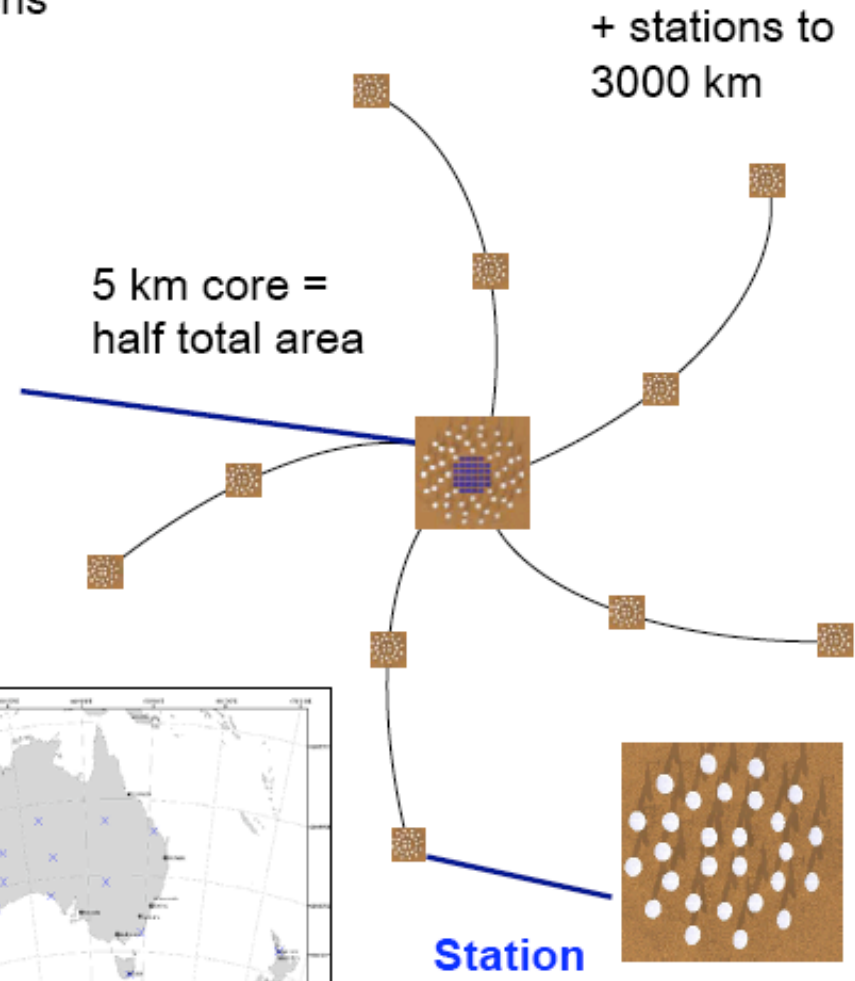
SKA: the big picture

Digital radio camera

Radio fish-eye lens

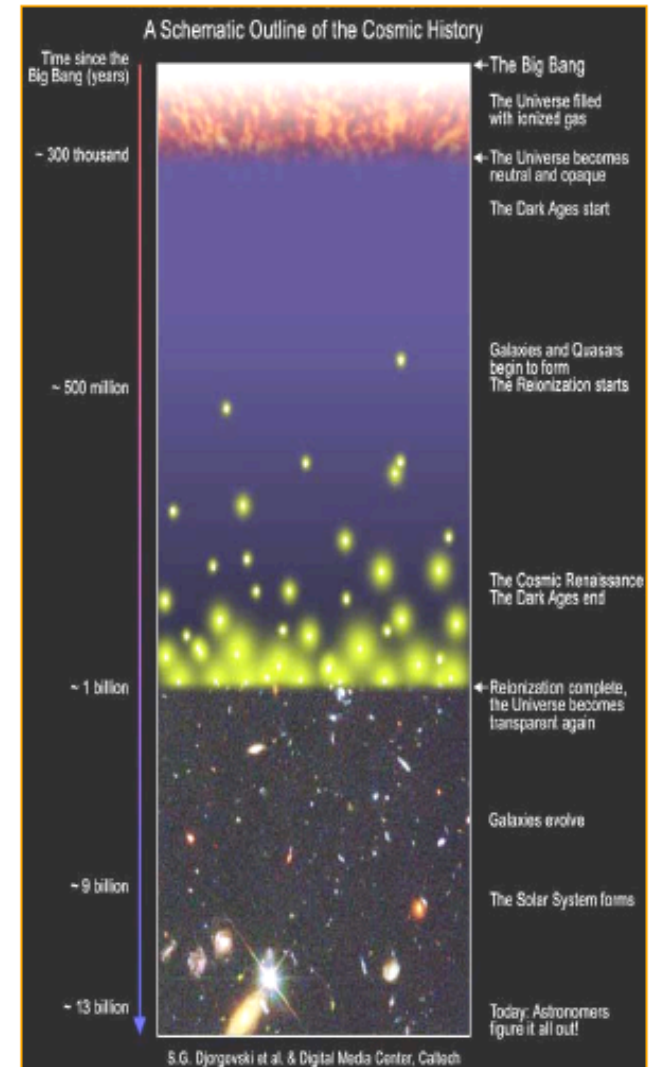


Inner core



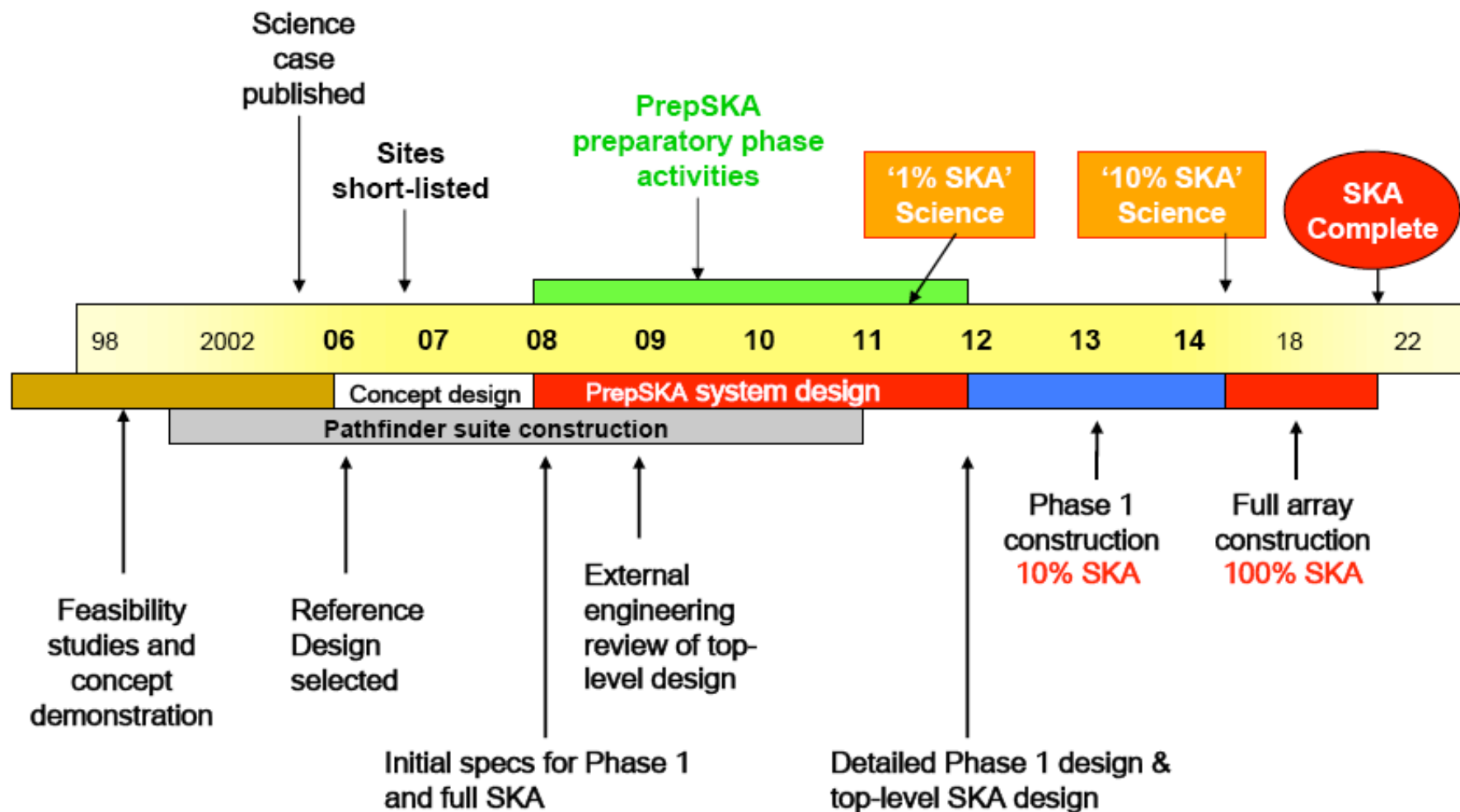
SKA science priorities

- **The first stars and galaxies in the Universe**
 - Emergence of structure
- **Large scale structure of the Universe**
 - “Dark energy”
- **Origin and evolution of cosmic magnetic fields**
 - “The magnetic Universe”
- **Gravity in the strong field case**
 - Gravitational wave detection
- **Planet formation**
 - Including search for extra-terrestrial intelligence (SETI)
- **EXPLORATION OF THE UNKNOWN**



SKA is the radio member of a suite of next-generation telescopes

SKA timeline

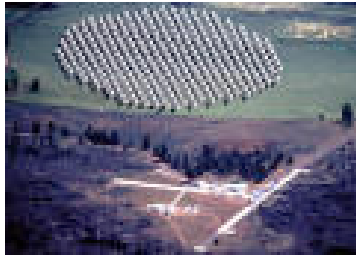
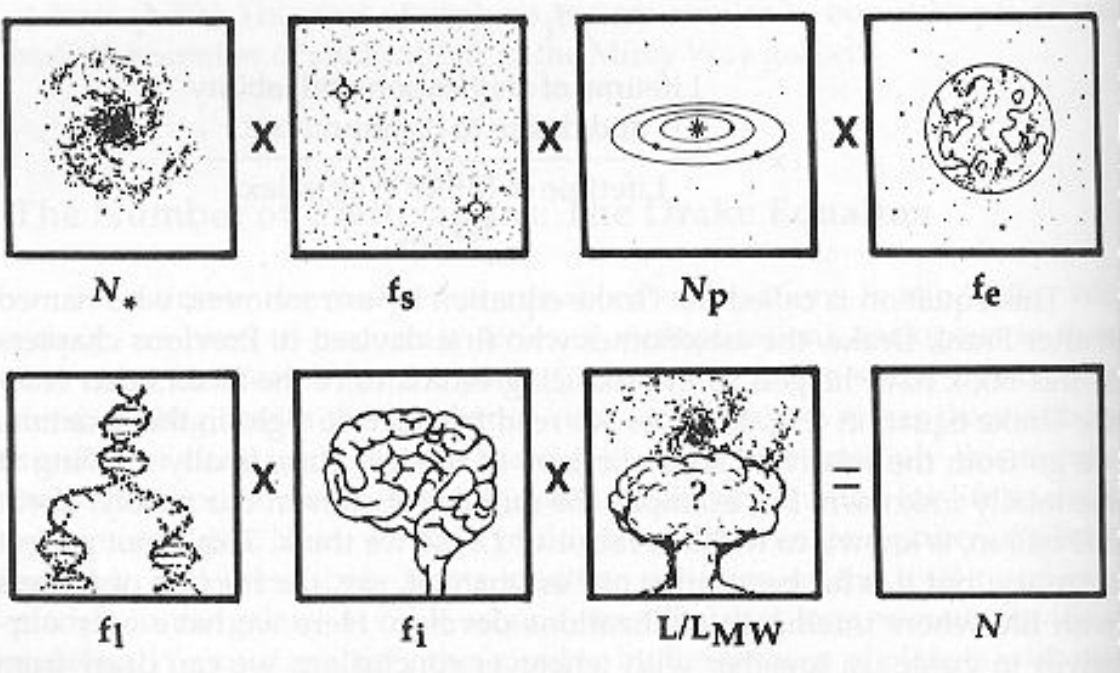




SKA – not just antennas

- **High speed data transport**
 - Tb/s from EACH station on scales of hundreds of km
 - 100 Gb/s trans-continental and trans-oceanic links
 - Longest links will rely on telcos and research networks
 - » **Need government initiatives for affordable access**
- **Signal processing**
 - Peta-ops per second
 - Need highly scaleable solutions
- **Post-processing, information management**
 - New super-computer architectures
 - Archive and sharing of data will be a major challenge
- **Infrastructure**
 - Civil, electrical (power, ...), communications
- **Operations and support**

What Becomes Possible in the 21st Century?



Connecting to the Computing

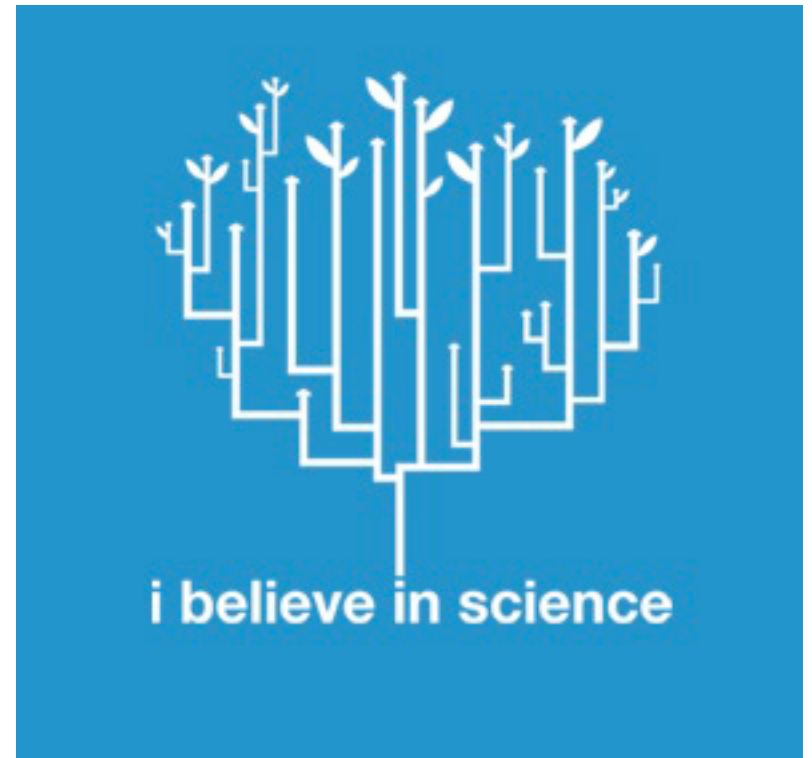
- Goal: Revolutionizing radio astronomy through largest-scale synthetic aperture telescope, multiple purpose user facility for cosmology, dark energy, planet finding, SETI search and discovery
- Software: Operational software environment included an embedded petascale system for signal analysis and beam forming, offline and near-line software systems for data mining, analysis and data assimilation for mechanistic models
- Novel integration of real-time petascale computing
- System environment also requires terabit networking and beyond and large-scale databases
- Systems: embedded systems for signal processing and filtering, digital synthesis (and perhaps also analog synthesis?) multi-user, multi-function, large-scale Exascale? analysis systems will be needed to manage data.
- Architecture: Embedded systems, HPC, Grid and Tier system for data management between partners
- **The BIG Opportunity: Progress on the big questions in cosmology, dark energy, start and planetary system formation and evolution, pushing out the detection limit for SETI and positions earth for discovery**



What Does it Mean?

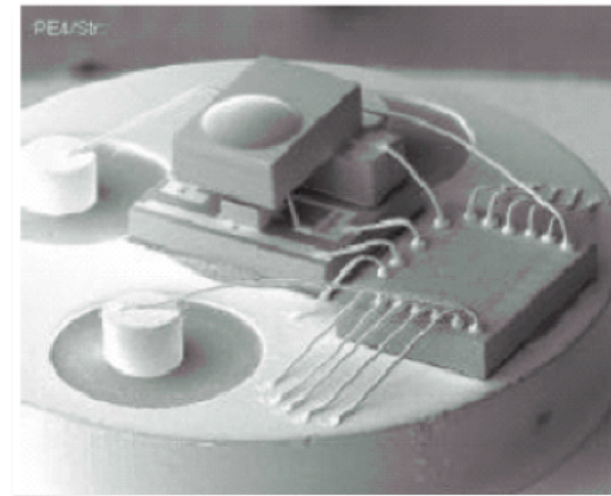
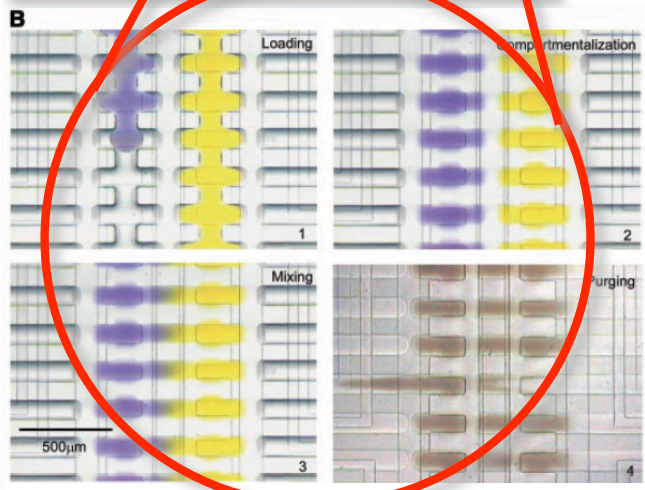
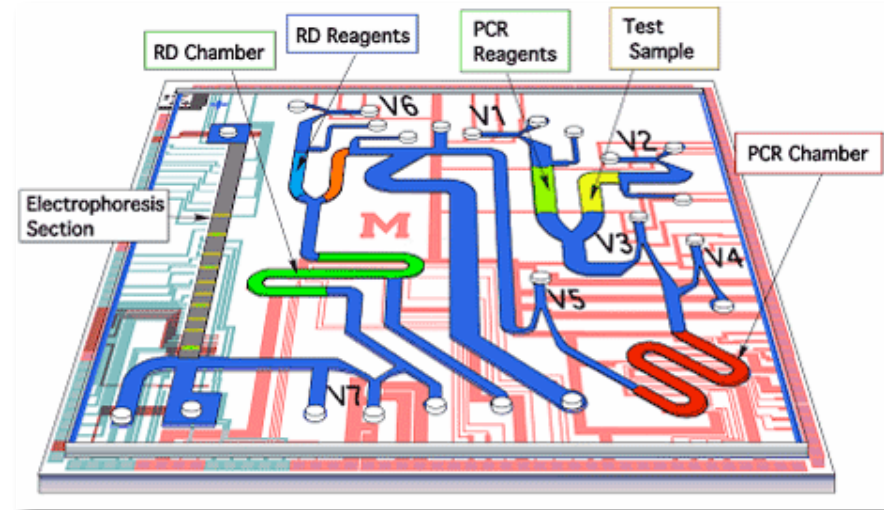
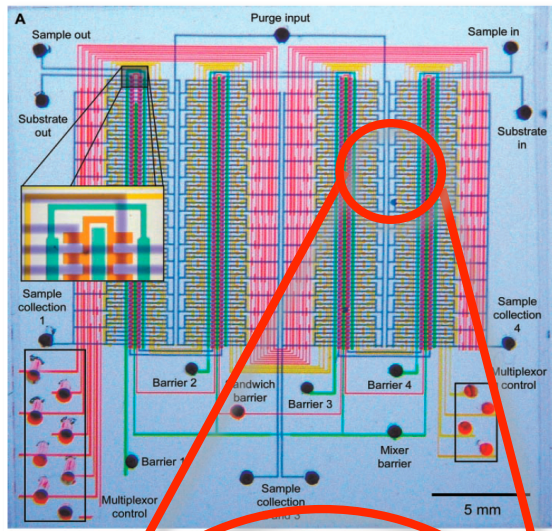
Roles in Science

- What is the sustained role of HPC in advancing science?
- How does this role differ from that of more general computing and computer science?



Molecular Biology Laboratory-on-a-Chip

What could you do with 100,000 of these?



The Great Big Roadmap

- 2010 , Sustained Open Petaflops on Real Application, First Synthetic Bacterial Organism Developed that is not booted from existing cell, commercial development of a biological CAD tool for microbial engineering, human genome sequenced for < \$100,000, human genetic screen (hapmap) for < \$10,000, 1000 genomes have been sequenced, major push to understand biological diversity of the lower eukaryotes, year of the metagenome, near real-time annotation of a bacterial genome
- 2015 Bioengineered plants demonstrated that contain multiple bacterial subsystems, Square Kilometer Array is deployed featuring production petascale system as an embedded systems, local deployment of regional sensor systems, autonomous roving environmental probes, protein folding problem essentially solved, programming models demonstrated that support large-scale parallel agent modeling systems, human genome sequenced for <\$10,000, 1km scale climate models used in insurance estimating, space tourism companies profitable, attempts at comprehensive models of global agriculture, climate and economy, attempts to recreate early evolutionary transitions in A-life worlds, use of global trade flux optimization software to improve economic performance
- 2020 Open Exascale systems in production, >TF personal systems available for < \$10,000, First Million Core systems available, first global earth sensor grid project, first deployment of self-deploying sensor networks, over 10,000 non-solar planets cataloged, first commercial human re-engineering service offered, human genome sequenced for <\$1,000, first serious attempts to create life forms from non-living materials, widespread neural implants for disease and injury treatments, widespread adoption of personalized genomics based medical treatments, widespread use of designer enzymes, investments in modeling and simulation exceed those in experiment on an annualized basis.
- 2025 PF workstations common, non-biologically derived synthetic living systems, First dedicated beacons deployed for extraterrestrial signaling, most humans sequenced at birth or in utero, predictive modeling of volcanic eruptions and earthquakes in use by national governments, development of a self-replicating ocean floor mining system
- 2030 Computers available with the computational power of a human in both language and vision skills, autonomous self-replicating space probes in development, first human born with optimized genome, first demonstration of laser propulsion of a space probe



Using Petaflops to Search for New Drugs



What should the community be doing?

- Enable a variety of "integrative" problem domains. Where the goal of the effort is to understand the emergent behavior of complex and richly interconnected systems.
- Expand by a factor of at least three (3x) the number of disciplines able to embrace the opportunities of HPC and exploit large-scale computing and data analysis capabilities.
- Expand by a factor of one hundred (100x) the size of the developer community that can develop applications for 1,000 cpus.
- Expand by a factor of ten (10x) the size of the developer community that can develop parallel applications able to run on 100,000 cpus.



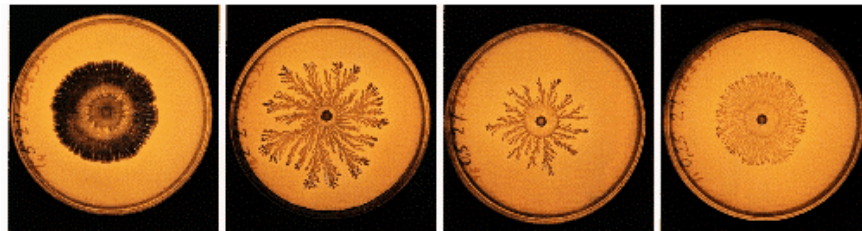
What should the community be doing?

- Developing new computer architectures/technologies/designs that can sustain a teraflop per watt and can be configured into a variety of systems ranging from handhelds to exascale supercomputers.
- Enable large-scale modeling and simulation to make significant impacts on the quality of life, on the quality of policies and on the pursuit of fundamental questions in science and humanities.
- Dramatically improve the adoption rate of new software technologies and algorithms by existing computational science communities.



Describe

Explain



Predict

Control

