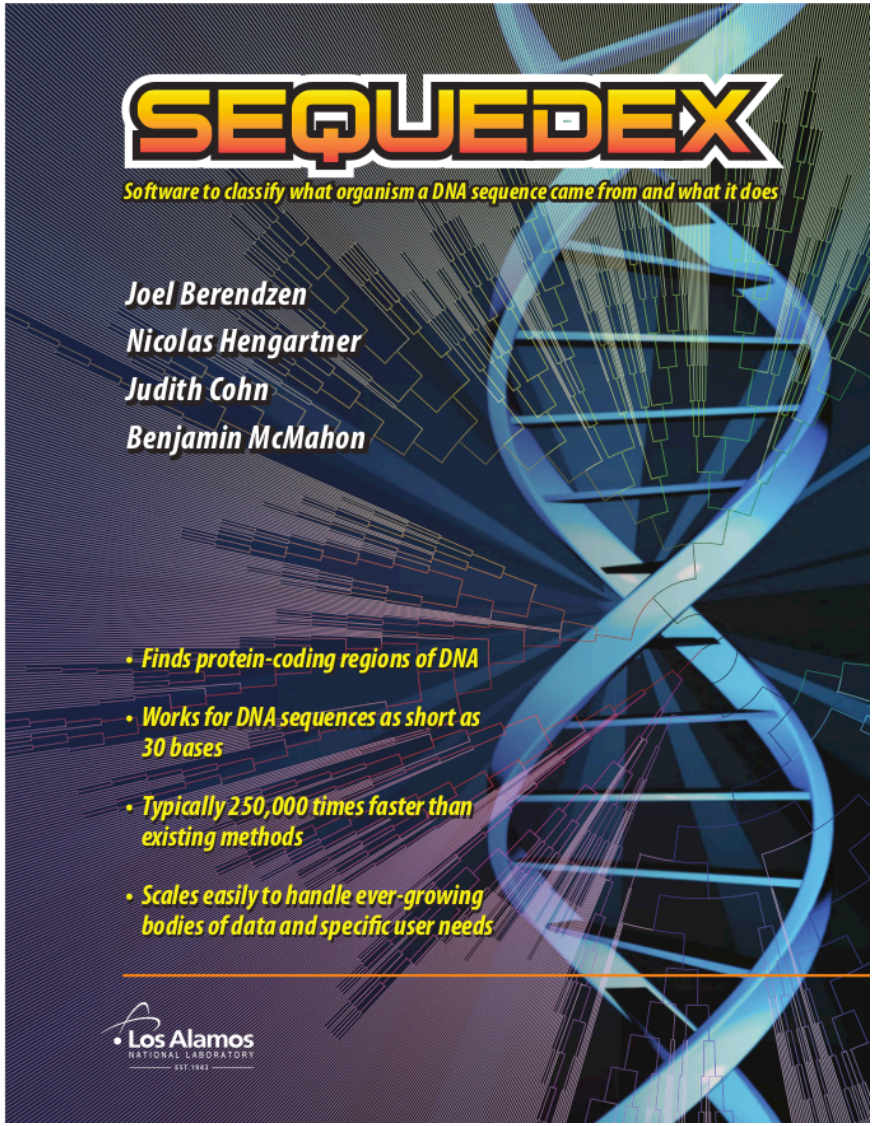
Signature peptide enabled metagenomics

Sequedex is next-generation bioinformatics

Nick Hengartner, Joel Berendzen, Judith Cohn, Ben McMahon
CCS, Physics, and Theoretical Divisions
Los Alamos National Laboratory



2012 R&D 100 submission

LA-UR 12-xxxxx

Licensable Technologies

SEQUEDEX: Classify Phylogeny & Function of Short DNA Sequences

Applications:

- Rapid ID of viruses and bacteria (medical, biodefense)
- Microbiome characterization (medical, pharmaceutical, consumer)
- Metagenome characterization (ecology, manufacturing)
- Transcriptome characterization (medical, pharmaceutical, algal biofuels)
- Epidemiology (public health, product safety)
- Enzyme mining (chemical manufacturing, biofuels)
- Cancer genomics (medical, pharmaceutical)

Benefits:

- Extremely rapid (~6 Gbp/hr on a desktop, ~250,000X faster than BLAST)
- High sensitivity, selectivity even for short reads (as short as 30 bp)
- Produces searchable database of classified reads
- Works for novel and known organisms, including eukaryotes and viruses
- Characterizes sample complexity (biodiversity)
- Characterizes similarities among samples (beta diversity)
- Built-in evolutionary theory simplifies bioinformatics
- Drops bioinformatics costs

Contact:

David Hadley
(505) 667-7539, dhadley@lanl.gov
sequedex@lanl.gov
Technology Transfer Division

Los Alamos NATIONAL LABORATORY ESTABLISHED 1943

06/12

SEQUEDEX ...ATG

Summary:

Sequedex is bioinformatics redesigned from the ground up to make use of the wealth of genomic data that has become available in the 20 years since the most commonly used algorithms like BLAST were written. The result is a remarkable increase in speed and sensitivity: Sequedex can chew through one human genome's worth of DNA analysis in 30 minutes on a laptop, getting usable results from reads as short as 30 base-pairs.

Sequedex also employs a science-of-signatures approach that combines a foundation in evolutionary theory with well-developed informaticstechnology taken from web search engines. Unlike conventional approaches that rely on "mapping" reads of DNA to a large and growing database of genomes, Sequedex uses evolutionary signatures to place reads directly on the Tree of Life. LANL scientists have constructed a Tree of Life that spans plants, animals, bacteria, archaea, and viruses, thus allowing Sequedex to accurately classify DNA sequences from organisms which have no close relatives in databases.

Development Stage:

A demo version of Sequedex will be available for free public download at http://sequedex.lanl.gov pending copyright and publication issues. Features beyond those in the demo version, such as writing a SQLite database of classified reads, can be enabled by purchase of a software license from Los Alamos. Contact sequedex@lanl.gov for details.

Patent Status:

A patent application on the underlying algorithm is pending. The software package is copyright-protected.

Licensing Status:

Los Alamos is seeking commercial partners for licensing and distribution. Exclusive licenses in limited markets are possible.

www.lanl.gov/partnerships/license/technologies/

An Equal Opportunity Employer / Operated by Los Alamos National Security LLC for DOE/NNSA

Software information and future release at:
www.lanl.gov/org/tt/pdf/techs/sequedex.pdf

Manuscript under (re)review at BMC Bioinformatics

6 Gbp/hr on single processor of laptop

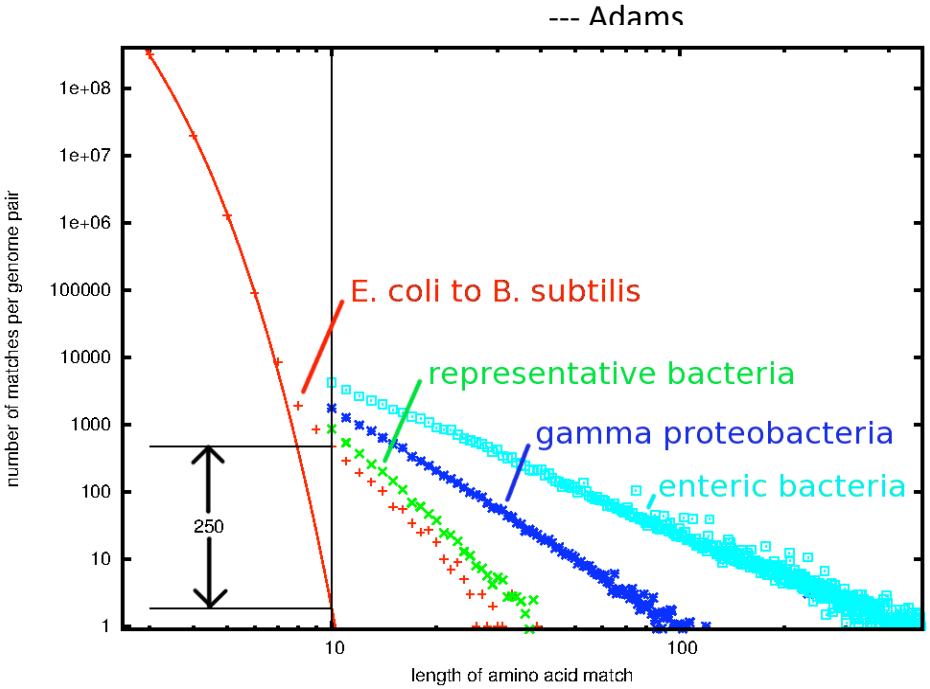
Our signatures are currently 10-mers of amino acids

- *k*-mers (*N*-grams, solid patterns) have a long history in both biological and natural-language research. In biology, they are physical objects.

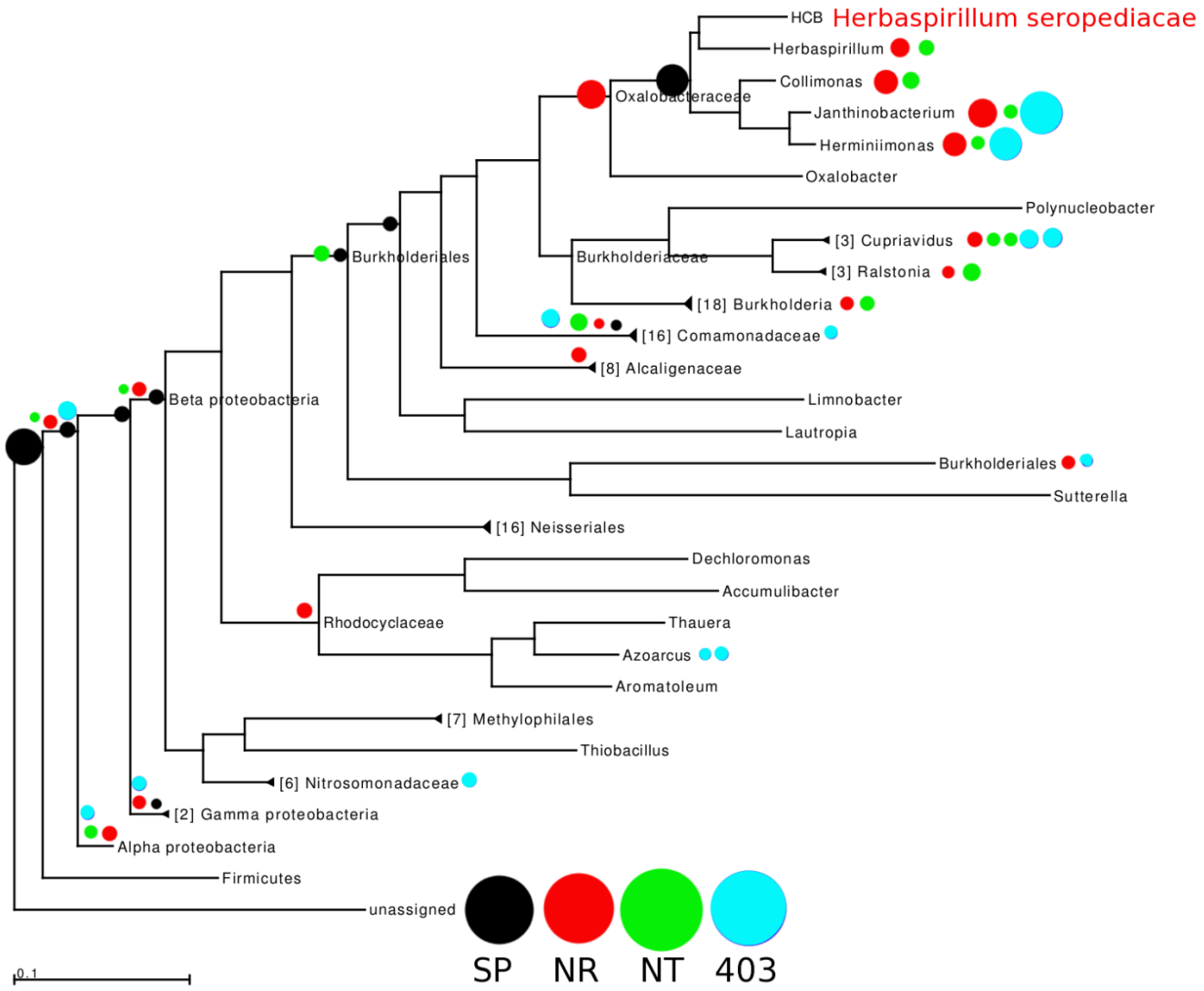
English	Science is the great antidote to the poison of superstition.
French	La Science est le grand antidote au poison de la superstition.
Spanish	La ciencia es el gran antidoto al veneno de la supersticion.
Dutch	Wetenschap is het grote tegengif aan het vergift van bijgeloof.
German	Wissenschaft ist das grosse Antidot zum Gift von Aberglauben.

Amino-acid (protein) rather than nucleotide (DNA):

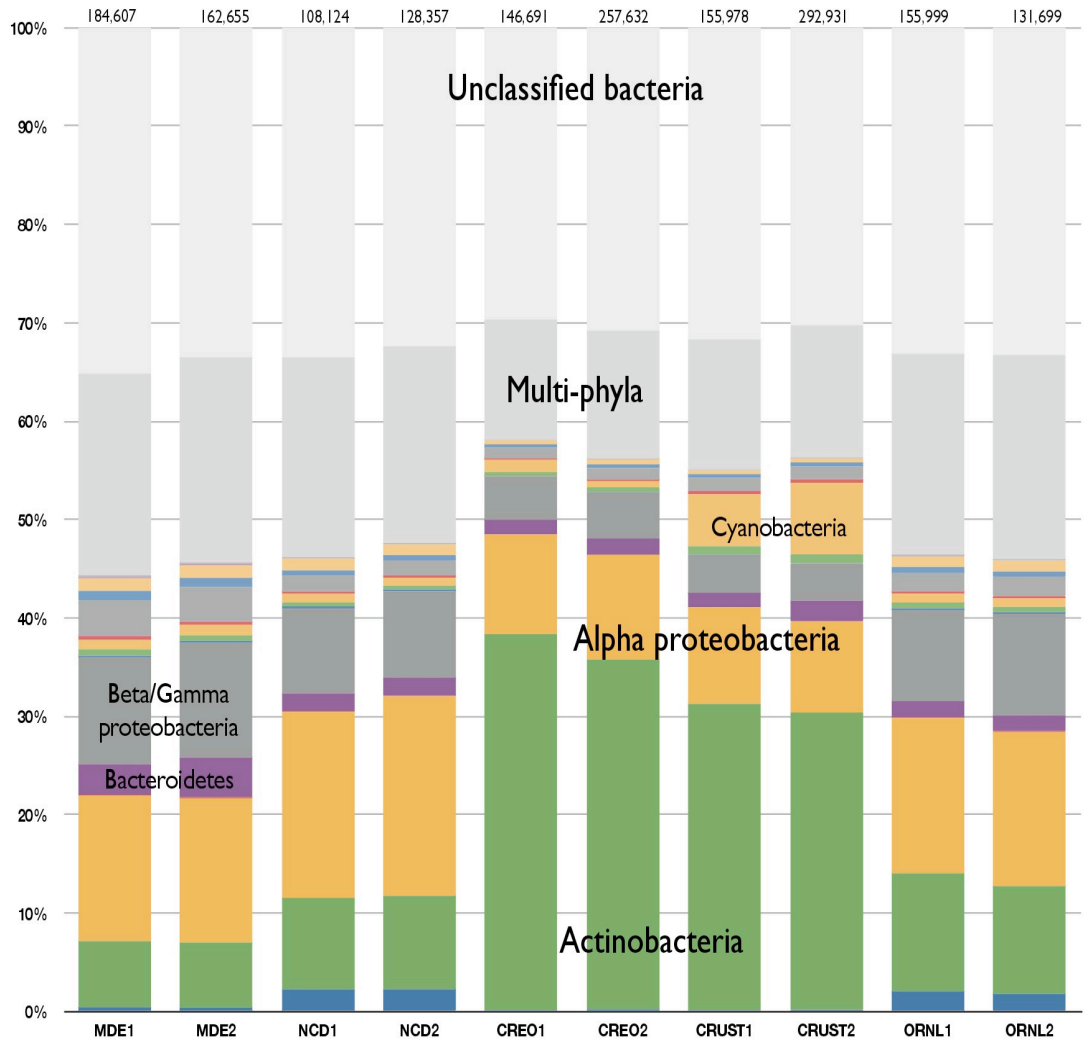
- Coding fraction is high for microbes
- No "silent" mutations-radius of convergence
- Functional pressures are mostly on proteins
- Protein sequences appear to be random strings
- 10-mers to achieve needed level of signficance (< 99.7% non-random matches)
- Shared AA 10-mers are shared for a reason

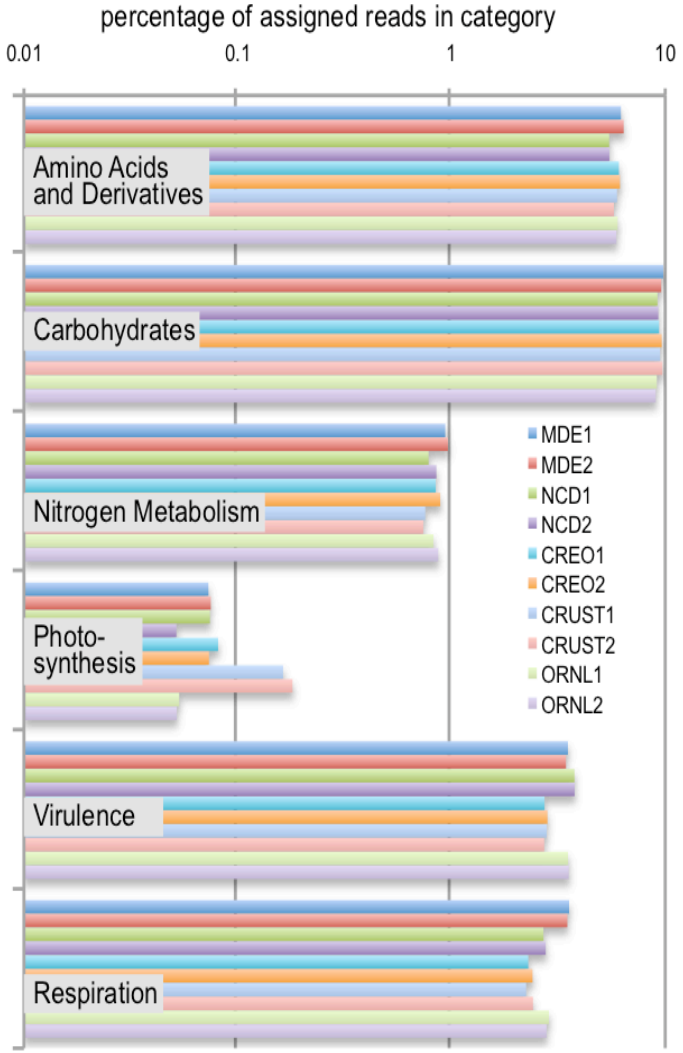


Sequedex assigns novel reads to nodes of tree, rather than leaves



Phylogenetic and functional profiles

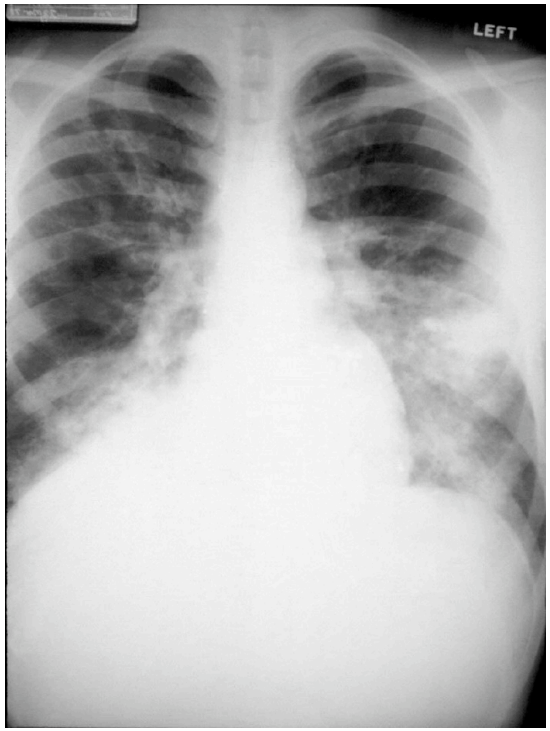




Applications

- Rapid ID of viruses and bacteria (medical, biodefense)
- Microbiome characterization (medical, pharmaceutical, consumer)
- Metagenome characterization (ecology, manufacturing)
- Transcriptome characterization (medical, pharmaceutical, algal biofuels)
- Epidemiology (Public health, product safety)
- Enzyme mining (chemical manufacturing, biofuels)
- Cancer genomics (medical, pharmaceutical)

Downstream Analysis: Assembly, Epi, Correlation with Outcomes, Diagnostics, Therapeutics



Acute Respiratory Tract Infection

15 – 30%

unknown cause

Gastroenteritis

30 – 50%

unknown cause

Meningitis / Encephalitis

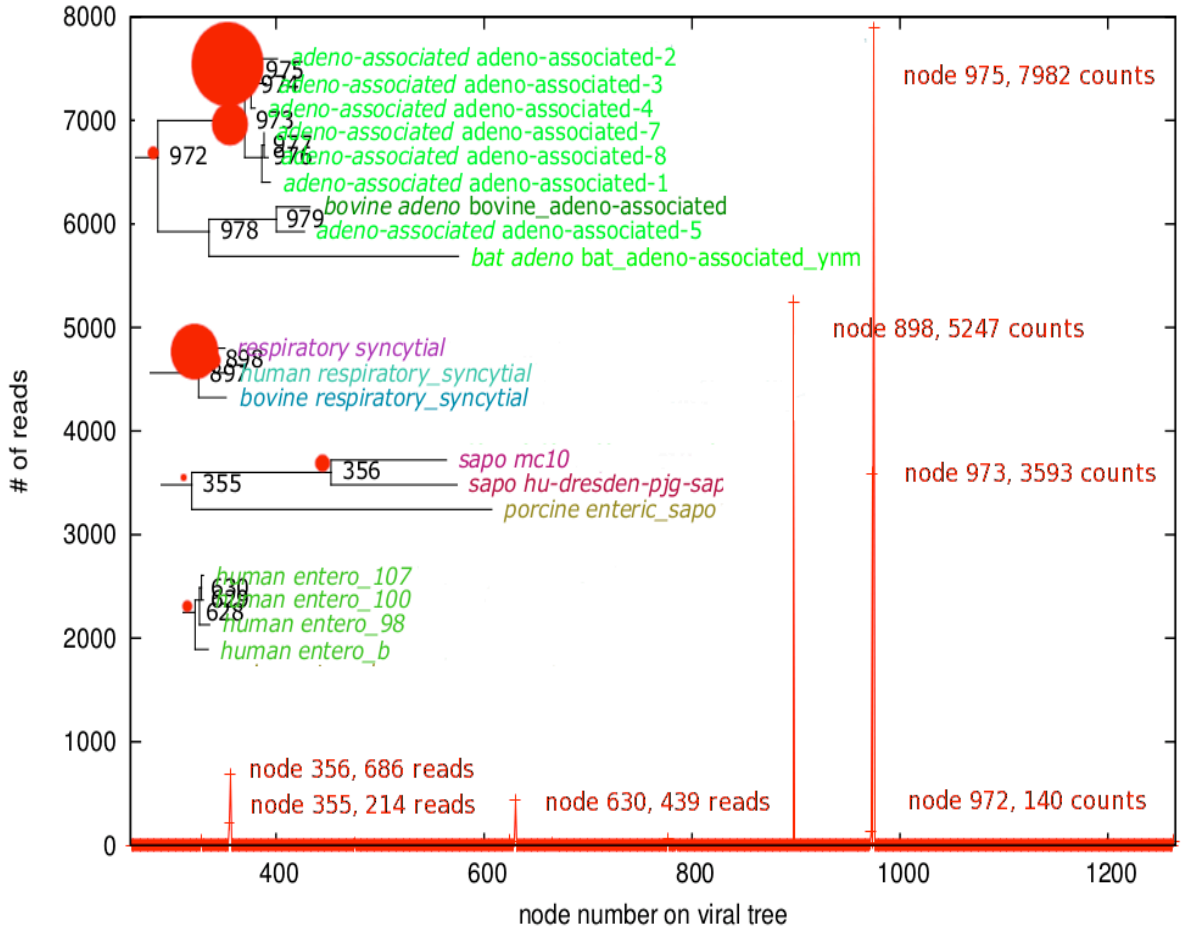
50 – 90%

unknown cause

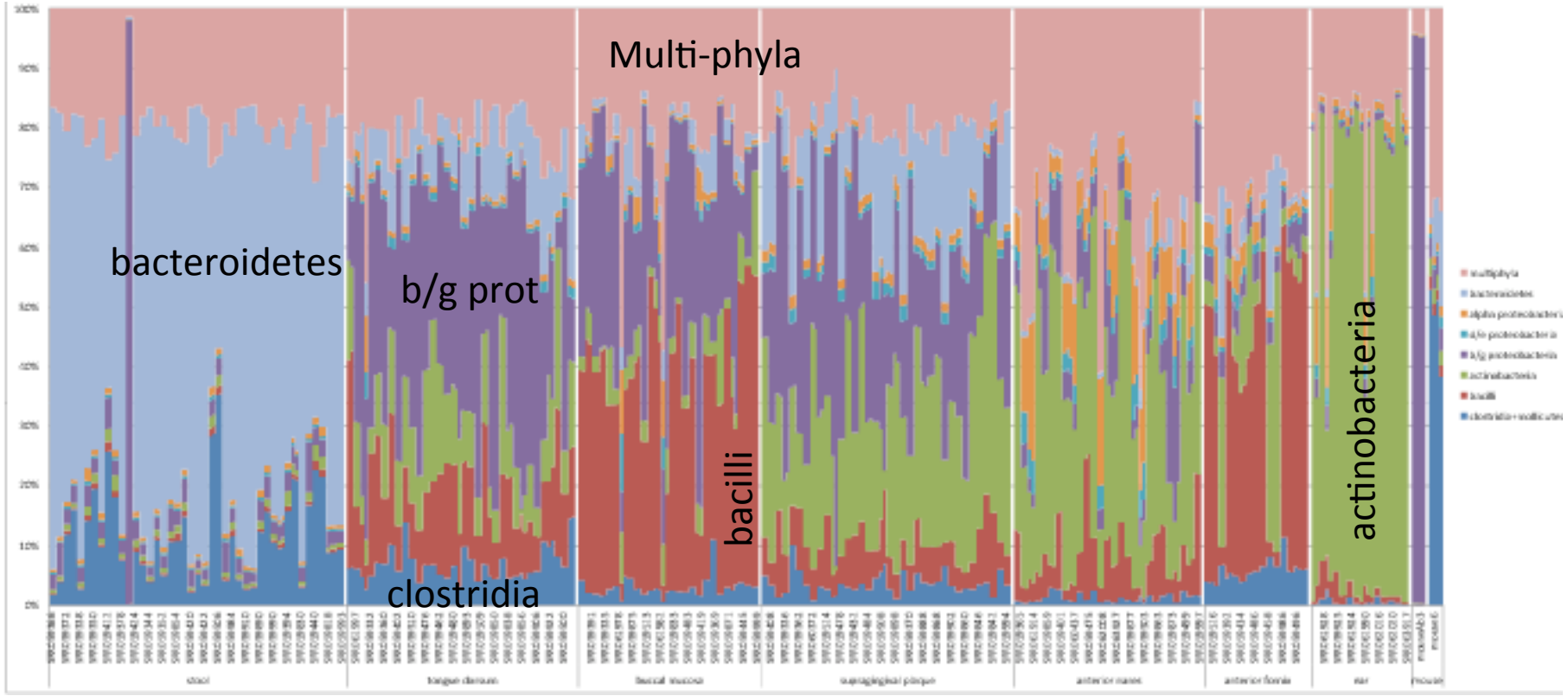
Directly addresses clinical needs in viral diagnostics
Can also be used to detect pathogenic bacteria, fungi, and parasites
Implementation of multiplex disease-specific panels amenable to clinical
Validation, CLIA waiver, and ultimately 510(k) FDA approval
(e.g. respiratory infections, febrile illness, diarrhea, meningitis, encephalitis)
Validation of Sequedex tool will be done in a CLIA laboratory with direct comparis
to FDA approved or in-house validated assays (Cepheid GenXpert, Luminex RVP,,
Etc.)

Preliminary Results (Clinical Data): ID All Viruses in a Sample

- Data from 200 μL clinical samples (diarrhea)
- ~100M reads (100-bp)
- $3000 reagent cost, next-gen sequencing
- ~1 hr analysis time, 1 CPU, 4 GB memory
- One assay ID's all viruses present as a series of peaks
- Automated readout and interpretation can be provided to the clinician or laboratorian
- Interesting reads can be automatically marked for upload to central server
- Virus sensitivity can be optimized by lowering signature significance & improved host knock-out

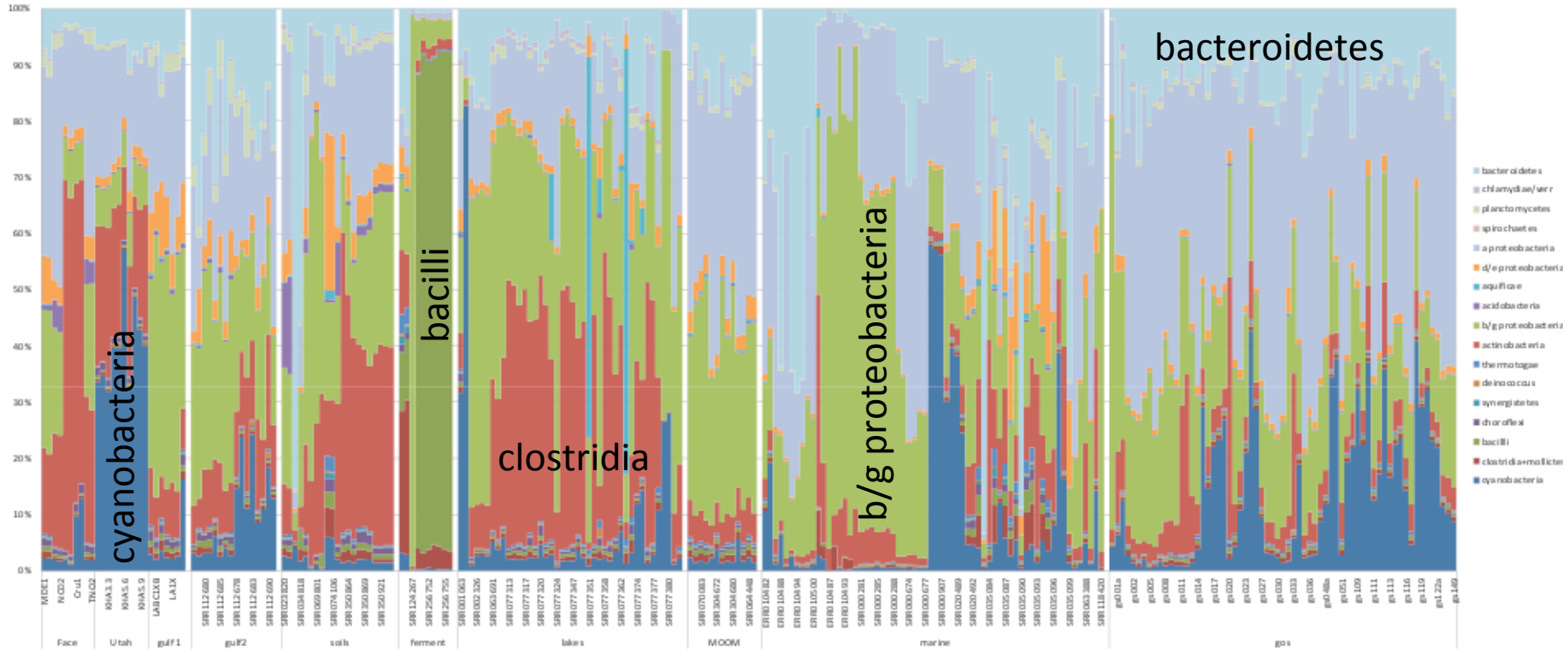


Phylogenetic profile across 400 ‘normal’ human microbiomes (and 8 mice)



Stool Tongue Cheek Plaque Nose Vagina Ear mice

Phylogenetic profile across 258 environmental metagenomes



Soils

Kimchi

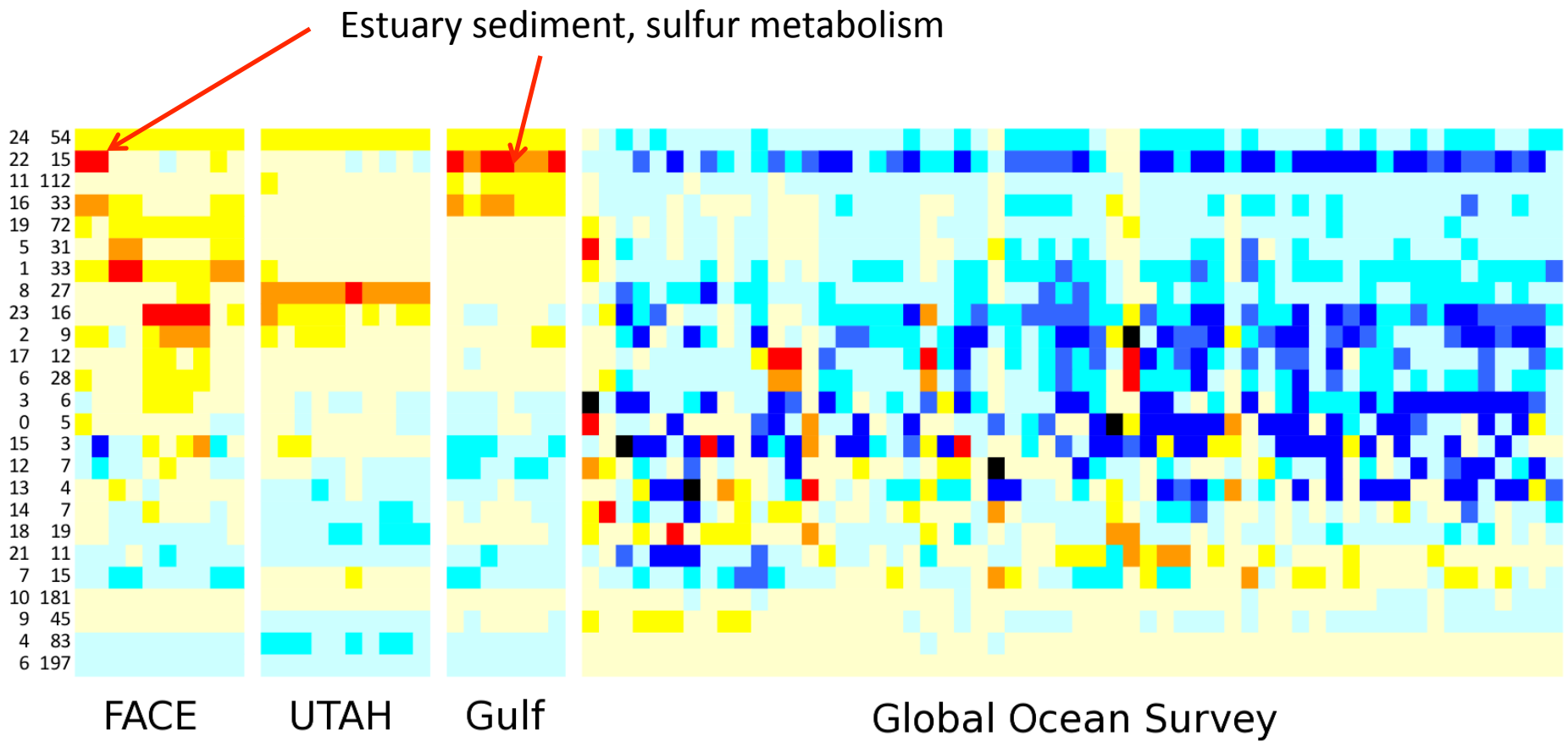
Lakes

Marine

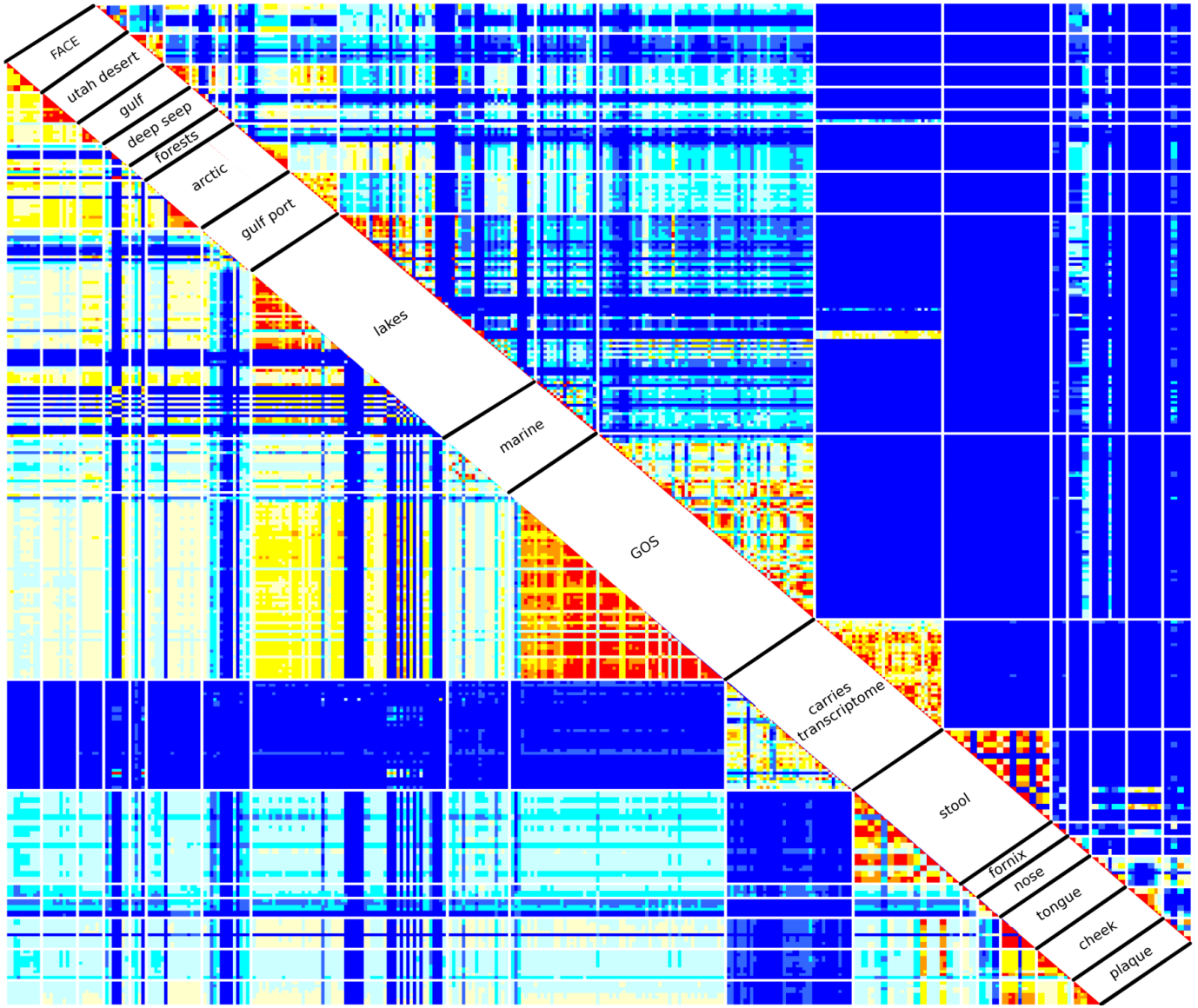
Global ocean survey

Identification of important niche genes

k-means clustering associates quantity of functional groups of genes with ecosystems



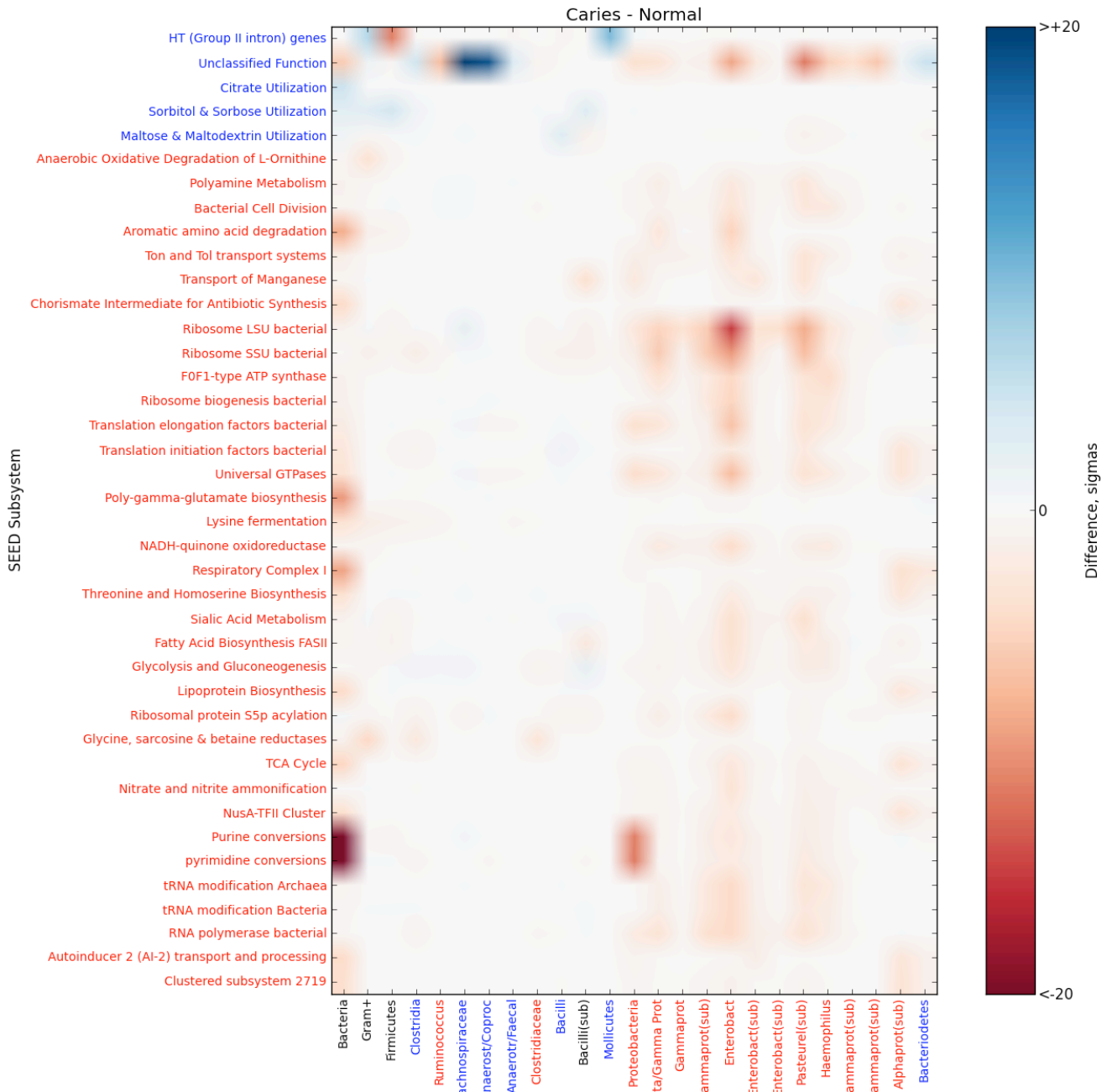
Microbial community similarity can be computed based on functional and phylogenetic profiles



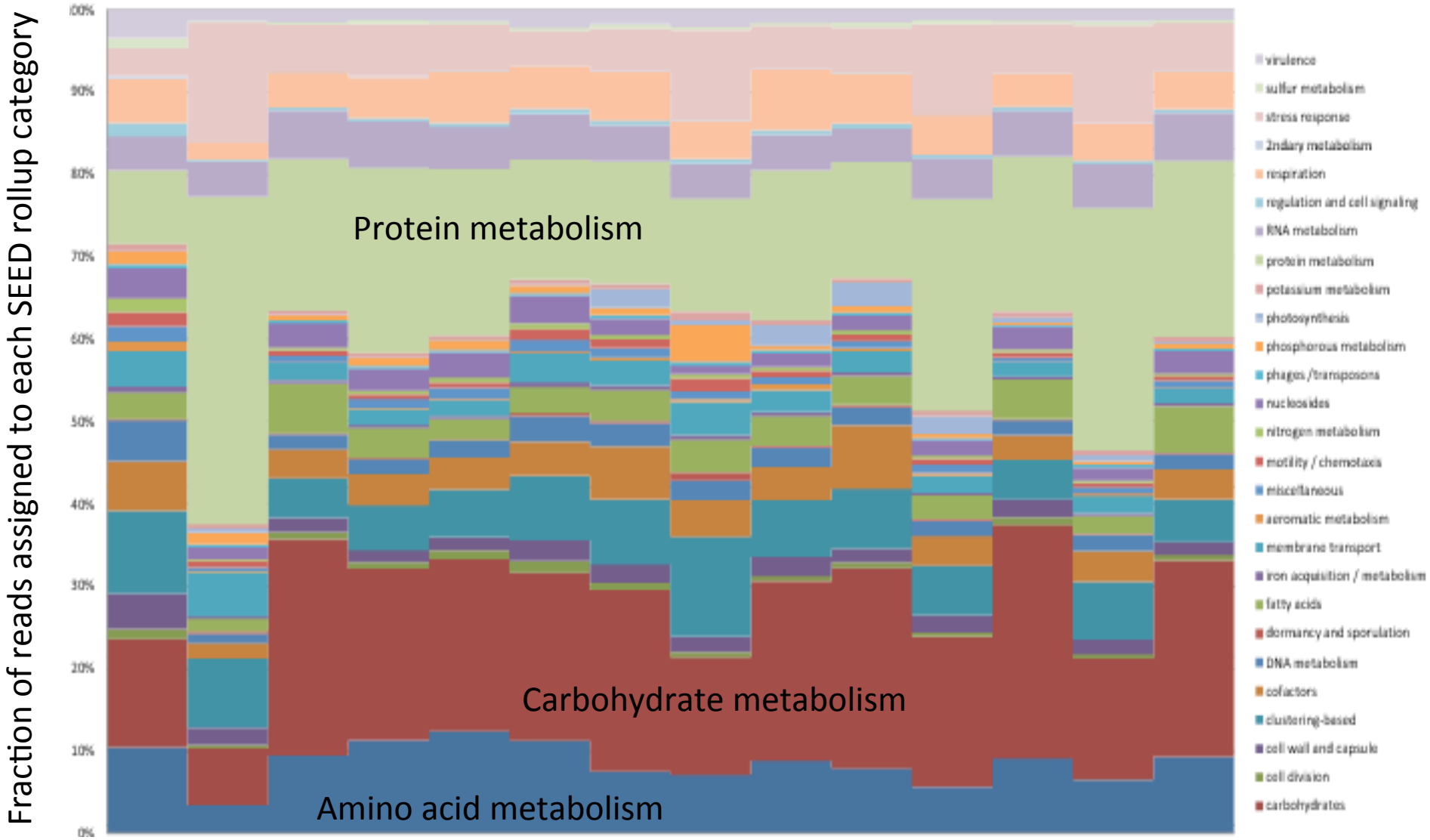
Functional similarity

Phylogenetic similarity

Human transcriptome data shows increased sugar metabolism in people with dental carries

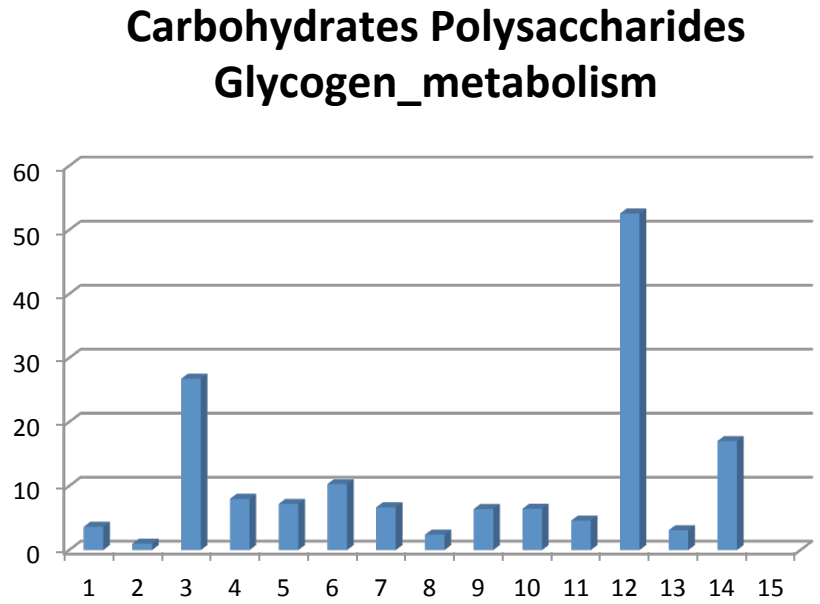


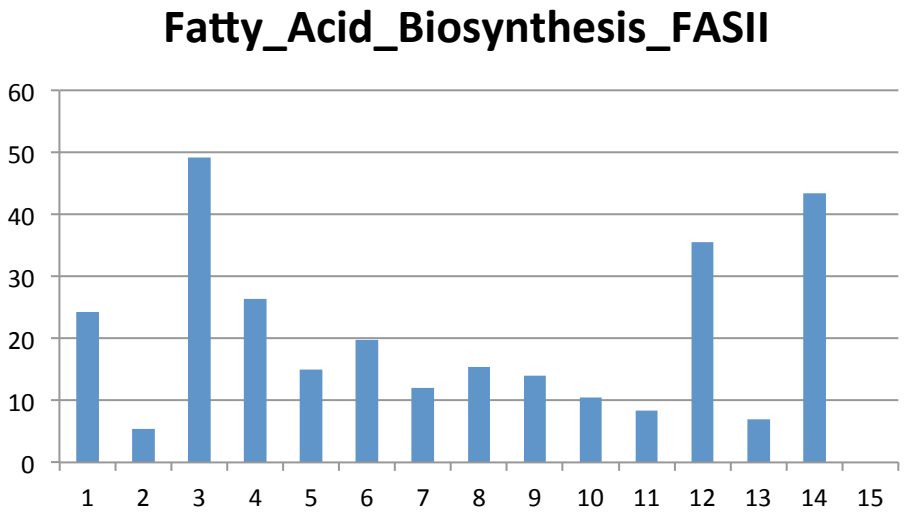
Functional profiling of algal growth in different conditions

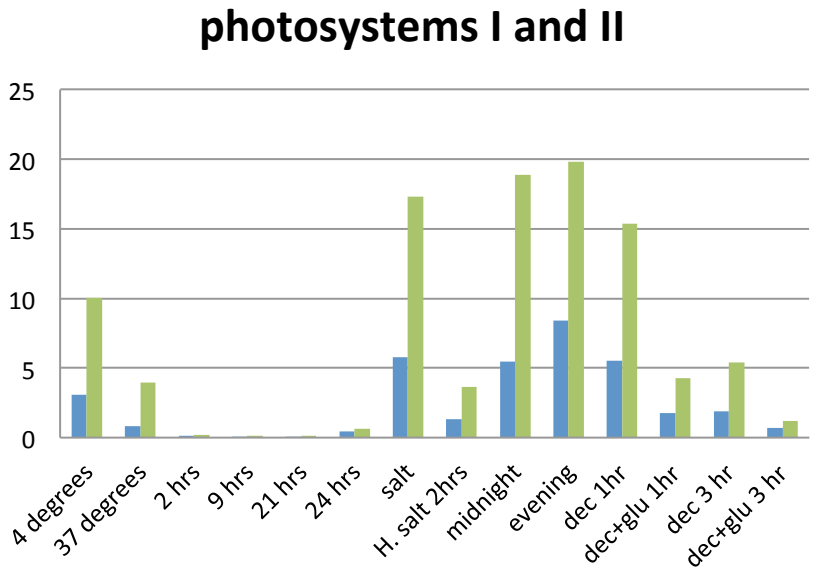


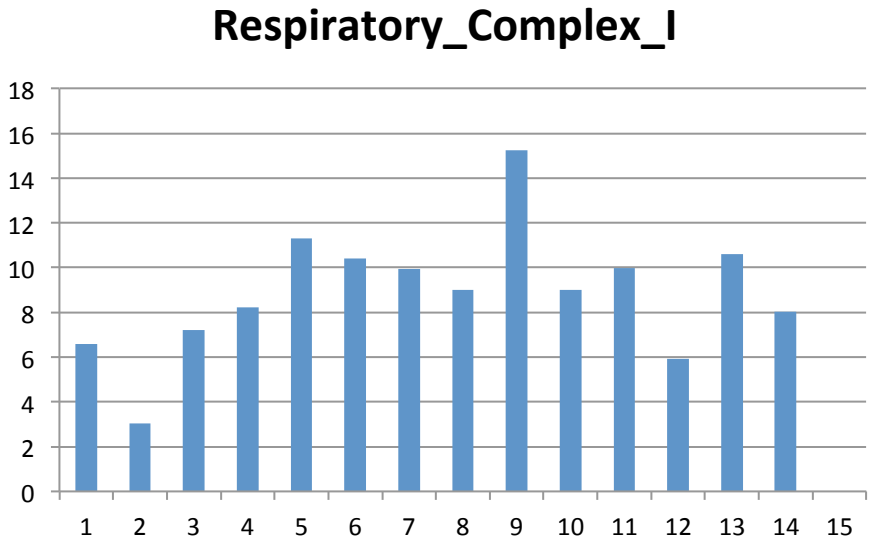
Growth conditions

Detailed profile of specific SEED subsystems across algal growth conditions









Acknowledgements

- Soil metagenomics: Cheryl Kuske, Cliff Han, Gary Xie, Blaire Steven, Bioscience, (LANL)
- Viral diagnostics: Charles Chiu, Narayanan Veeraraghavan (UCSF)
- Algal transcriptomics: Dick Sayre, Mustapha Samrakandi (LANL, NMC)
- Environmental metagenomics: Michael Blum, Alex Kolker (Tulane, Lumcon)