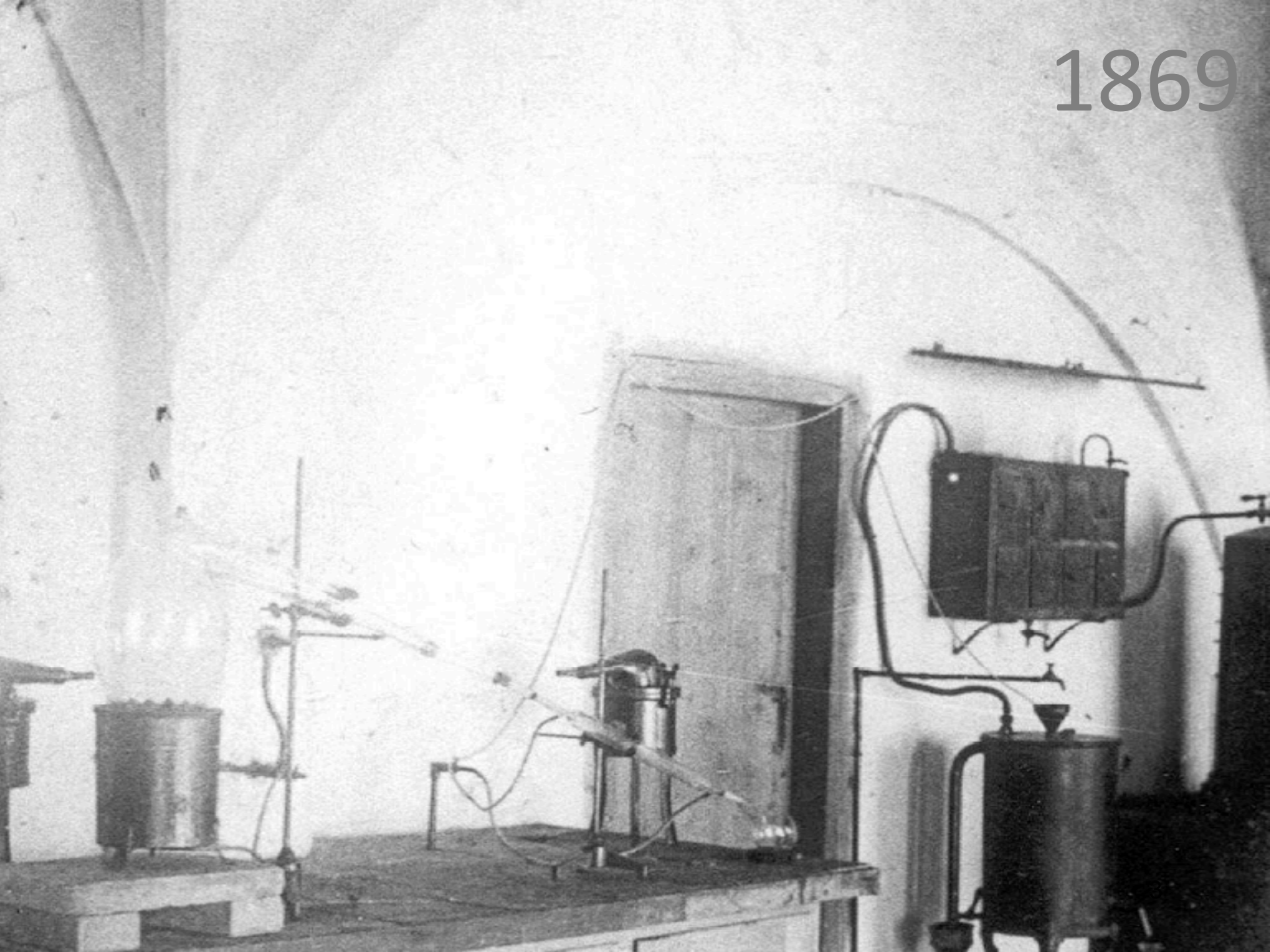
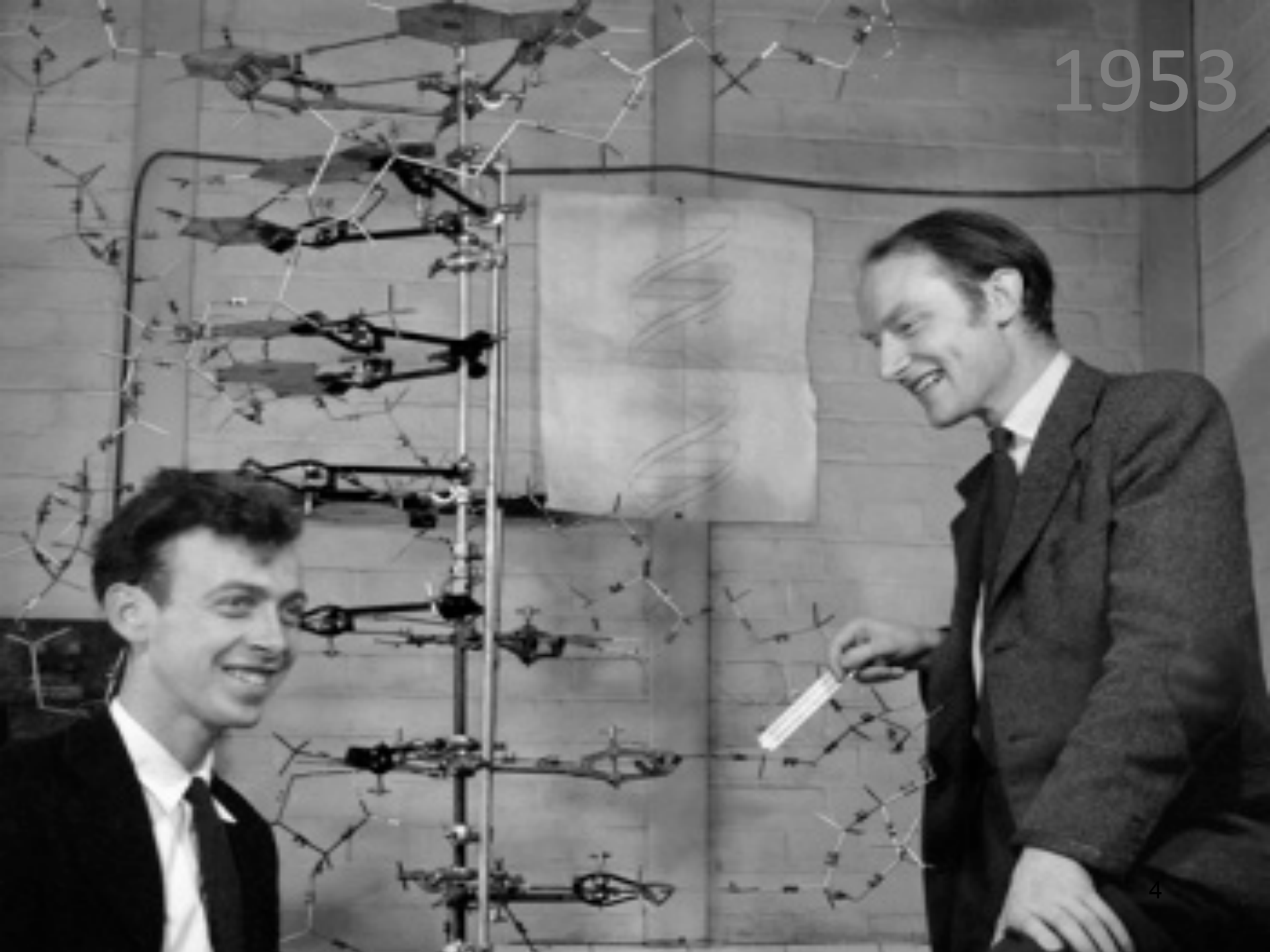
Analysis of environmental genomes using Pathway Tools

Steven Hallam | University of British Columbia
Hydrocarbon MetaPathways Tutorial, 2014

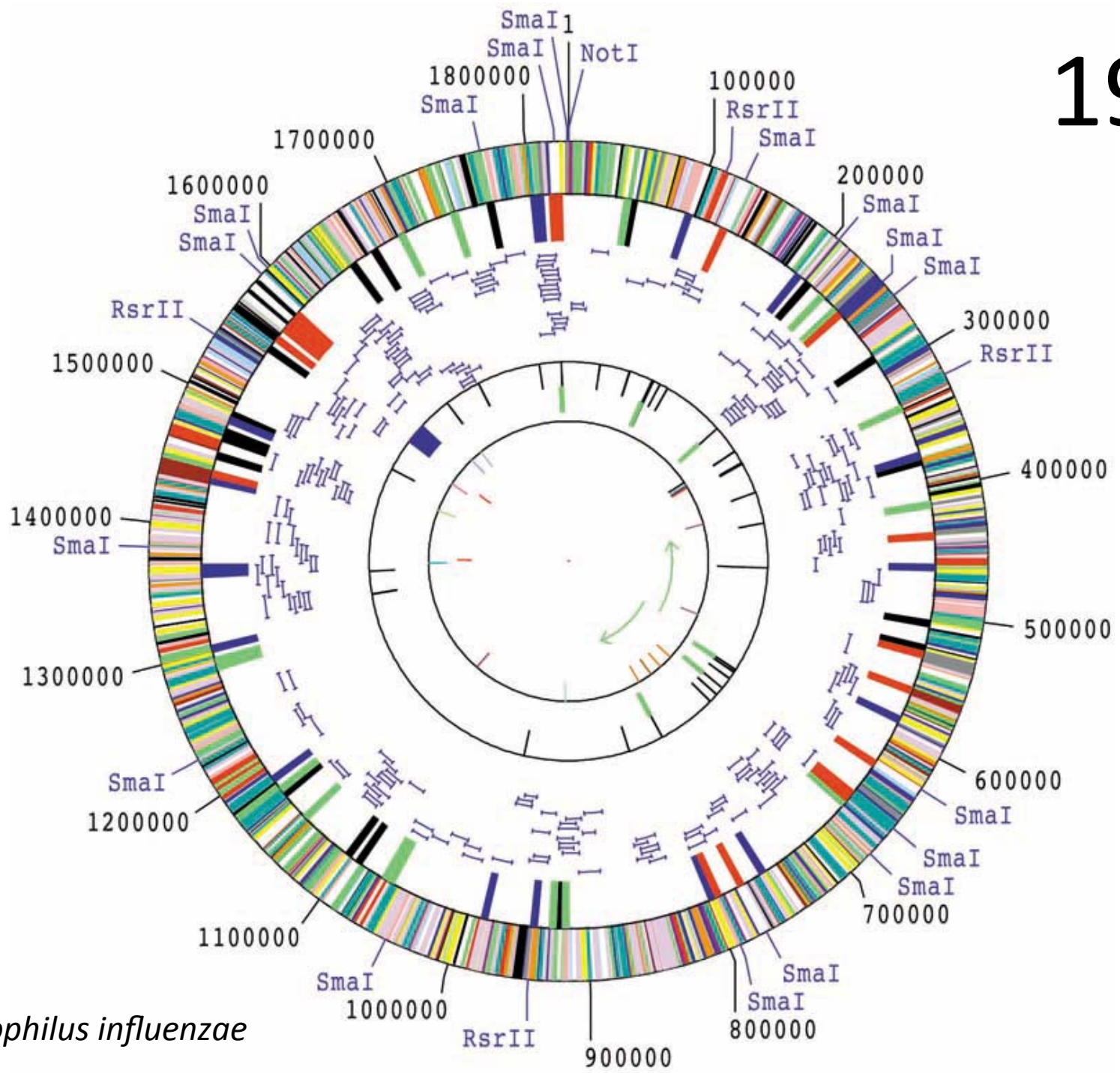
Overview

- Through the looking glass…
- MetaPathways Pipeline Development
- Hawaii Ocean Time Series


1869

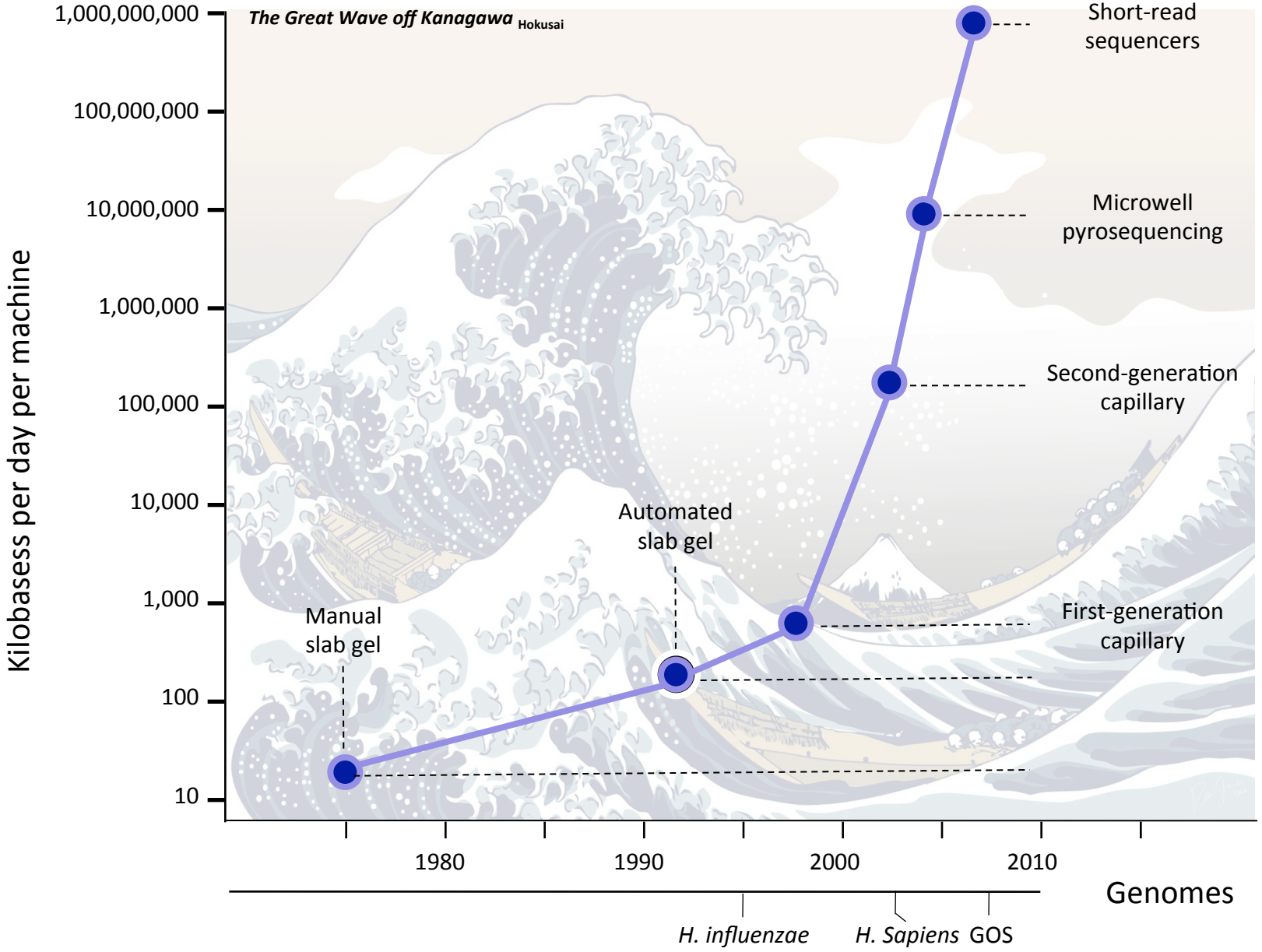

1953

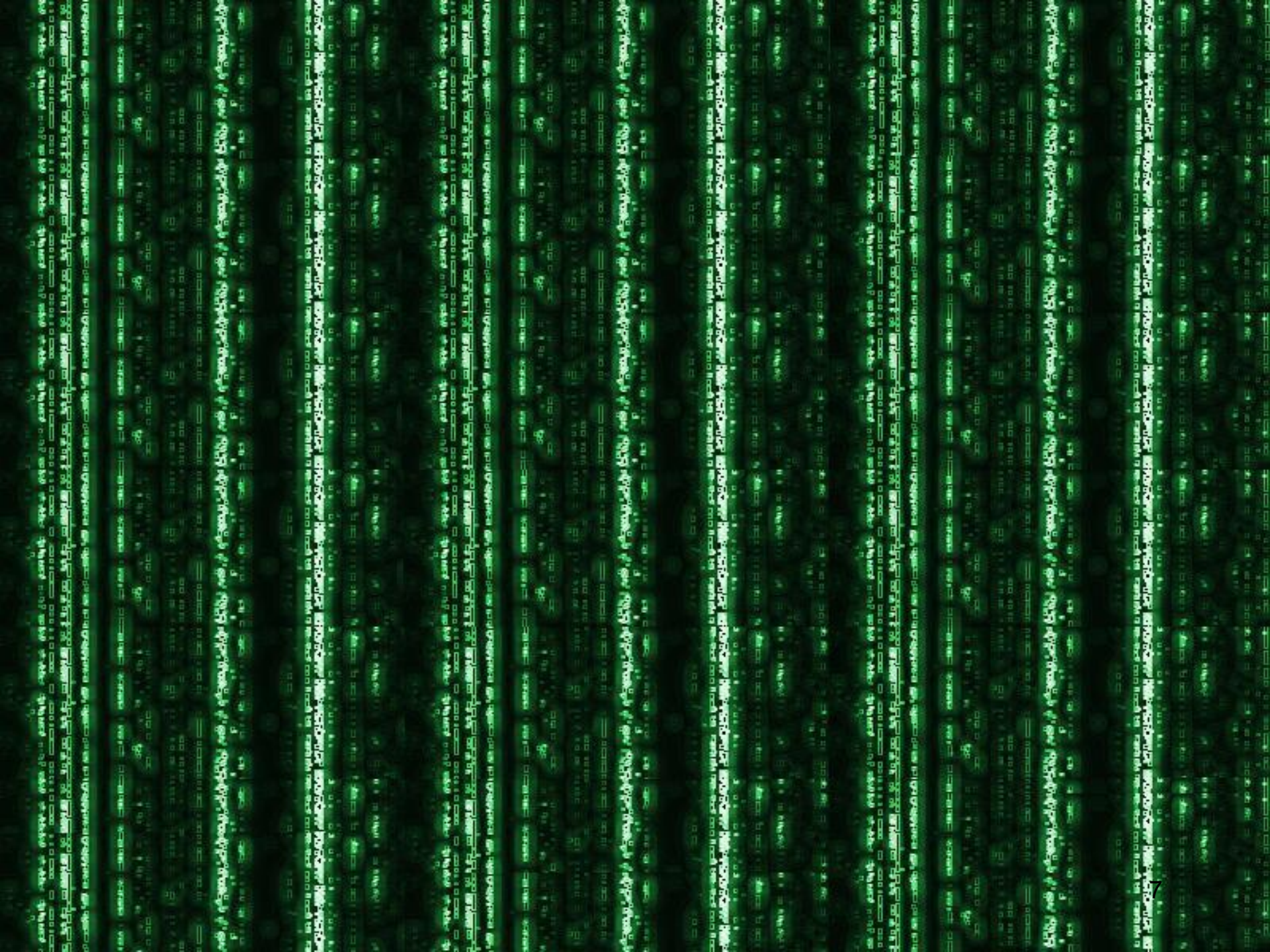
1995



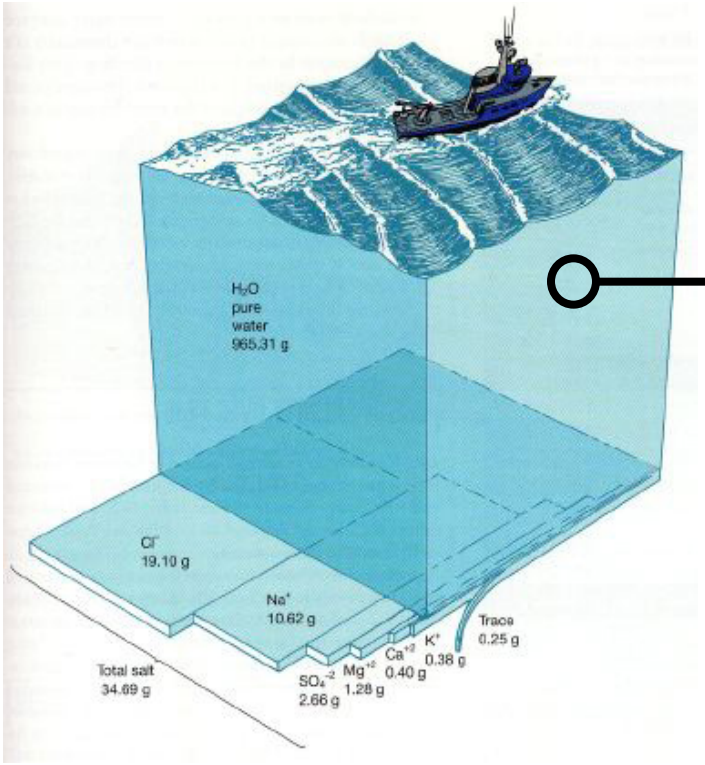
Haemophilus influenzae

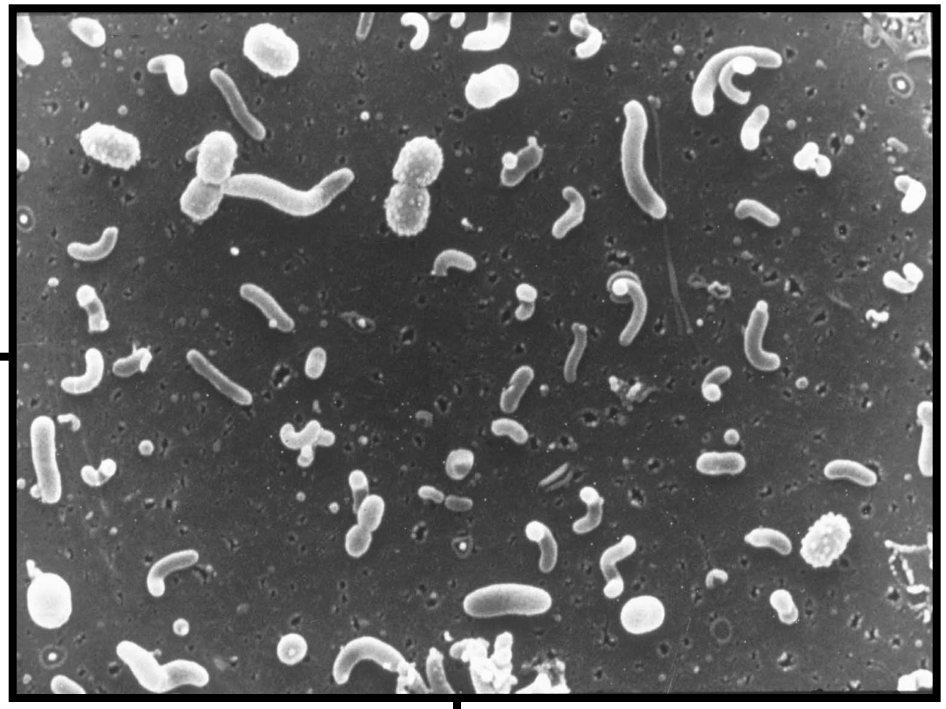
Sequencing Innovation





Like cells in a body…





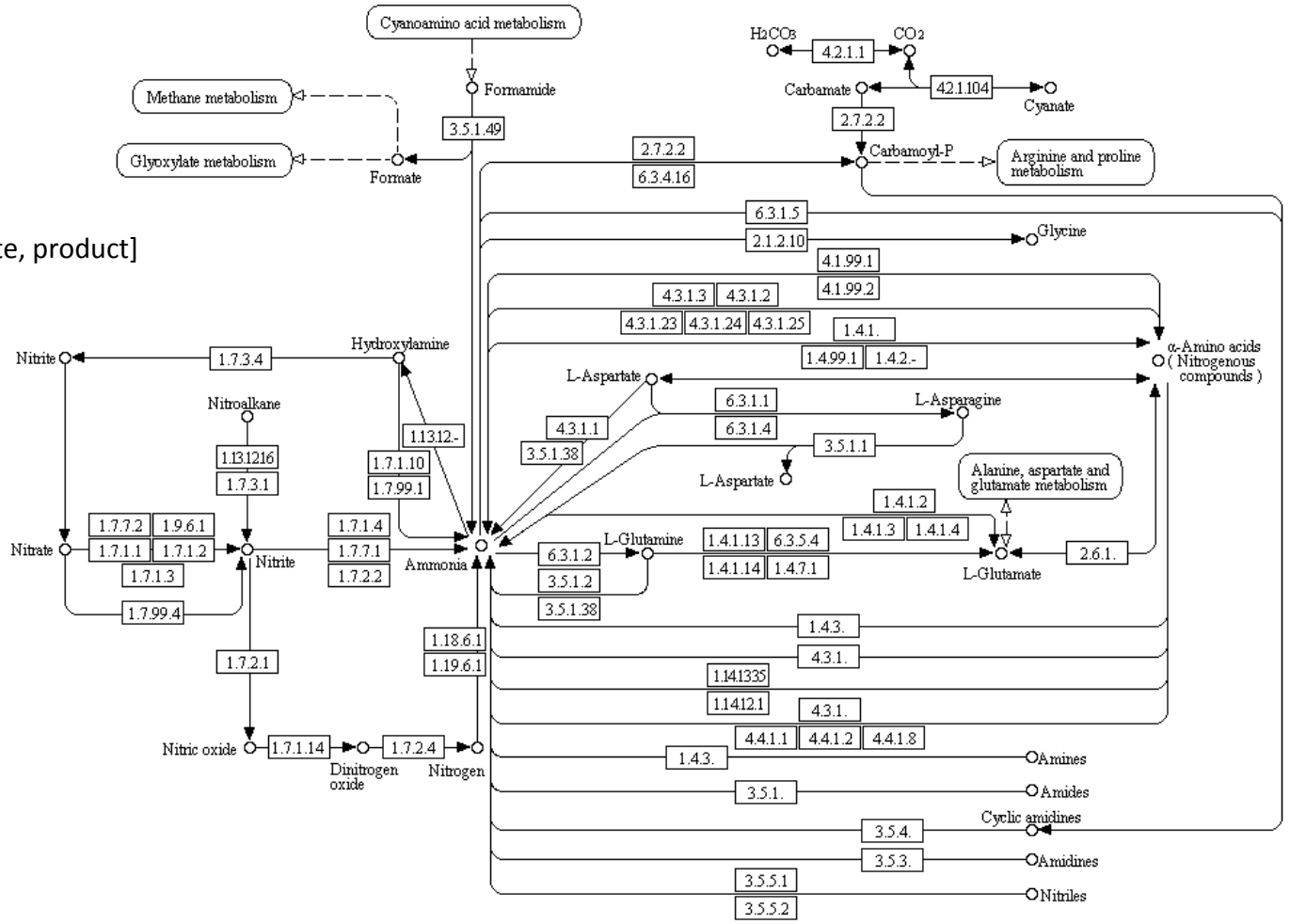
~10^6 bacterial cells per ml

- Defined by genomic diversity and metabolic potential rather than morphological diversity

Metabolism

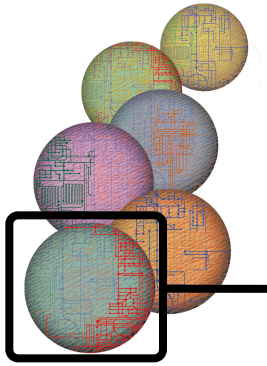
Vertex = chemical [substrate, product]

Edge = enzyme

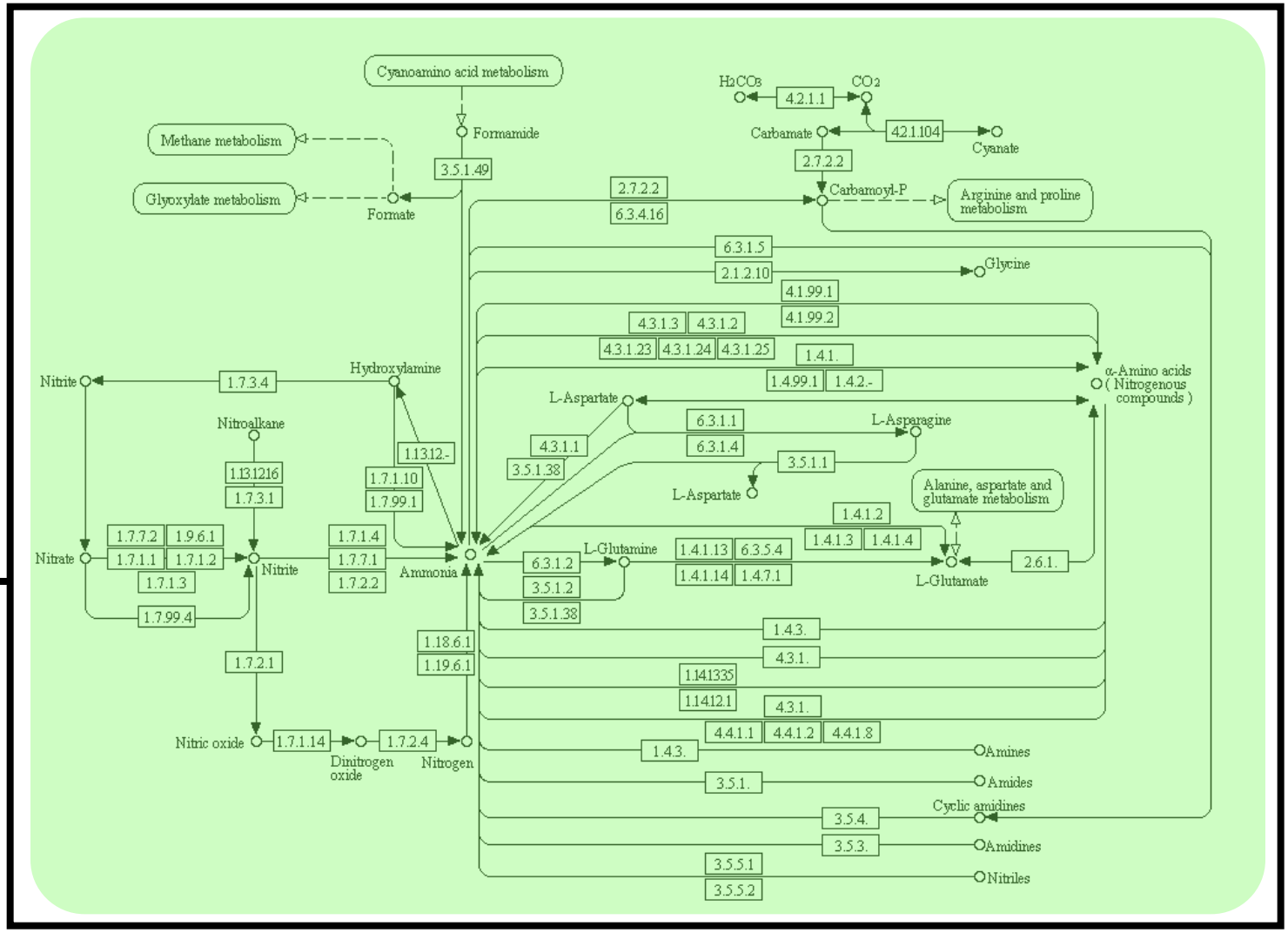


- Metabolism, or the synthesis and decomposition of chemicals in a cell can be organized into pathways represented by graphs.

Cellular Pathways

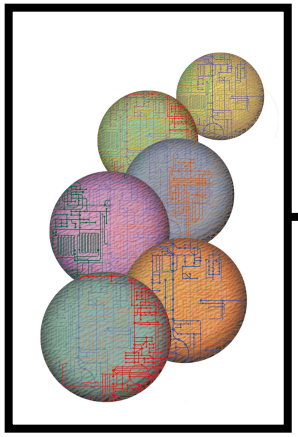


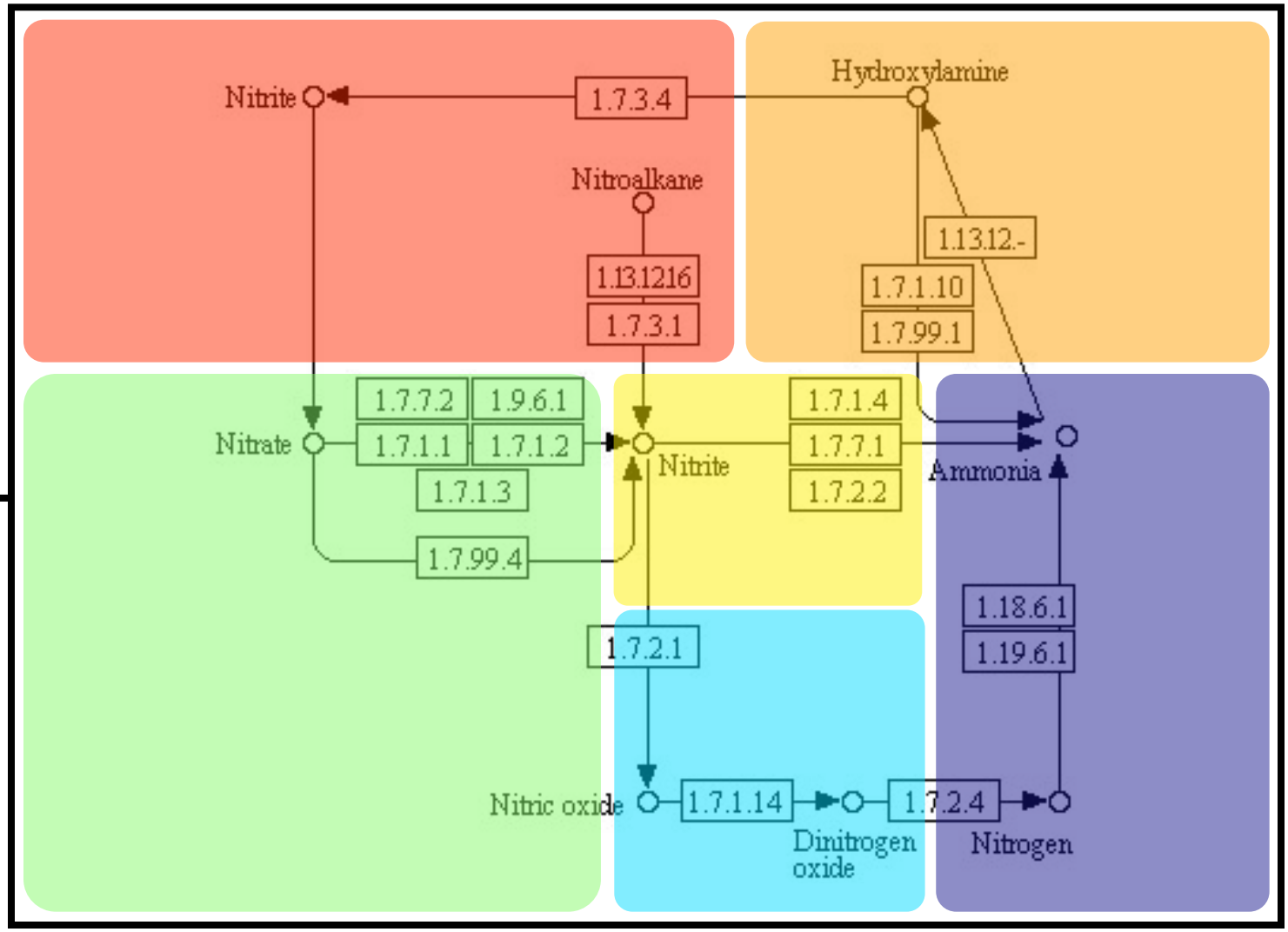
Genome Management Information System, Oak Ridge National Laboratory



- Our genetic and biochemical understanding of metabolism is based largely on the study of complete pathways within cells.

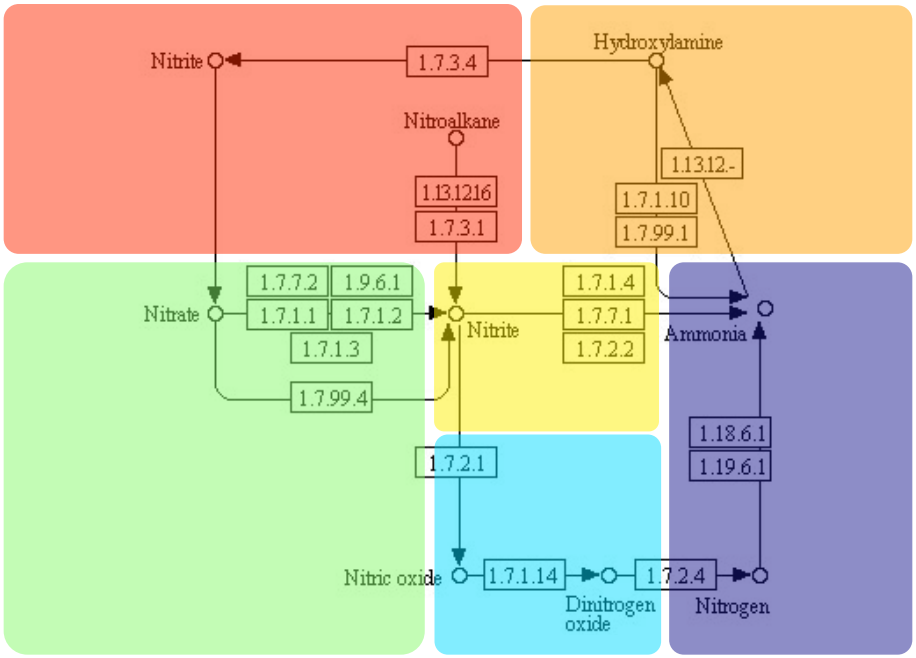
Distributed Pathways

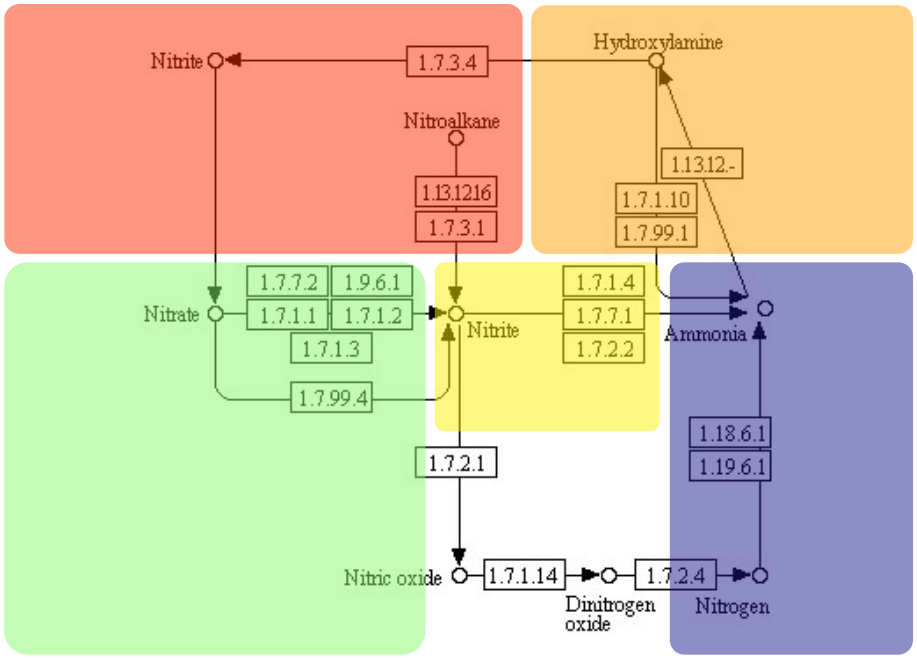




- However, microbial communities form distributed metabolic pathways directing matter and energy exchange.

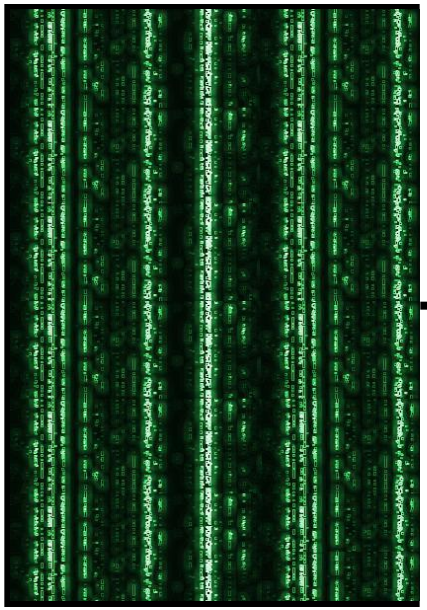
Community Metabolism

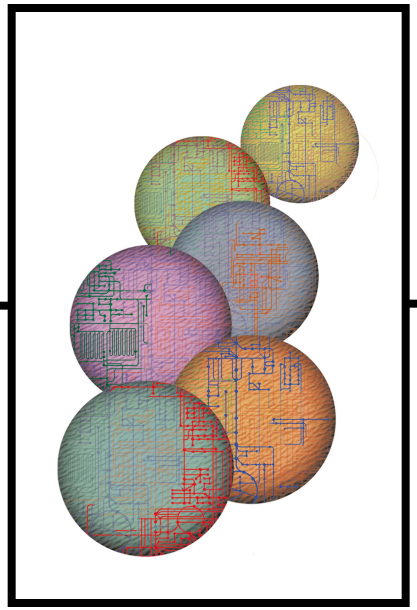


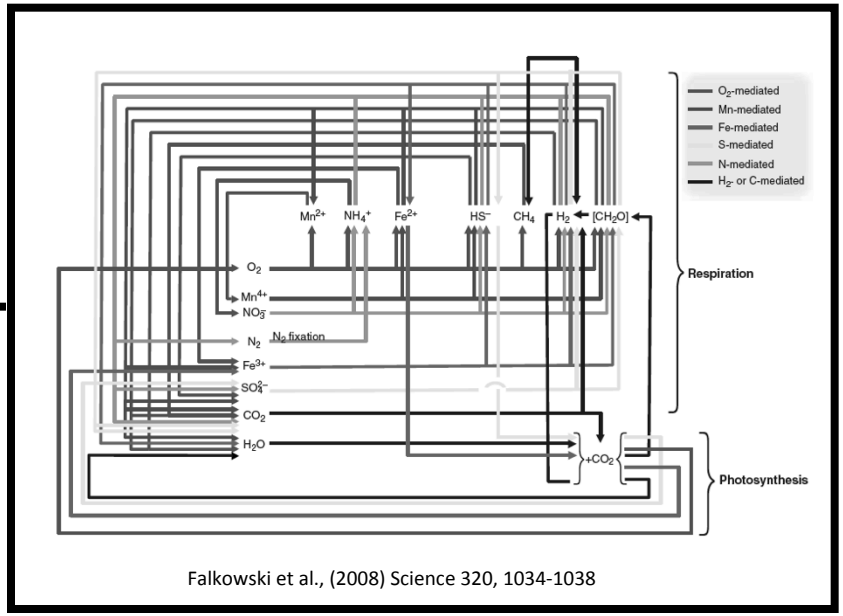


- The goal is to predict and compare distributed pathways to better understand coupled biogeochemical cycles in natural and human engineered ecosystems.

Through the Looking Glass…







Falkowski et al., (2008) Science 320, 1034-1038

Biological Information Community Metabolism Ecosystem Function

- *"The regulation of the pools and fluxes in biogeochemical cycles have their origins in the genetic inventory of individual microbes, and the regulation of these genes within the organism is determined by the environment. As such, one can look at the microbial food web as a collection of genomes whose expression and replication is coordinated through complex feedback loops at the organismal, population, and ecosystem level. "* *Chisholm*

Foundational Questions

- *What is the taxonomic and functional structure of the ecosystem?*
- *How does this structure change in response to environmental perturbation?*
- *What are the ecological consequences of this change?*
- *How can this information be used to develop biomarkers or models for monitoring industrial performance and environmental impacts?*

$$G_m = \sum_{i=1}^{l} n_i G_i$$

G_m = metagenome size in bases

l = number of genomes in sample

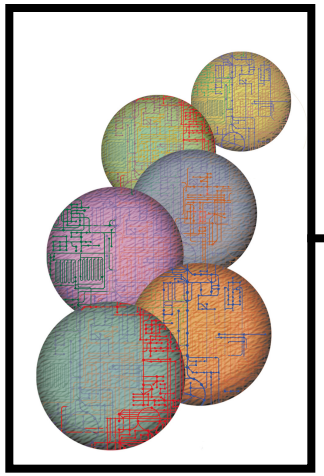
n_i = number of copies of genome G_i

G_i = size of any given genome in sample of l genomes

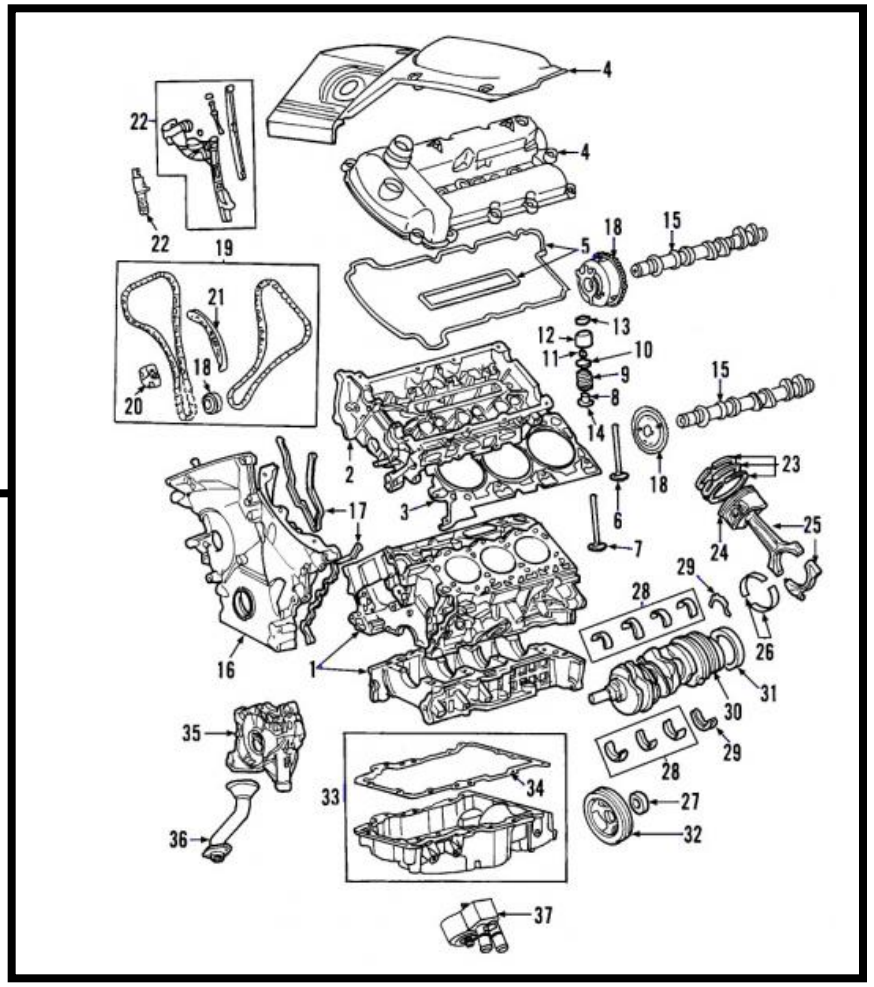
- In any given metagenome sample genotypes appear at different frequencies (evenness). Therefore a metagenome of size Gm composed of genomes of sizes G_i through G_k can be viewed as a sum of fractions where each component genome of size G_i constitutes a fraction of $G_{m:}$

$$\hat{G}_m = p_1 G_m + p_2 G_m + \ldots + p_l G_m$$

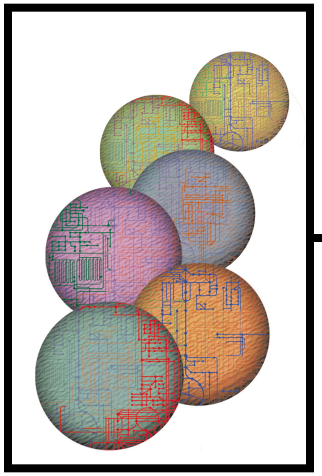
Parts



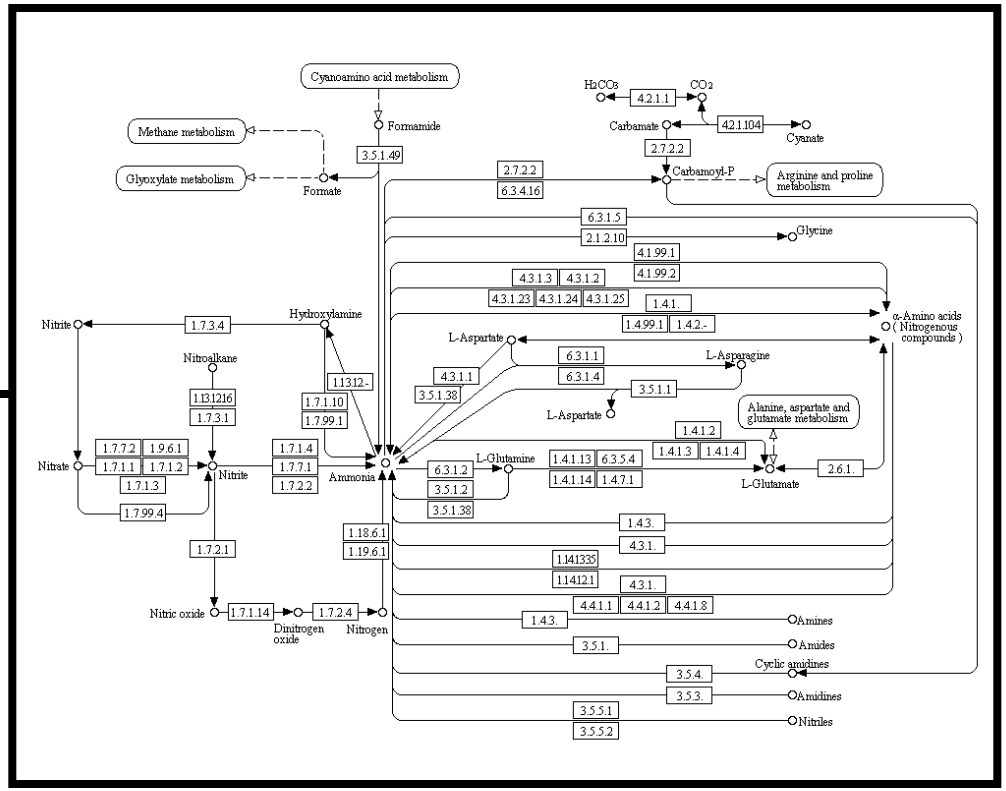
Molecular Machines



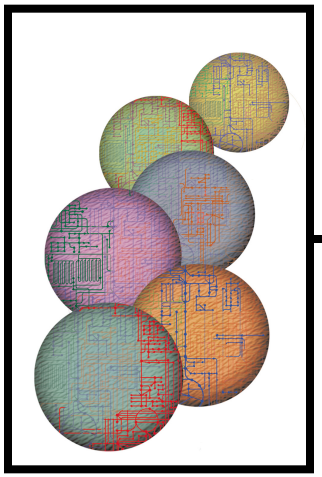
Paths



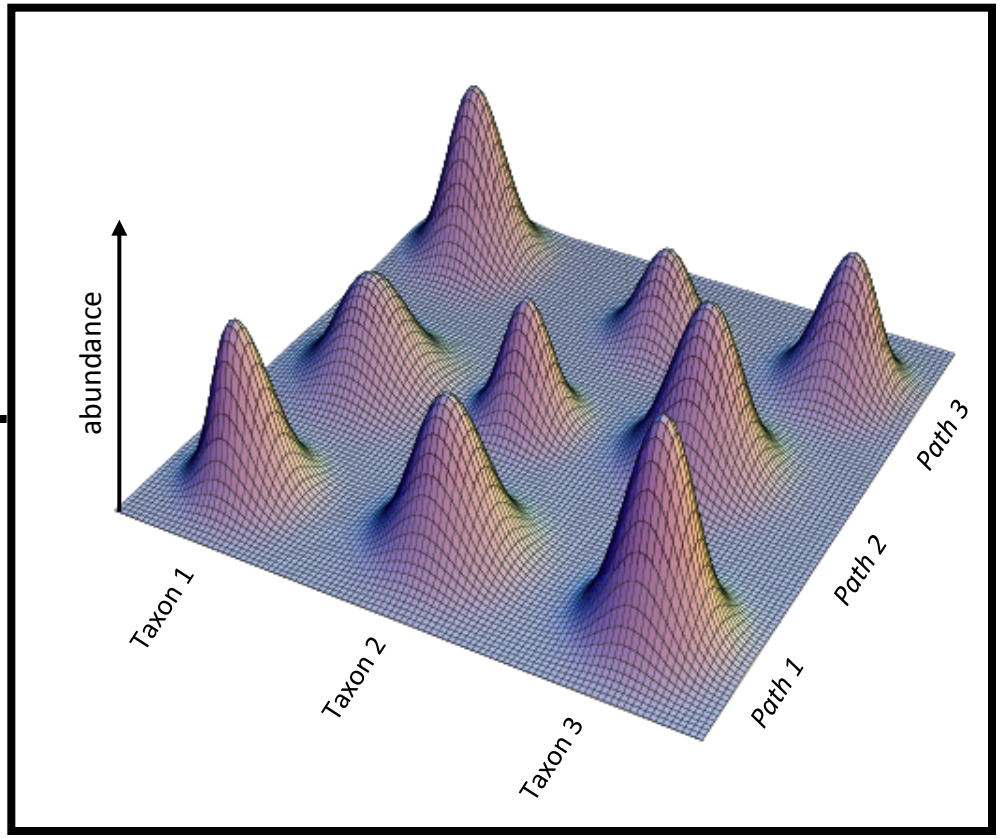
Nitrogen Metabolism



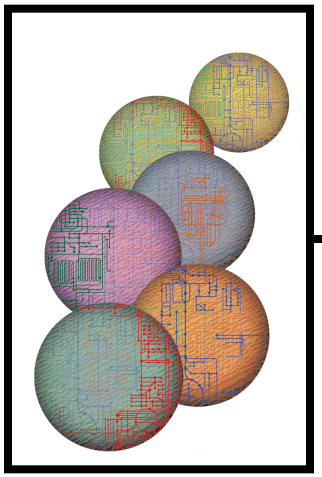
Panoramas



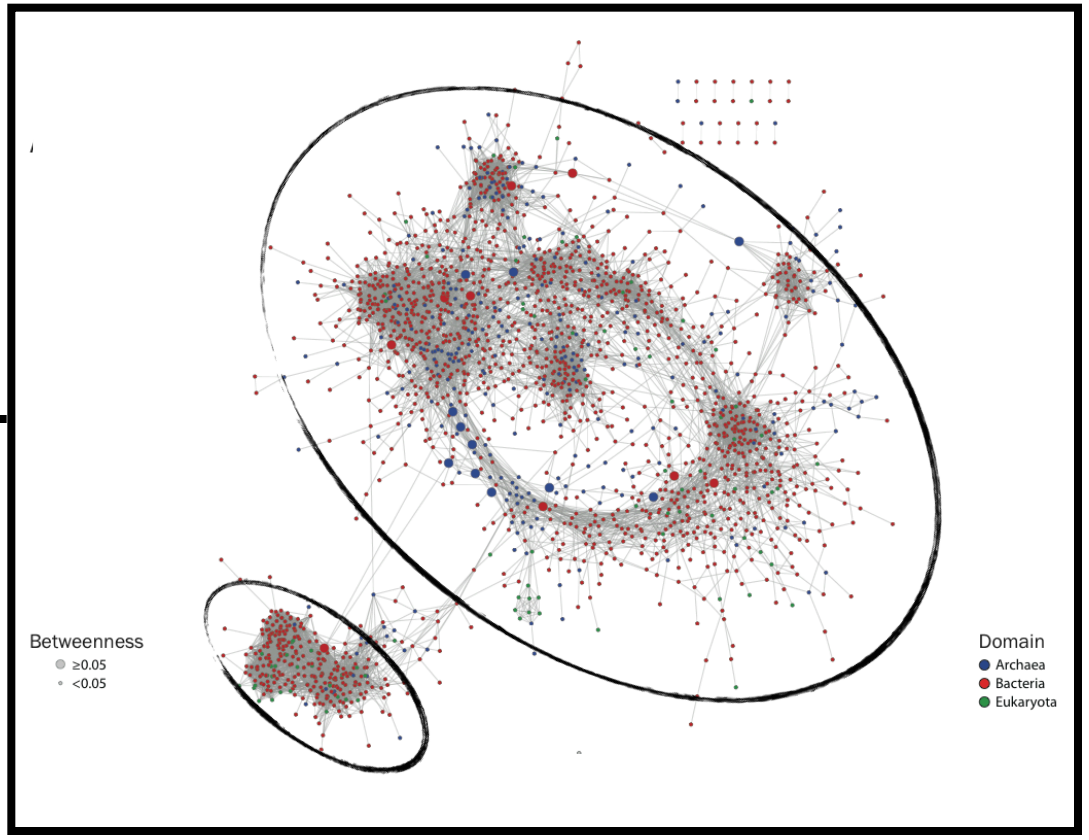
Taxonomic or
Functional Bins



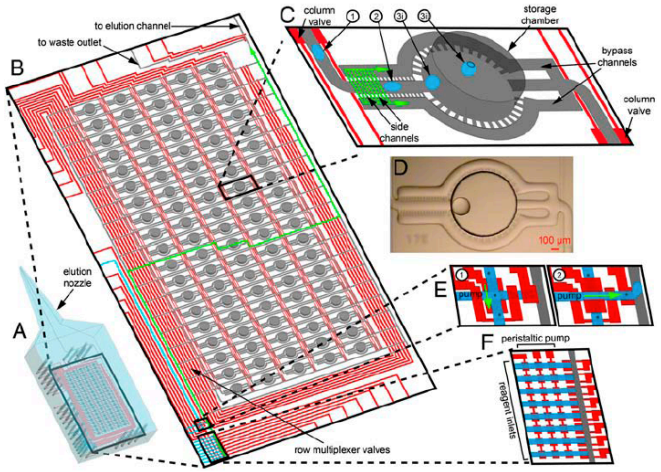
Co-occurrence Networks

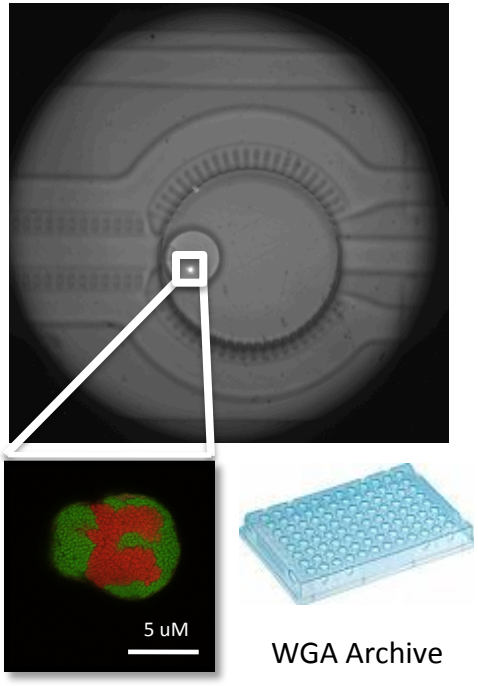


Ecological Interactions



Single-cell Genomics

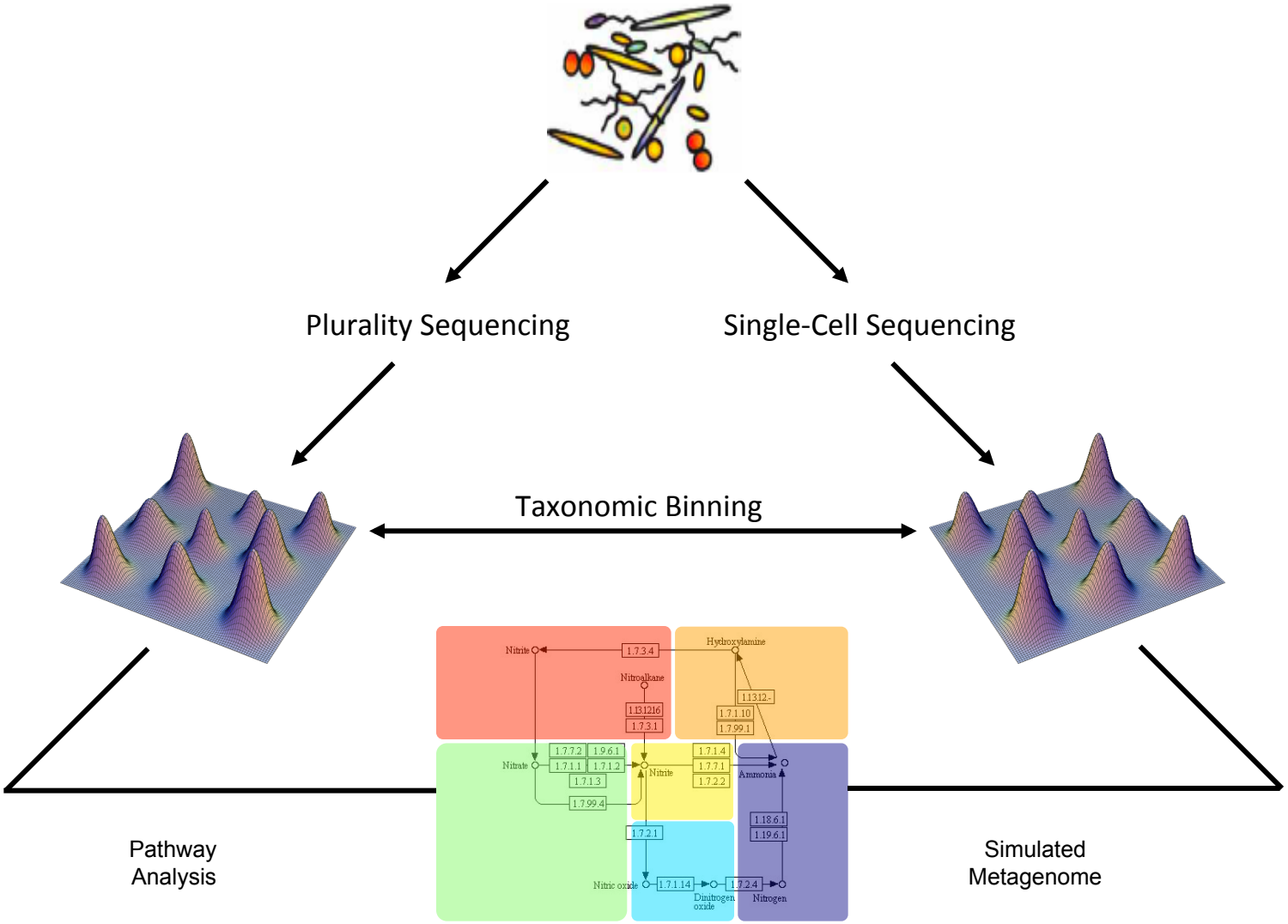




WGA Archive

:eung et al. 2012 *A programmable droplet-based microfluidic device applied to multiparameter analysis of single microbes and microbial communities*

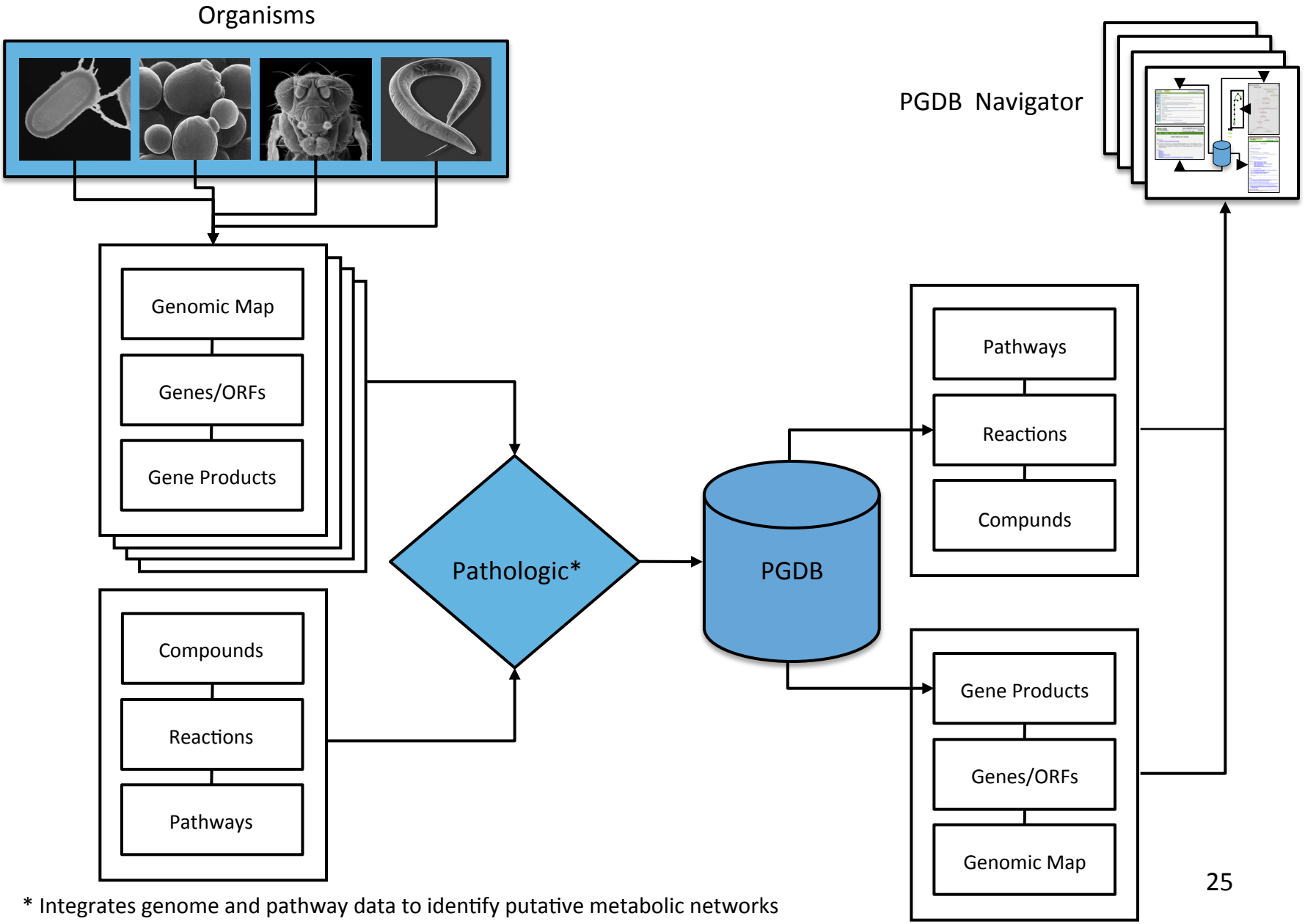
Predicting Community Metabolism



Overview

- Through the looking glass…
- MetaPathways Pipeline Development
- Hawaii Ocean Time Series

Pathway/Genome Database



* Integrates genome and pathway data to identify putative metabolic networks

http://biocyc.org/

- The BioCyc collection of Pathway/Genome Databases (PGDBs) provides an electronic reference source on the genomes and metabolic pathways of sequenced organisms.

- The BioCyc databases are divided into three tiers, based on their quality. Tier 1 databases have received person-decades of literature-based curation, and are the most accurate. Tier 2 and Tier 3 databases contain computationally predicted metabolic pathways, predictions as to which genes code for missing enzymes in metabolic pathways, and predicted operons.

Caspi, R. et al., 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases, Nucleic Acids Research , 36:D623-D631

http://metacyc.org/

- MetaCyc is a multi-organismal member of the BioCyc collection that serves as a high-quality reference DB for predicting metabolic pathways in other organisms. MetaCyc entries are based on experimental evidence cited from literature searches.

- MetaCyc does not seek to model metabolism of individual organisms, which is the role of individual BioCyc PGDBs.

- MetaCyc contains **1,846** base pathways, compared to **179** pathway modules in KEGG. Furthermore MetaCyc assigns 10,262 reactions to pathways and contains over 36,796 pathway-associated literature citations.

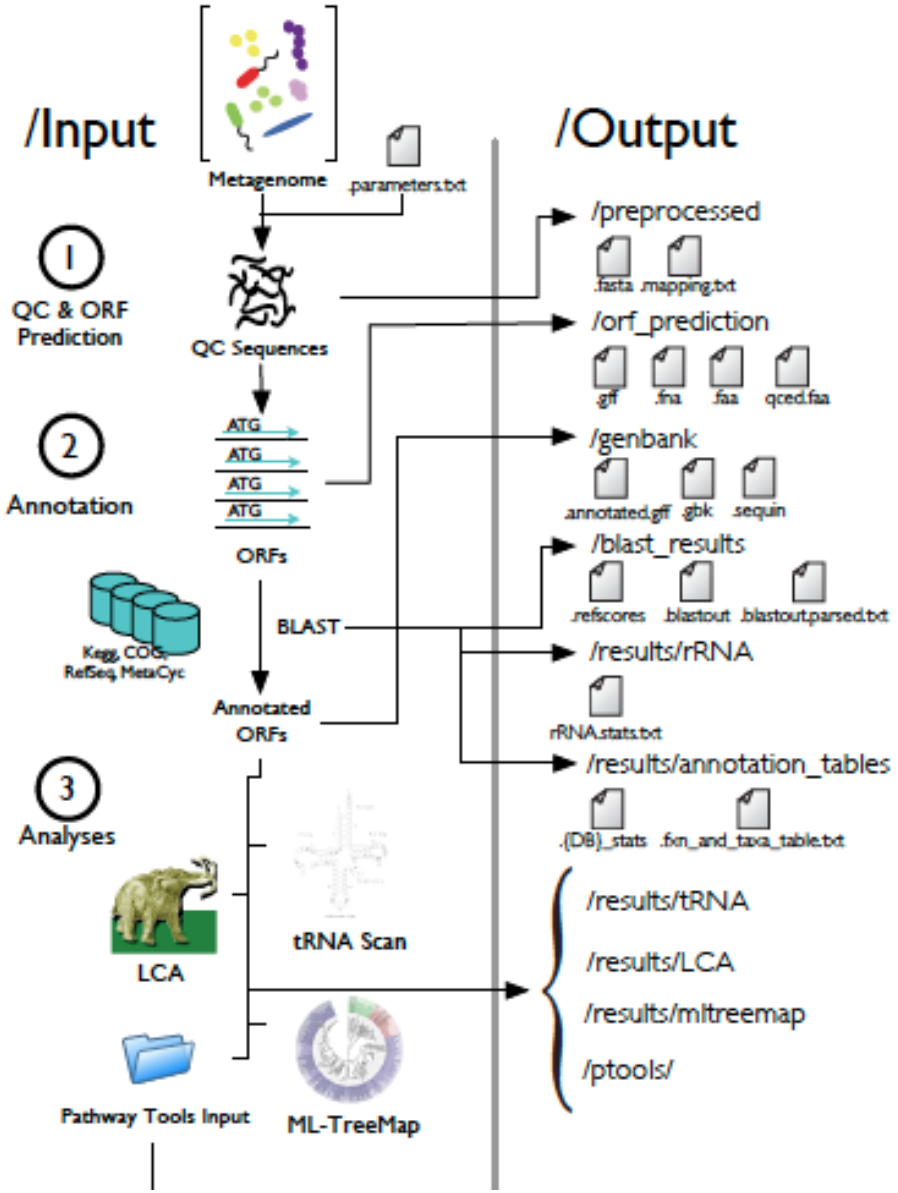
Caspi, R. et al., 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases, Nucleic Acids Research , 36:D623-D631

http://bioinformatics.ai.sri.com/ptools

- Pathway Tools
  - PathoLogic: Creates a new PGDB containing the inferred metabolic pathways of an organism, given a Genbank entry as input.
  - Pathway/Genome Navigator: Supports query, visualization, and analysis of PGDBs.
  - Pathway/Genome Editors: Provide interactive editing capabilities for PGDBs.
  - MetaFlux: Supports flux balance analysis (FBA) modeling.

Karp, P.D. and Paley, S. and Romero, P.2002. The Pathway Tools Software. Bioinformatics, vol. 18, pp. S225-S232.

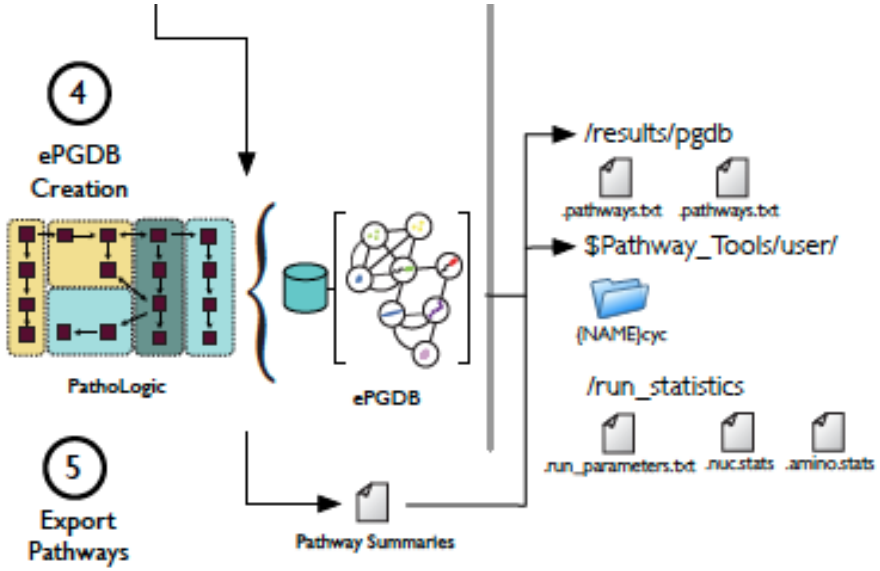
MetaPathways



- A modular pipeline for constructing Pathway/Genome Databases (PGDBs) from environmental sequence information
- MetaPathways currently supports four "data products" including i) GenBank submission, ii) LCA, iii) MLTreeMap, and iv) PGDBs with associated feature summary tables and GFF files
- MetaPathways externalizes compute-intensive processes onto a user defined cluster using Sun Grid Engine or the Amazon elastic cloud

Konwar, K. M., N. W. Haetal., 2013. MetaPathways: A modular pipeline for constructing Pathway/Genome Databases from environmental sequence information. BMC Bioinformatics 21;14(1):202.

Environmental PGBDs



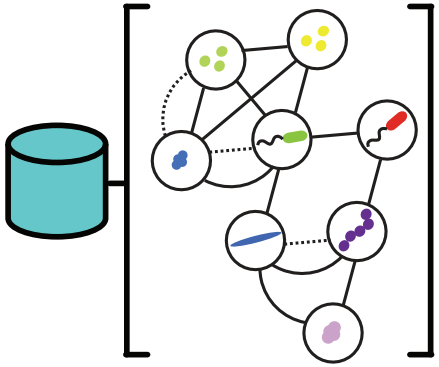
- ePGDBs facilitate pathway-centric exploration of environmental sequence information using Pathway Tools and the MetaCyc web interface

- Provides inference-based approach to metabolic reconstruction based on explicit computational rules to predict presence or absence of distributed metabolic networks

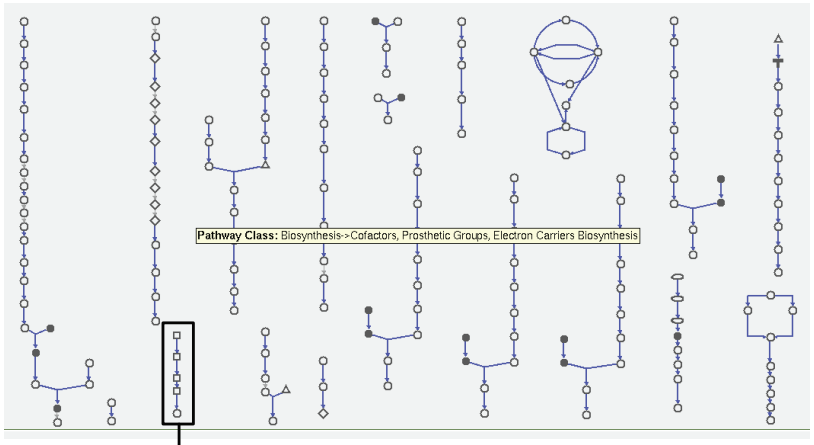
- MetaPathways can be used with multi-molecular data sets (DNA, RNA or protein) sourced from cultured isolates, single-cells and natural or human engineered ecosystems

http://www.github.com/hallamlab/MetaPathways
http://hallam.microbiology.ubc.ca/MetaPathways

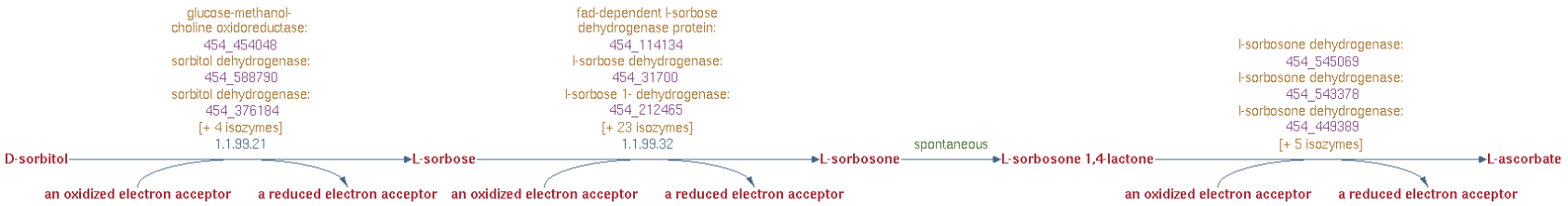
ePGDB Navigation



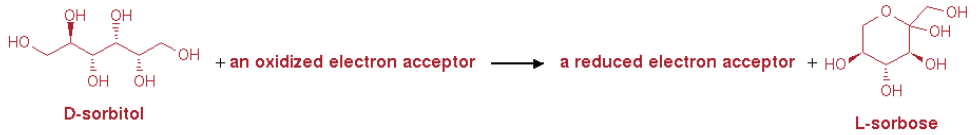
ePGDB



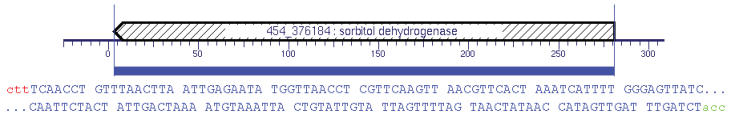
Celllar Overview



Metabolic Pathway

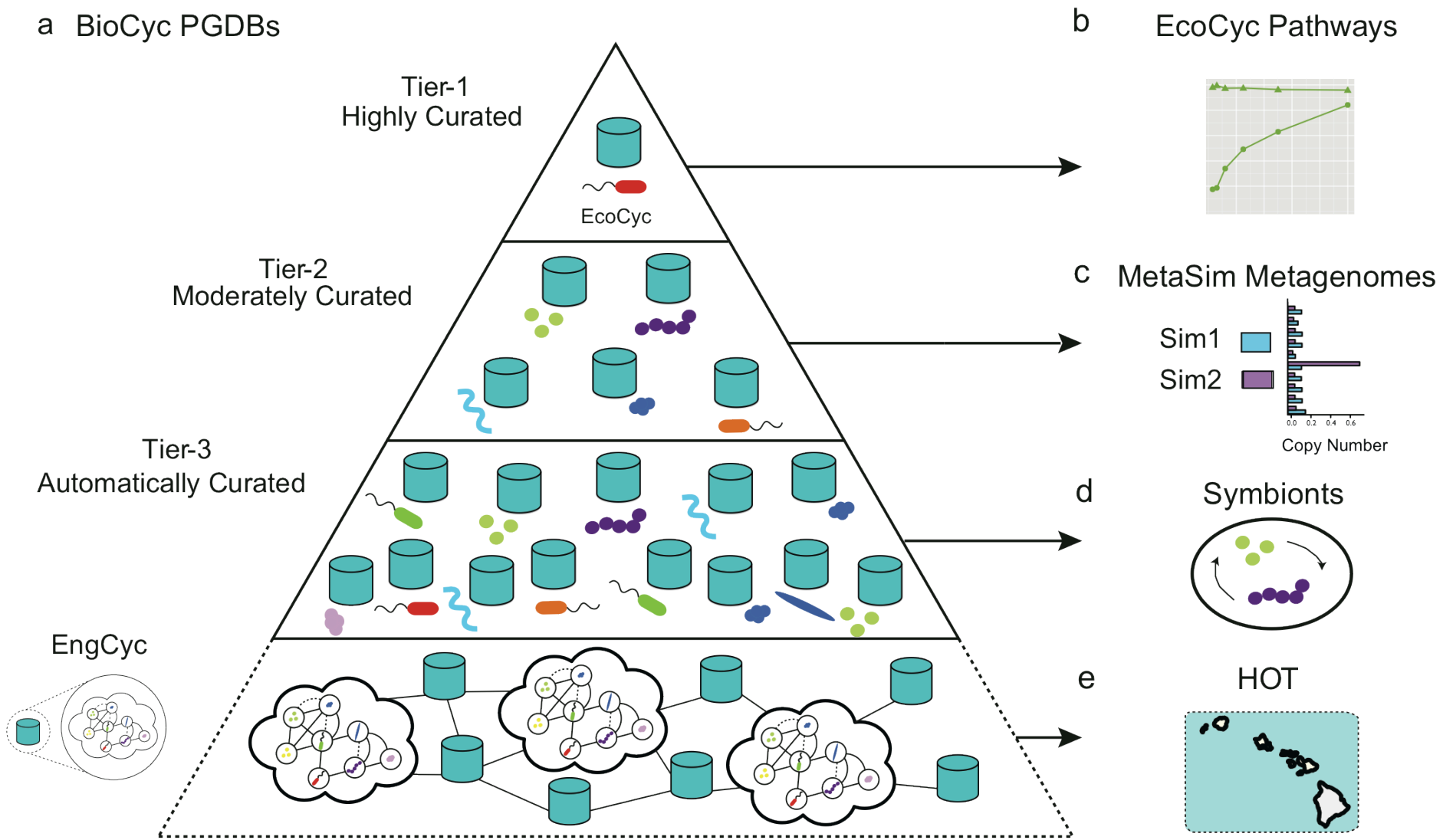


Reaction

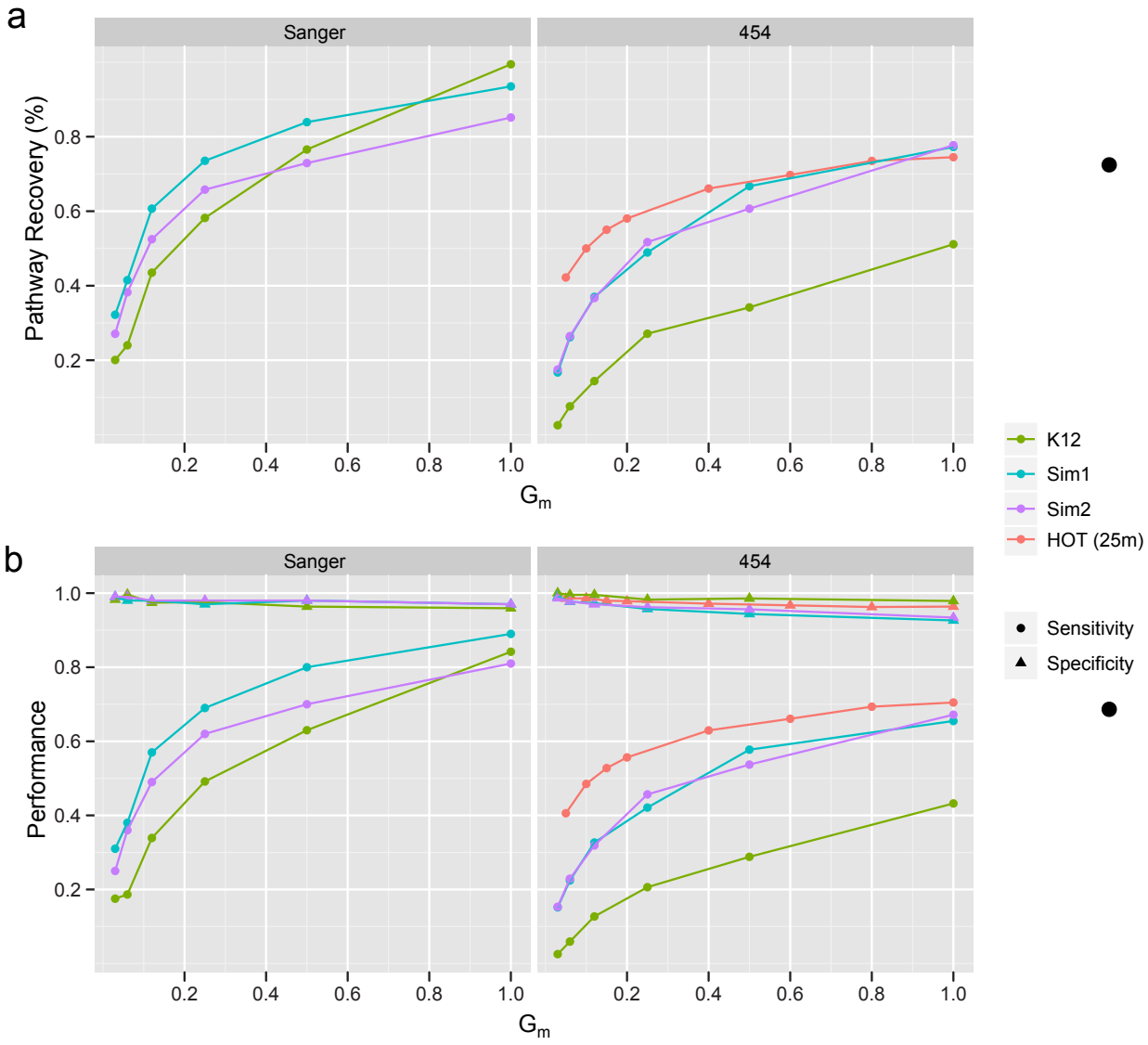


Open Reading Frame

ePGDB Validation



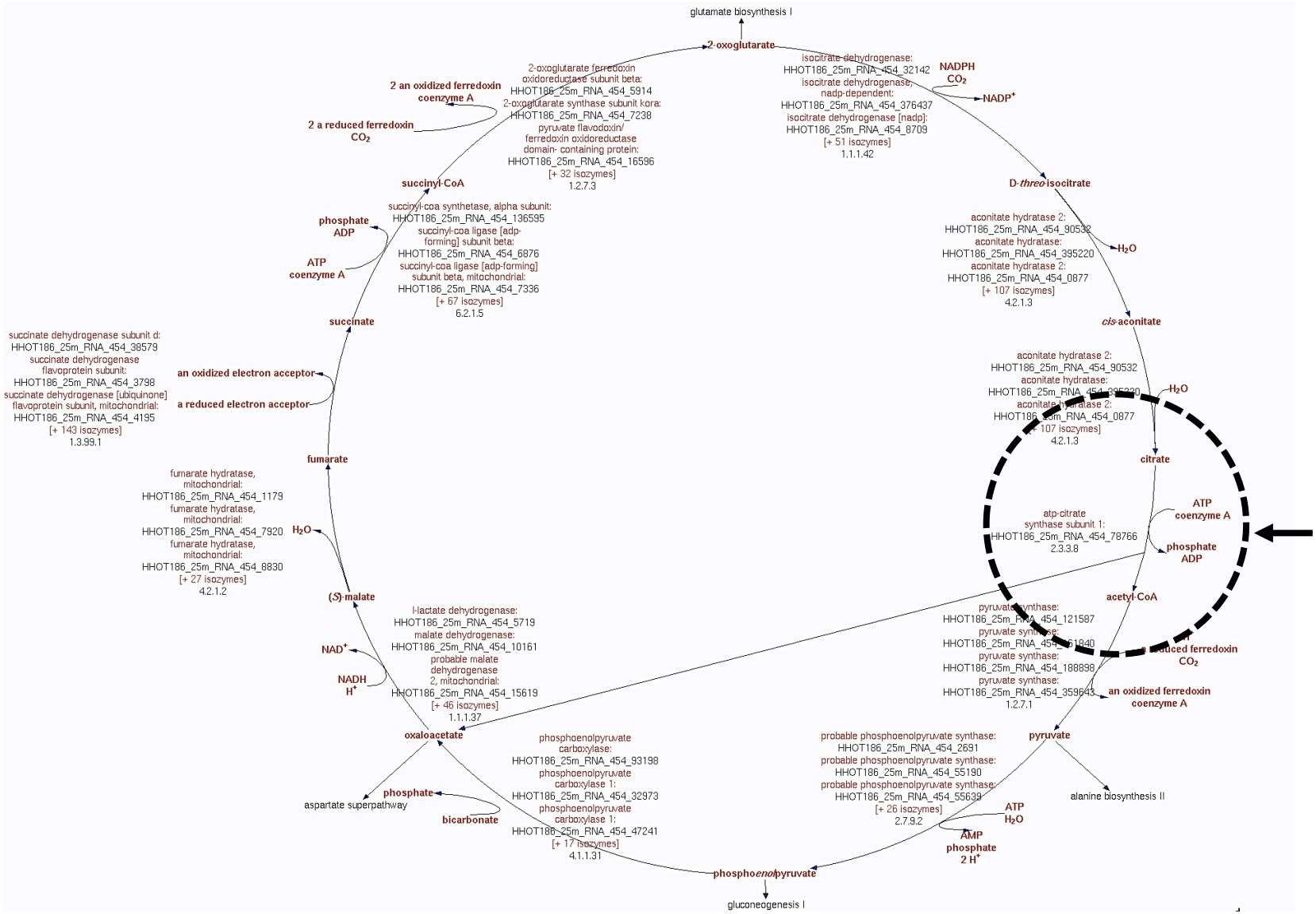
Simulated Metagenomes



- Pathway recovery increases as a function of read length and sample coverage

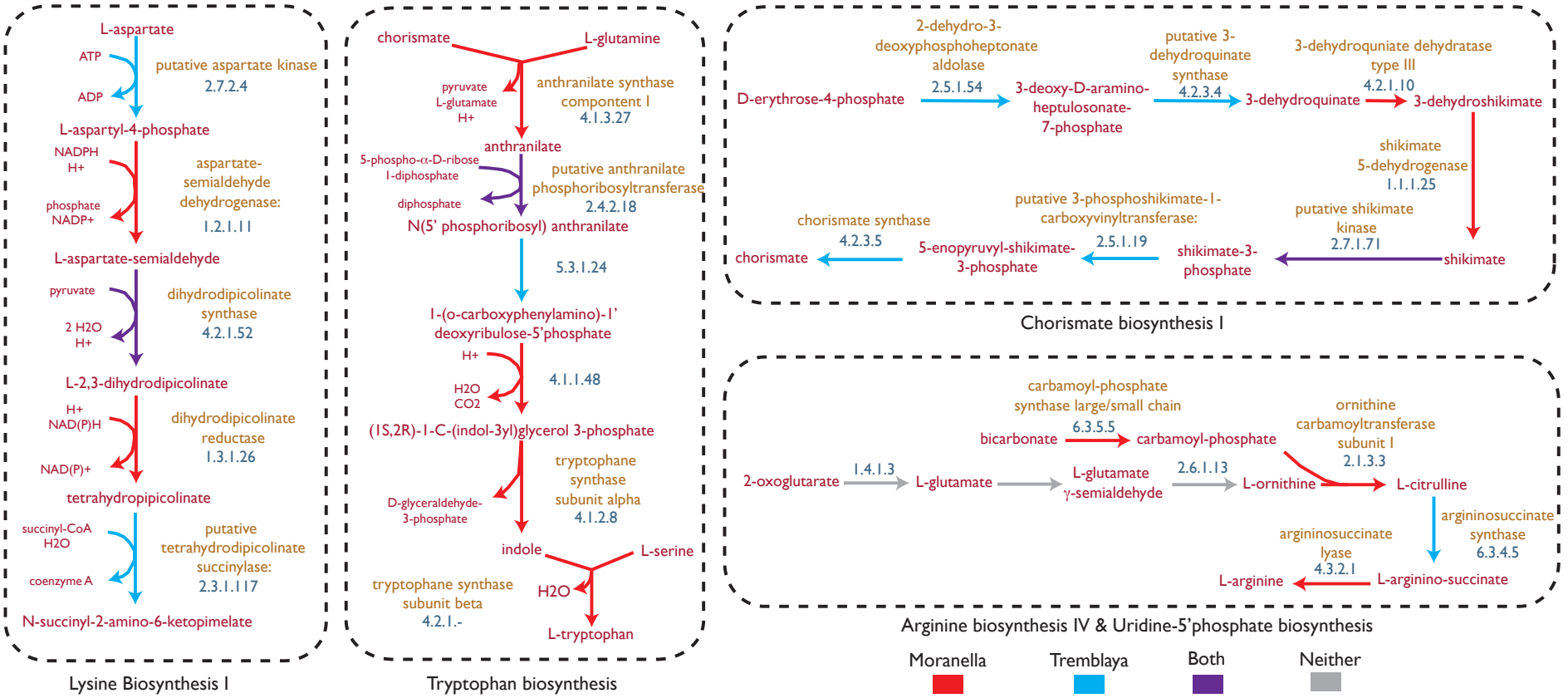
- Specificity remains high irrespective of read length or sample coverage, but sensitivity increases with both parameters

Discerning Hazards



- Missing ATP citrate lyase indicates false positive for rTCA

Symbiotic Systems



- An ePGDB constructed for the Mealybug symbionts *Tremblaya princeps* and *Moranella endobia* predicted interpathway complementarity in essential amino acid biosynthetic pathways

McCutcheon, J.P. and von Dohlen, C.D. 2011. "An interdependent metabolic patchwork in the nested symbiosis of mealybugs." Current Biology, doi: 10.1016/j.cub.2011.06.051

Overview

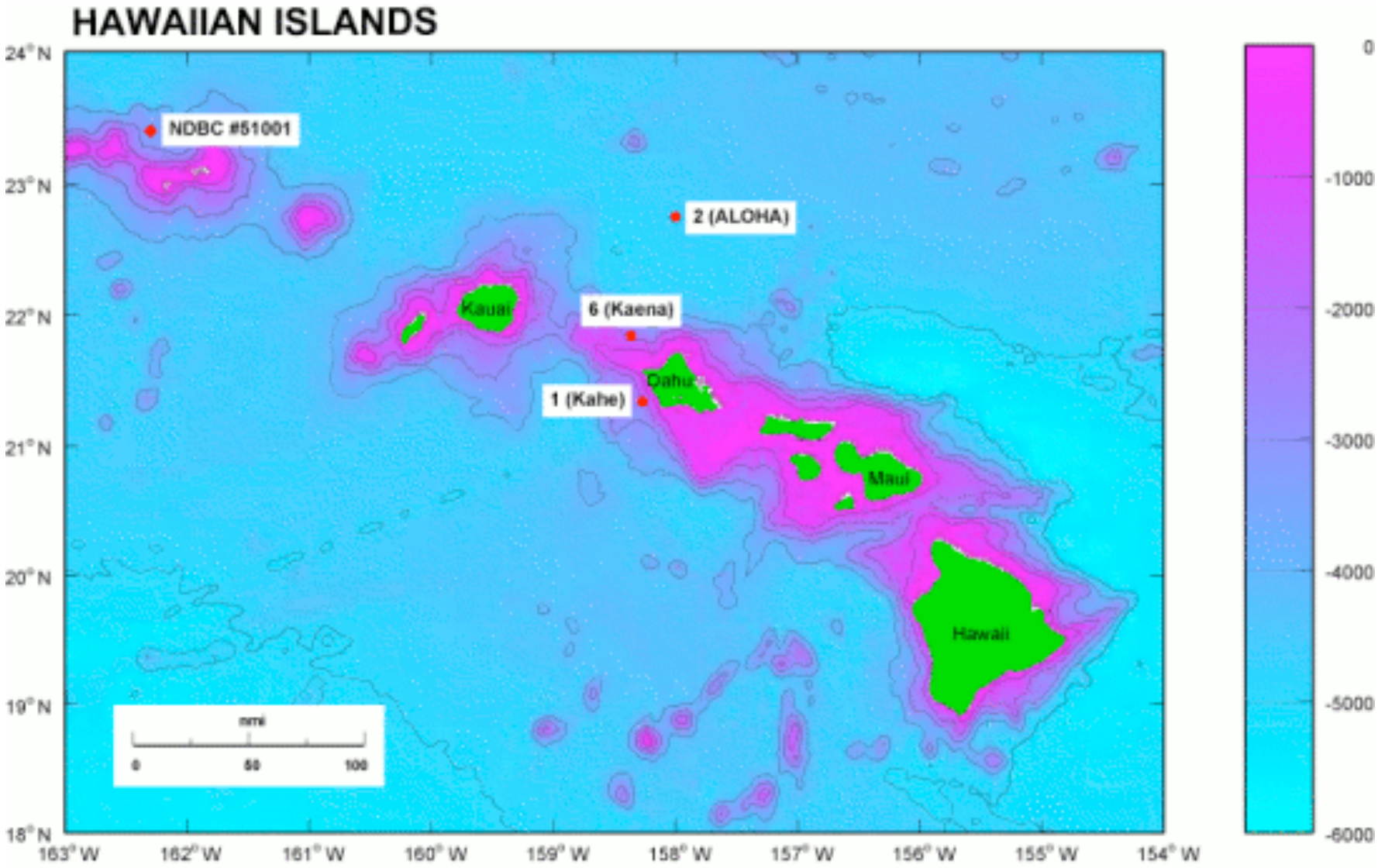
- Through the looking glass…
- MetaPathways Pipeline Development
- Hawaii Ocean Time Series

Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior

Edward F. DeLong,[1*] Christina M. Preston,[2] Tracy Mincer,[1] Virginia Rich,[1] Steven J. Hallam,[1] Niels-Ulrik Frigaard,[1] Asuncion Martinez,[1] Matthew B. Sullivan,[1] Robert Edwards,[3] Beltran Rodriguez Brito,[3] Sallie W. Chisholm,[1] David M. Karl[4]

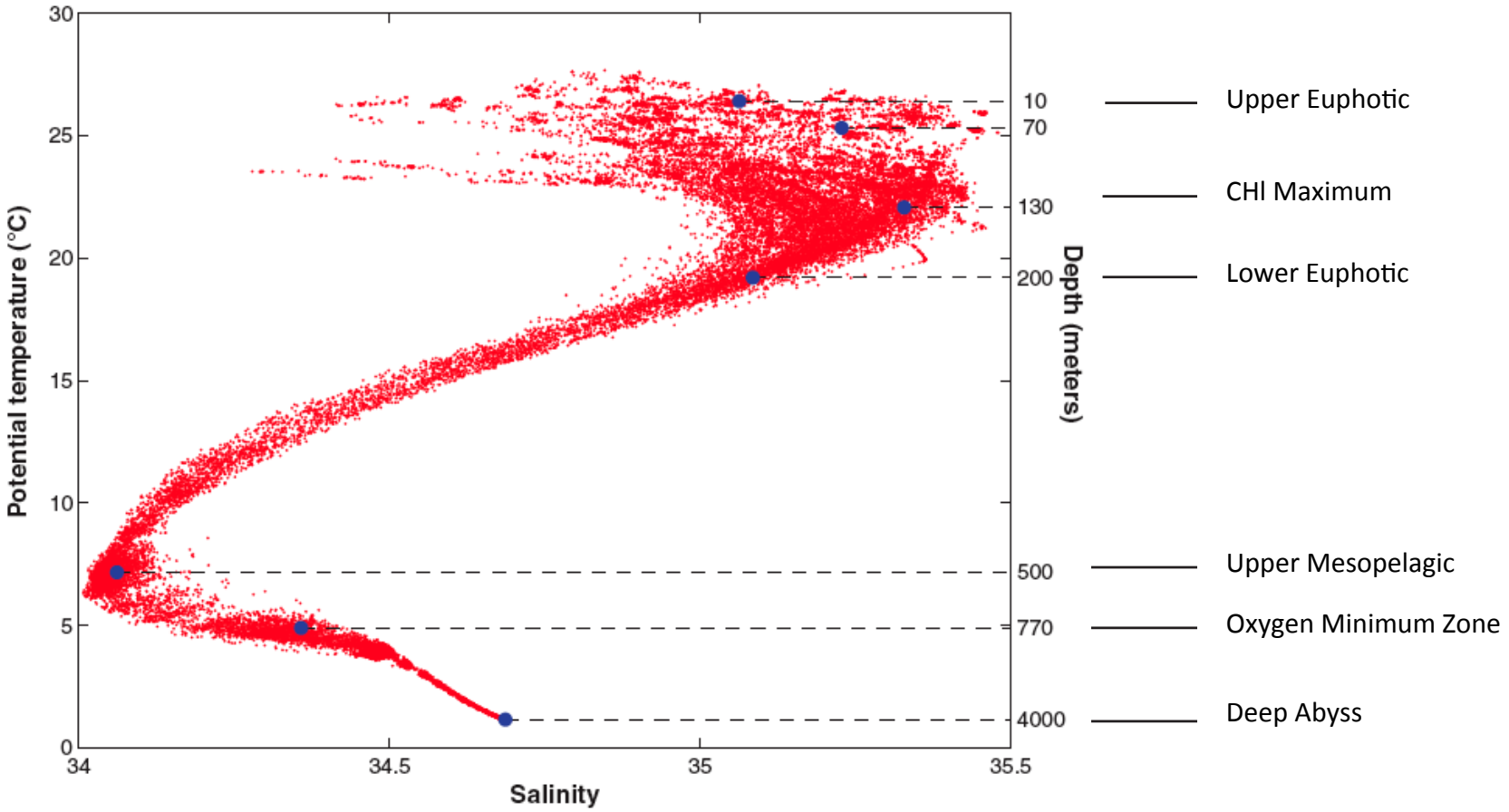
Microbial life predominates in the ocean, yet little is known about its genomic variability, especially along the depth continuum. We report here genomic analyses of planktonic microbial communities in the North Pacific Subtropical Gyre, from the ocean's surface to near–sea floor depths. Sequence variation in microbial community genes reflected vertical zonation of taxonomic groups, functional gene repertoires, and metabolic potential. The distributional patterns of microbial genes suggested depth-variable community trends in carbon and energy metabolism, attachment and motility, gene mobility, and host-viral interactions. Comparative genomic analyses of stratified microbial communities have the potential to provide significant insight into higher-order community organization and dynamics.

Hawaii Ocean Time Series (HOT)



http://hahana.soest.hawaii.edu/hot/hot_jgofs.html

Vertical Transect



DeLong et al. 2006. Community Genomics Among Stratified Assemblages in the Ocean's Interior. *Science 311*

c1988-2013

Environmental Parameters

Depth (m)	Temp. (°C)	Salinity	Chl a (μg/kg)	Biomass* (μg/kg)	DOC (μmol/kg)	N + N (nmol/kg)	DIP (nmol/kg)	Oxygen (μmol/kg)	DIC (μmol/kg)
10	26.40	35.08	0.08	7.21 ± 2.68	78	1.0	41.0	204.6	1,967.6
	(24.83 ± 1.27)	(35.05 ± 0.21)	(0.08 ± 0.03)	[78]	(90.6 ± 14.3)	(2.6 ± 3.7)	(56.0 ± 33.7)	(209.3 ± 4.5)	(1,972.1 ± 16.4)
	[2,104]	[1,611]	[320]		[140]	[126]	[146]	[348]	[107]
70	24.93	35.21	0.18	8.51 ± 3.22	79	1.3	16.0	217.4	1,981.8
	(23.58 ± 1.00)	(35.17 ± 0.16)	(0.15 ± 0.05)	[86]	(81.4 ± 11.3)	(14.7 ± 60.3)	(43.1 ± 25.1)	(215.8 ± 5.4)	(1,986.9 ± 15.4)
	[1,202]	[1,084]	[363]		[79]	[78]	[104]	[144]	[84]
130	22.19	35.31	0.10	5.03 ± 2.30	69	284.8	66.2	204.9	2,026.5
	(21.37 ± 0.96)	(35.20 ± 0.10)	(0.15 ± 0.06)	[90]	(75.2 ± 9.1)	(282.9 ± 270.2)	(106.0 ± 49.7)	(206.6 ± 6.2)	(2,013.4 ± 13.4)
	[1,139]	[980]	[350]		[86]	[78]	[68]	[173]	[69]
200	18.53	35.04	0.02	1.66 ± 0.24	63	1,161.9 ± 762.5	274.2 ± 109.1	198.8	2,047.7
	(18.39 ± 1.29)	(34.96 ± 0.18)	(0.02 ± 0.02)	[2]	(64.0 ± 9.8)	[7]	[84]	(197.6 ± 7.1)	(2,042.8 ± 10.5)
	[662]	[576]	[97]		[113]			[190]	[125]
500	7.25	34.07	ND	0.48 ± 0.23	47	28,850	2,153	118.0	2197.3
	(7.22 ± 0.44)	(34.06 ± 0.03)		[107]	(47.8 ± 6.3)	(28,460 ± 2210)	(2,051 ± 175.7)	(120.5 ± 18.3)	(2,200.2 ± 17.8)
	[1,969]	[1,769]			[112]	[326]	[322]	[505]	[134]
770	4.78	34.32	ND	0.29 ± 0.16	39.9	41,890	3,070	32.3	2323.8
	(4.86 ± 0.21)	(34.32 ± 0.04)		[107]	(41.5 ± 4.4)	(40,940 ± 500)	(3,000 ± 47.1)	(27.9 ± 4.1)	(2,324.3 ± 6.1)
	[888]	[773]			[34]	[137]	[135]	[275]	[34]
4,000	1.46	34.69	ND	ND	37.5	36,560	2,558	147.8	2325.5
	(1.46 ± 0.01)	(34.69 ± 0.00)			(42.3 ± 4.9)	(35,970 ± 290)	(2,507 ± 19)	(147.8 ± 1.3)	(2,329.1 ± 4.8)
	[262]	[245]			[83]	[108]	[104]	[210]	[28]

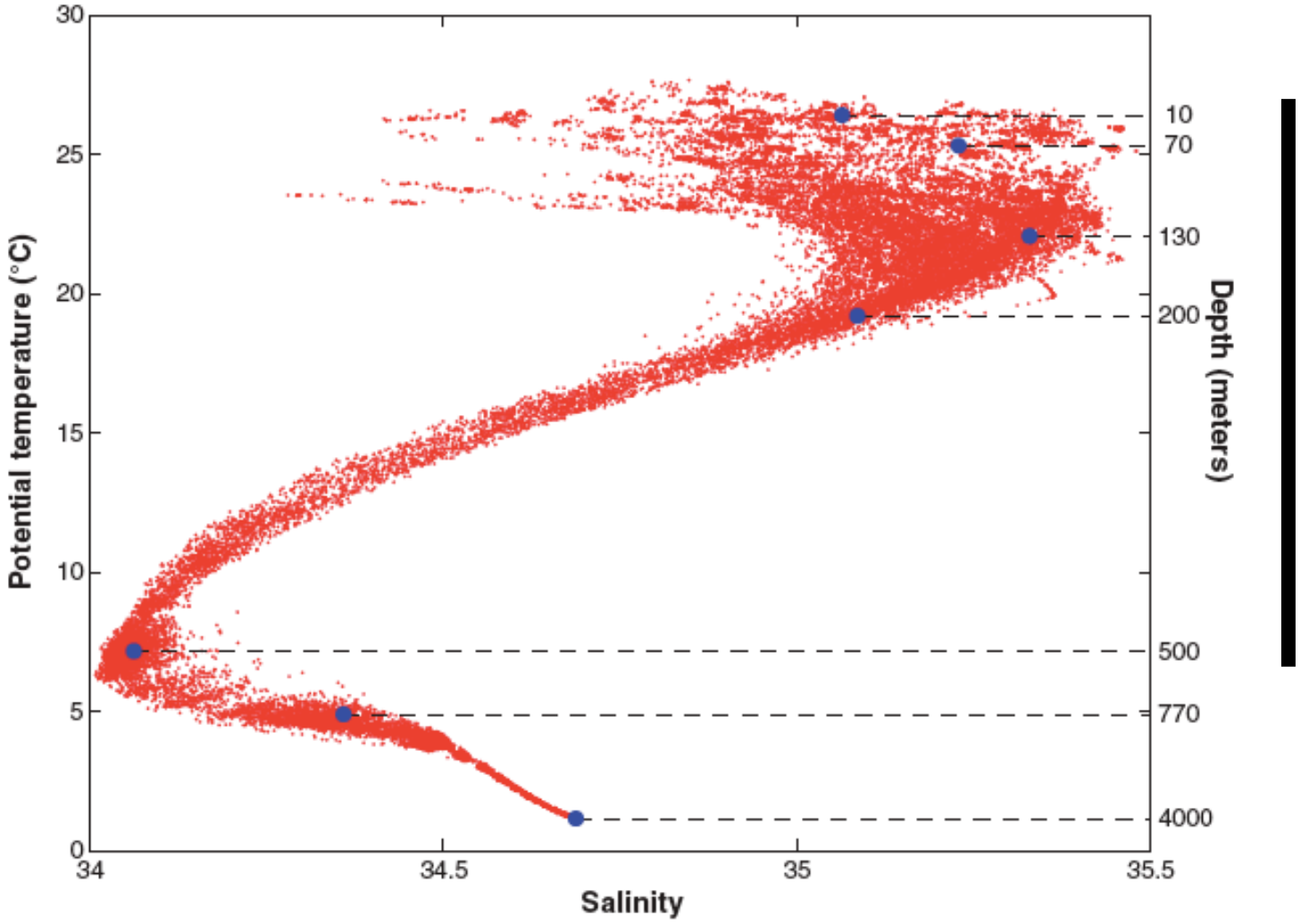
- *Oceanographic parameters were measured at Station ALOHA; values shown are those from the same CTD casts used in fosmid library construction, where available.*

Sequencing Statistics

Depth (m)	Sample date	Volume filtered (liters)	Total fosmid clones	Total DNA (Mbp)	
				Archived	Sequenced
10	10/7/02	40	12,288	442	7.54
70	10/7/02	40	12,672	456	11.03
130	10/6/02	40	13,536	487	6.28
200	10/6/02	40	19,008	684	7.96
500	10/6/02	80	15,264	550	8.86
770	12/21/03	240	11,520	415	11.18
4,000	12/21/03	670	41,472	1,493	11.10

- *"The vertical distribution of microbial genes from the ocean's surface to abyssal depths was determined by shotgun sequencing of fosmid clone termini. Applying identical collection, cloning, and sequencing strategies at seven depths (ranging from 10 m to 4000 m), we archived large-insert genomic libraries from each depth-stratified microbial community."*

Coupled “Omic” Information



DeLong et al. 2006. Community Genomics Among Stratified Assemblages in the Ocean’s Interior. *Science 311*
T. Danhorn, C. R. Young, E. F. Delong, 2012. Comparison of large-insert, small-insert and pyrosequencing libraries for metagenomic analysis, ISME JJ doi:10.1038/ismej.2012.35.

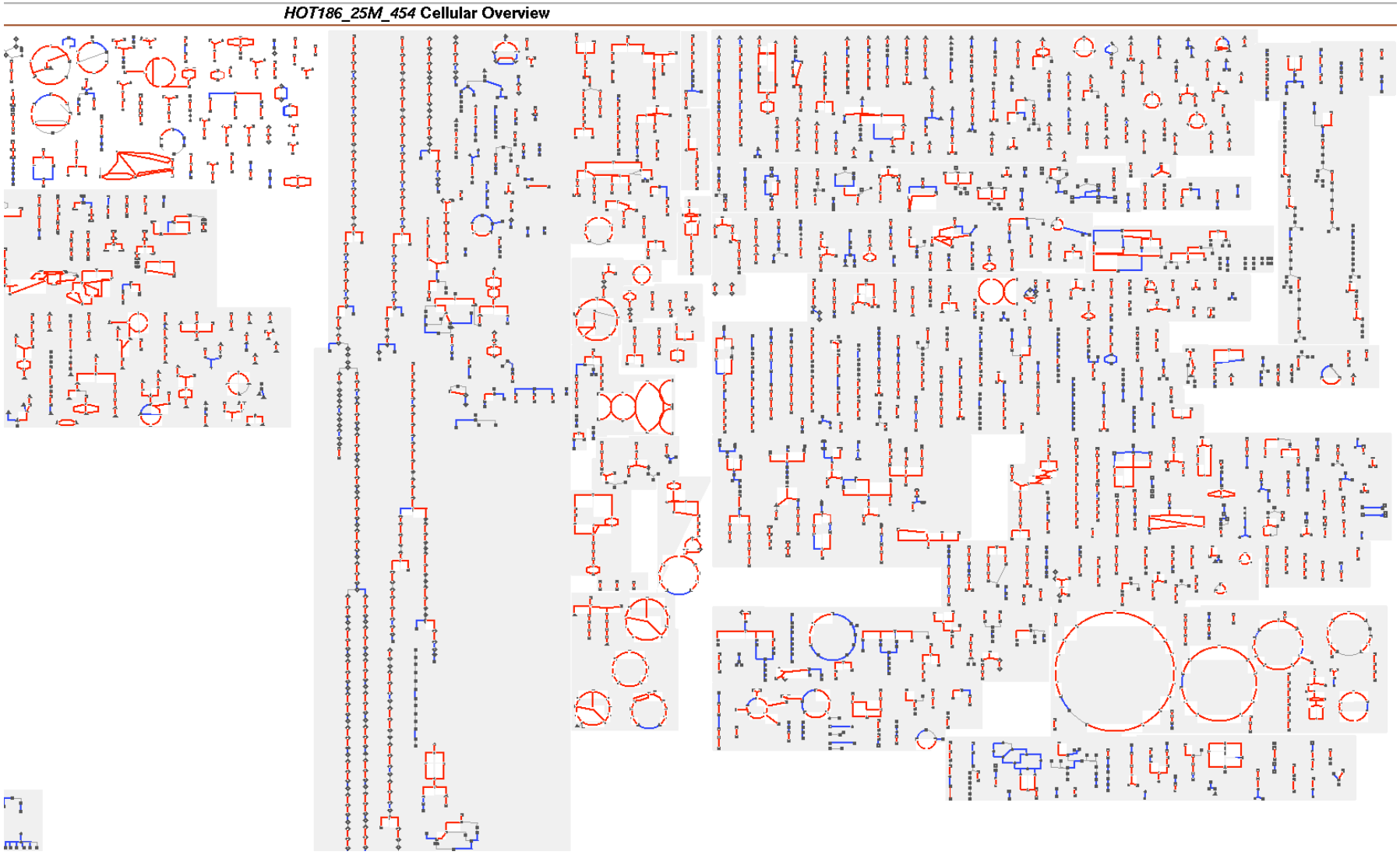
c1988-2013

HOT ePGDBs

Sample Depth (m)	Sample Description	Library Type	Sequencing Platform	Number of Sequences	Average Sequence Length	Protein Coding Sequences	Annotated Coding Sequences	MetaCyc Reactions	MetaCyc Pathways
25	upper euphotic	DNA	Roche 454	623559	257	405613	214149	4138	864
75	upper euphotic	DNA	Roche 454	673674	244	430689	222572	4052	854
110	chlorophyll max	DNA	Roche 454	473166	270	336035	165775	4133	860
500	mesopelagic	DNA	Roche 454	995747	276	714743	361193	4464	949
25	upper euphotic	RNA	Roche 454	561821	248	234404	85781	3433	723
75	upper euphotic	RNA	Roche 454	557718	239	203359	66855	3208	669
110	chlorophyll max	RNA	Roche 454	398436	228	135107	36912	2549	532
500	mesopelagic	RNA	Roche 454	479661	266	207465	71400	3034	641

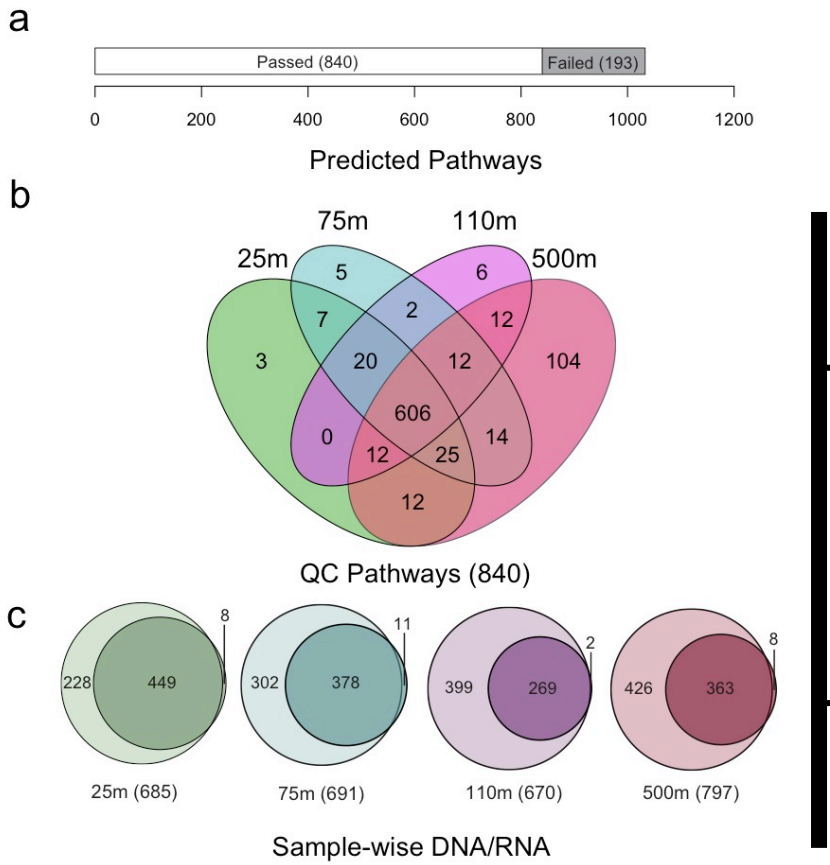
- Over 1,000 MetaCyc Pathways were predicted in combined DNA and RNA datasets from the HOT water column.

Cellular Overview



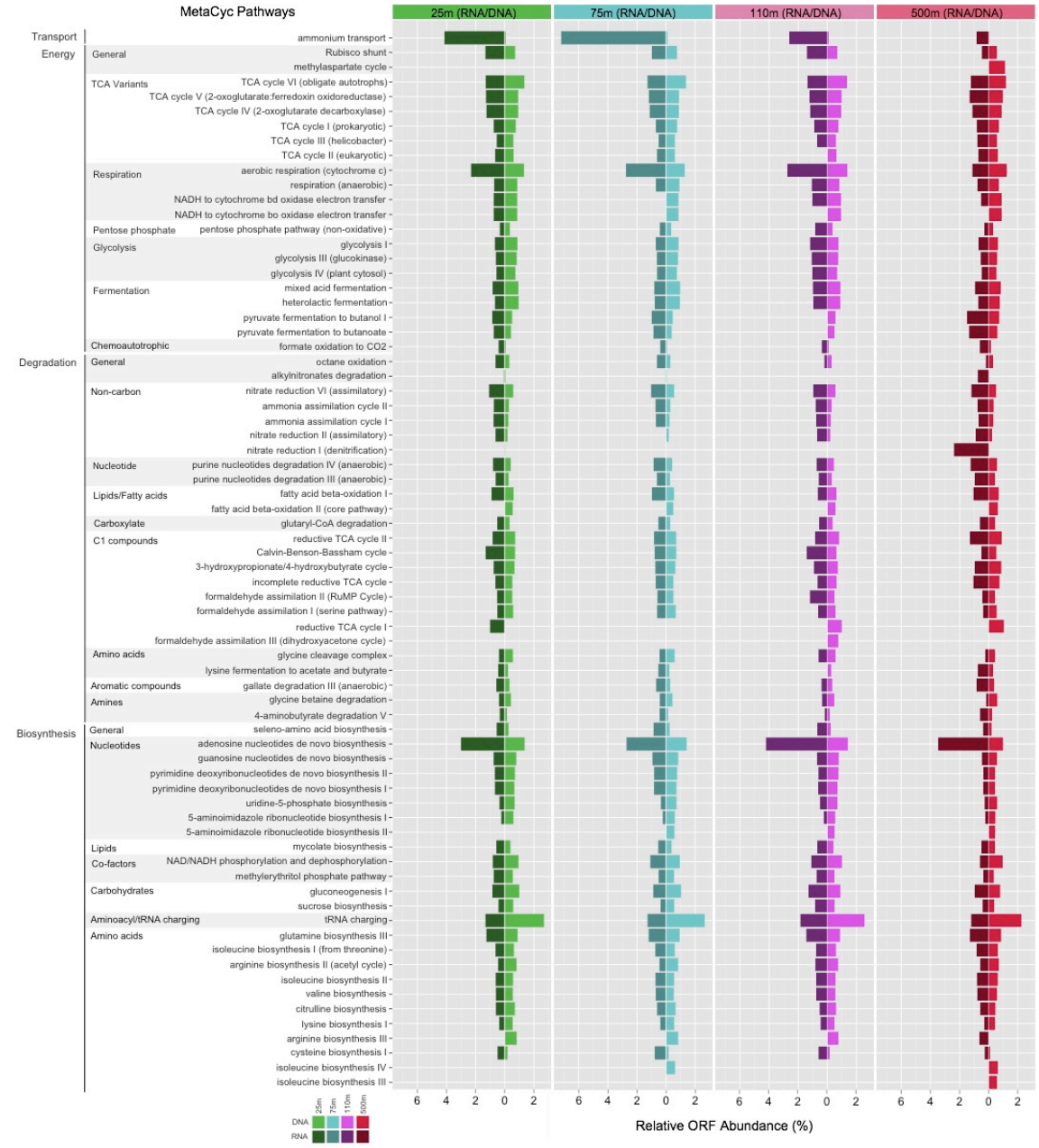
- Comparison of DNA (Blue) and RNA +DNA (Red) pathway predictions

Four-way Set Difference



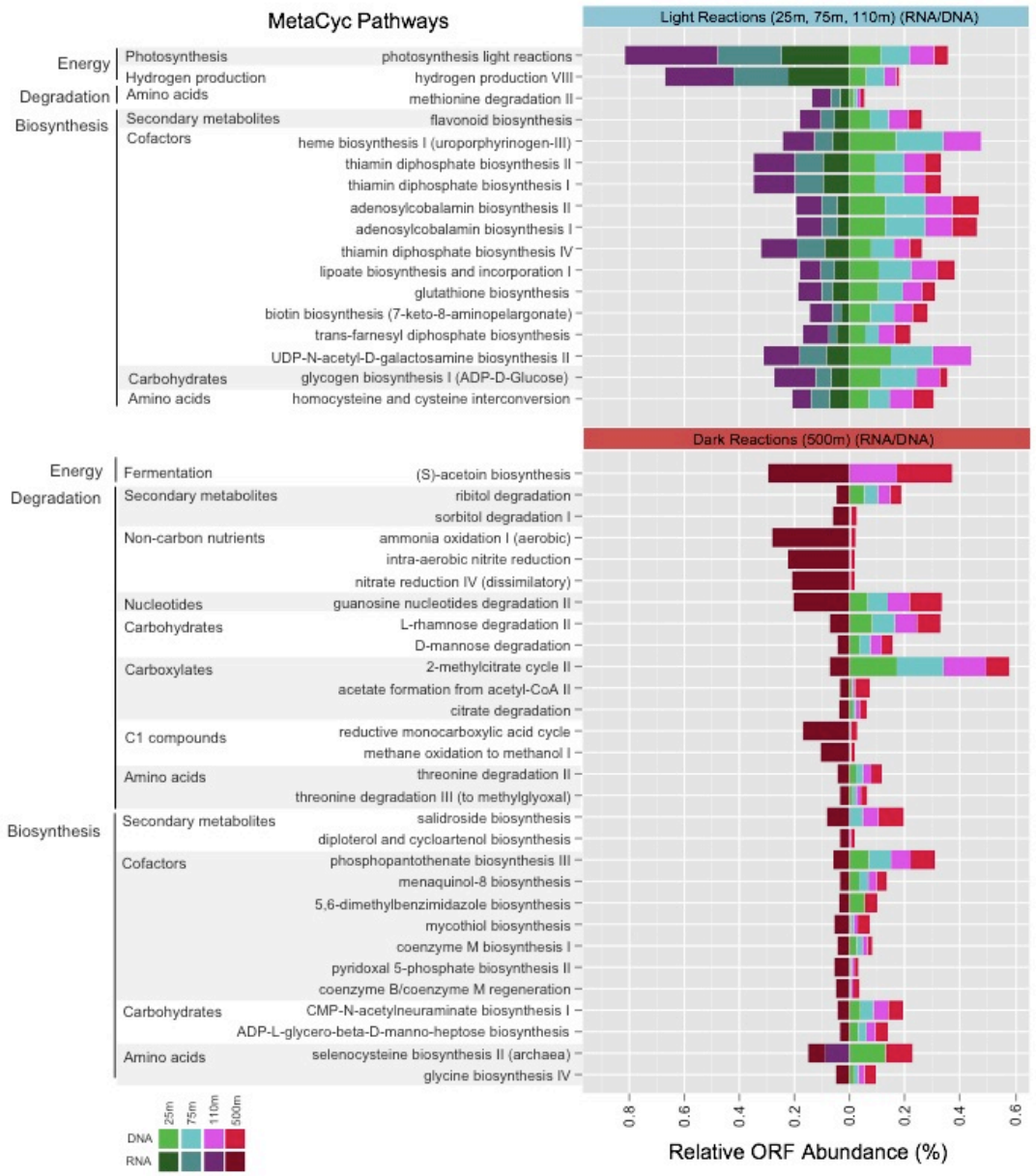
- Core metabolic pathways and those differentially expressed between sunlit and dark ocean samples
- Pathways identified in RNA but not DNA datasets

Core Pathways



- Top 70 pathways in the combined DNA and RNA datasets. ORF counts shown as a two way bar plot for each depth. The order of the pathways is based on the pathway hierarchy and then by abundance across each depth interval.

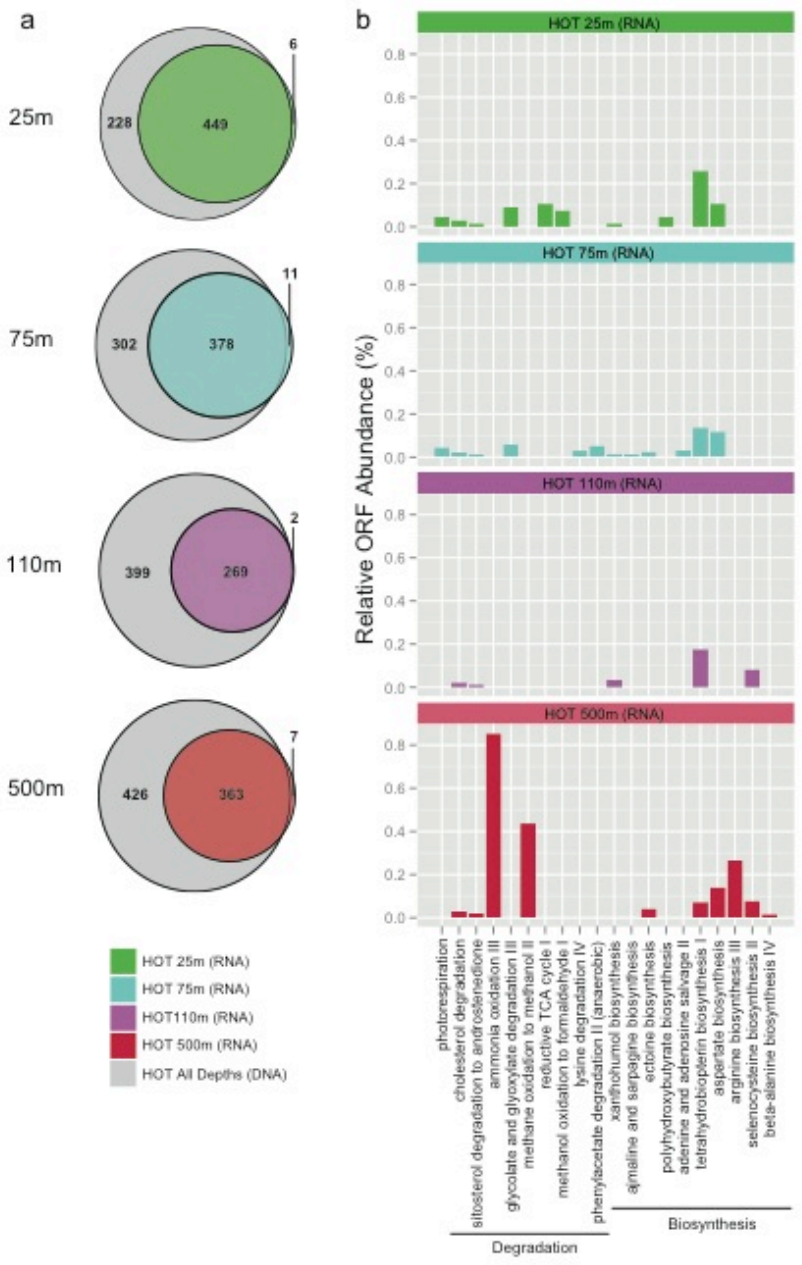
Regulated Gene Expression



- A number or pathways were present throughout the water column but differentially expressed between sunlit and dark ocean waters

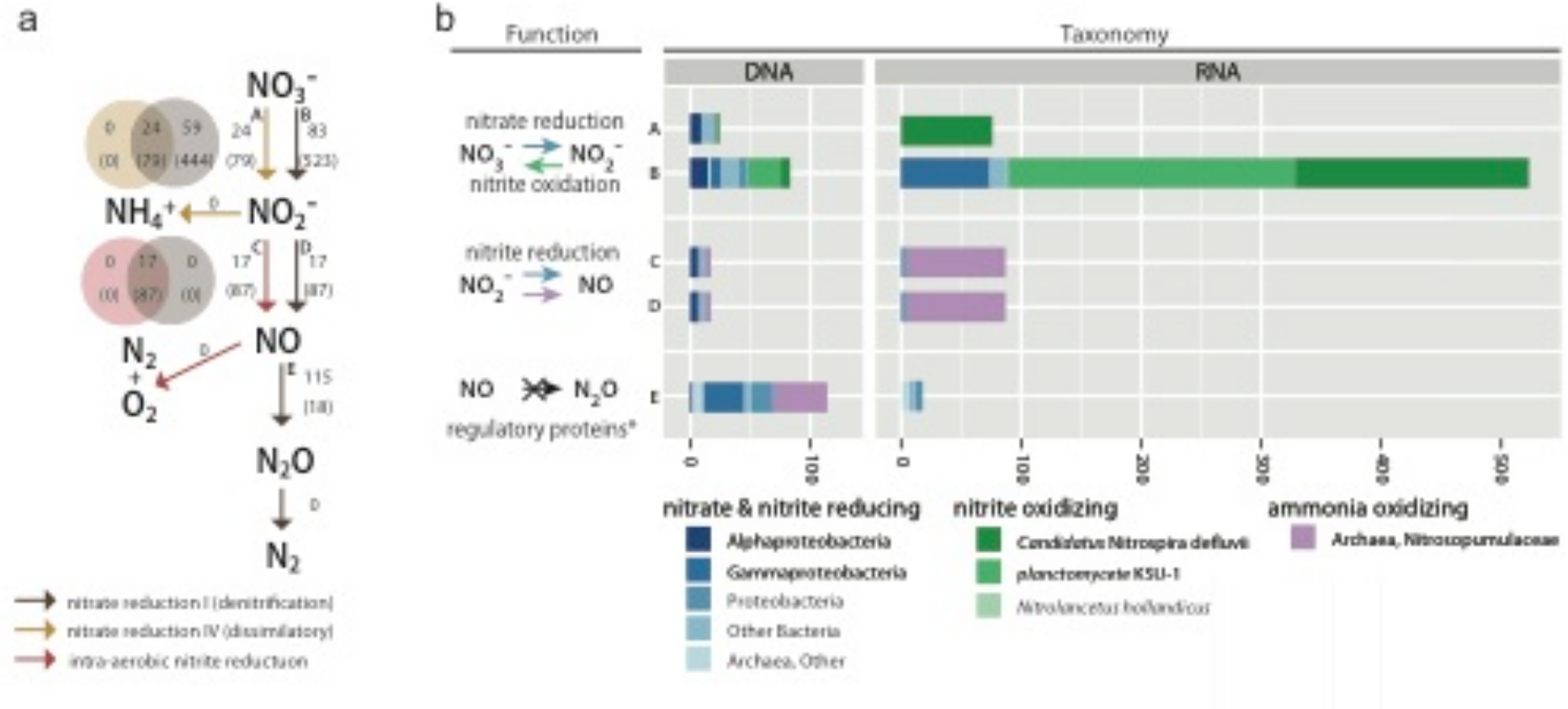
- These included pathways for photosynthesis, hydrogen production and vitamin B synthesis in sunlit waters and pathways for organic matter degradation, ammonia oxidation and “denitrification” in dark ocean waters

Cryptic Pathways



- For each depth interval, a small number of pathways were predicted in RNA but not DNA datasets
- These RNA-specific pathways including photorespiration and ammonia and methane oxidation showed depth distributions consistent with niche-partitioning between sunlit and dark ocean waters

Nitrogen Cycle



- a) Nitrate reduction I (denitrification) (brown), nitrate reduction IV (dissimilatory) (yellow), and intra-aerobic nitrite reduction (red). b) nitrate and nitrite reducing taxa (blue), nitrite oxidizing taxa (green), and ammonia oxidizing taxa (purple).

Things to Keep in Mind…

- Pathway Tools cannot predict pathways not present in MetaCyc
- Evidence for short pathways is hard to interpret
- False positives due to shared enzymes in multiple pathways or incorrect annotations create hazards
- Currently no taxonomic assignment or coverage information is mapped onto identified pathways in Pathway Tools
- Limited functional validation for pathways in metagenomes

“One gene is many hypotheses” *Julian Davies*


BAKER HUGHES
Baker Petrolite


ConocoPhillips


ENCANA


QUICKSILVER
RESOURCES


ARC ENERGY TRUST


TRIDENT
EXPLORATION CORP


SUNCOR
ENERGY


Syncrude
Securing Canada's Energy Future


PROFERO
ENERGY INC


iseee
INSTITUTE for SUSTAINABLE
ENERGY, ENVIRONMENT and ECONOMY


GenomeAlberta


GenomeBritishColumbia


ALBERTA WATER
RESEARCH INSTITUTE
water
AN INGENUITY INSTITUTE


Canada Foundation for Innovation


GenomeCanada