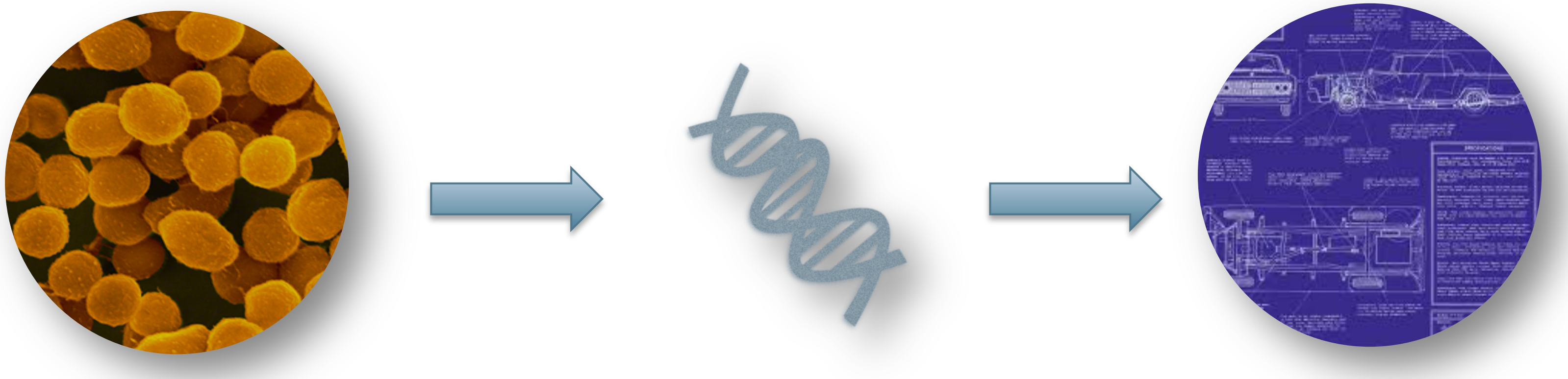
JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

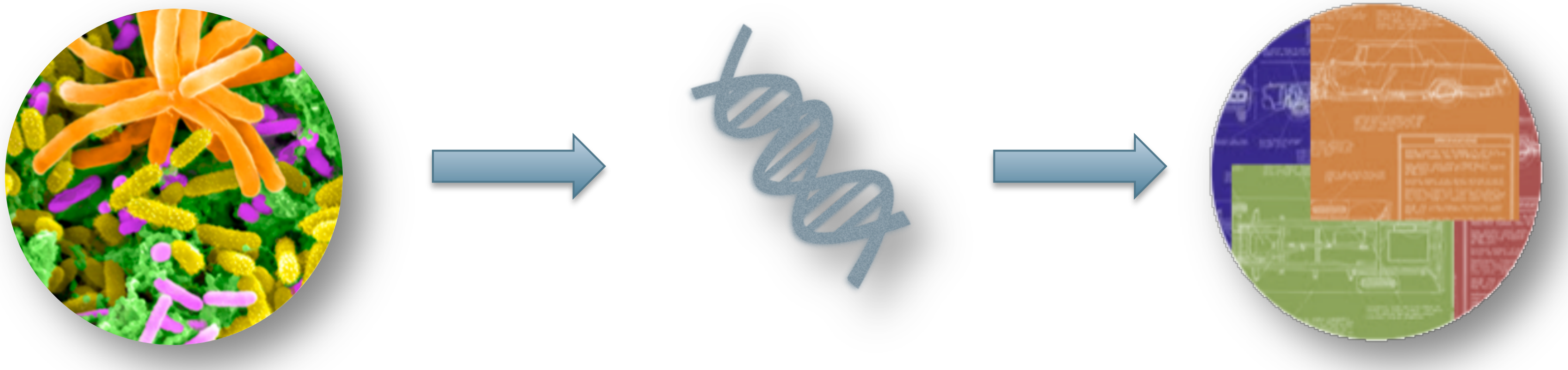
Metagenome and metatranscriptome sequencing: overview and case study from San Francisco Bay Wetlands

Susanna Theroux
Postdoctoral Fellow
Metagenome Group

What is metagenomics?



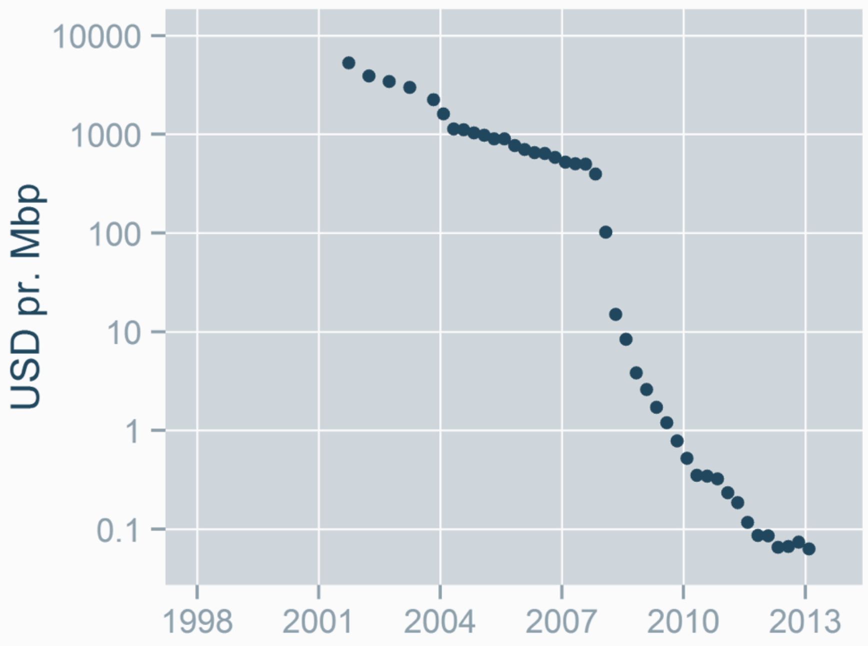
Genomics = pure, cultivated organisms



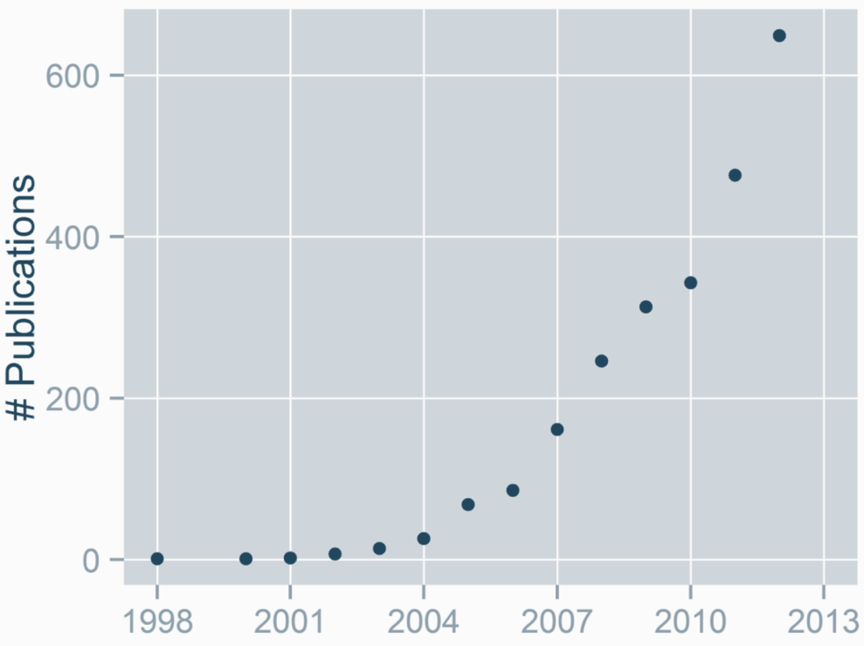
Metagenome = uncultivated communities

Photo: D. Kunkel; color, E. Latypova; slide modified from Mads Albertsen

Popularity of metagenomics



Sequencing costs

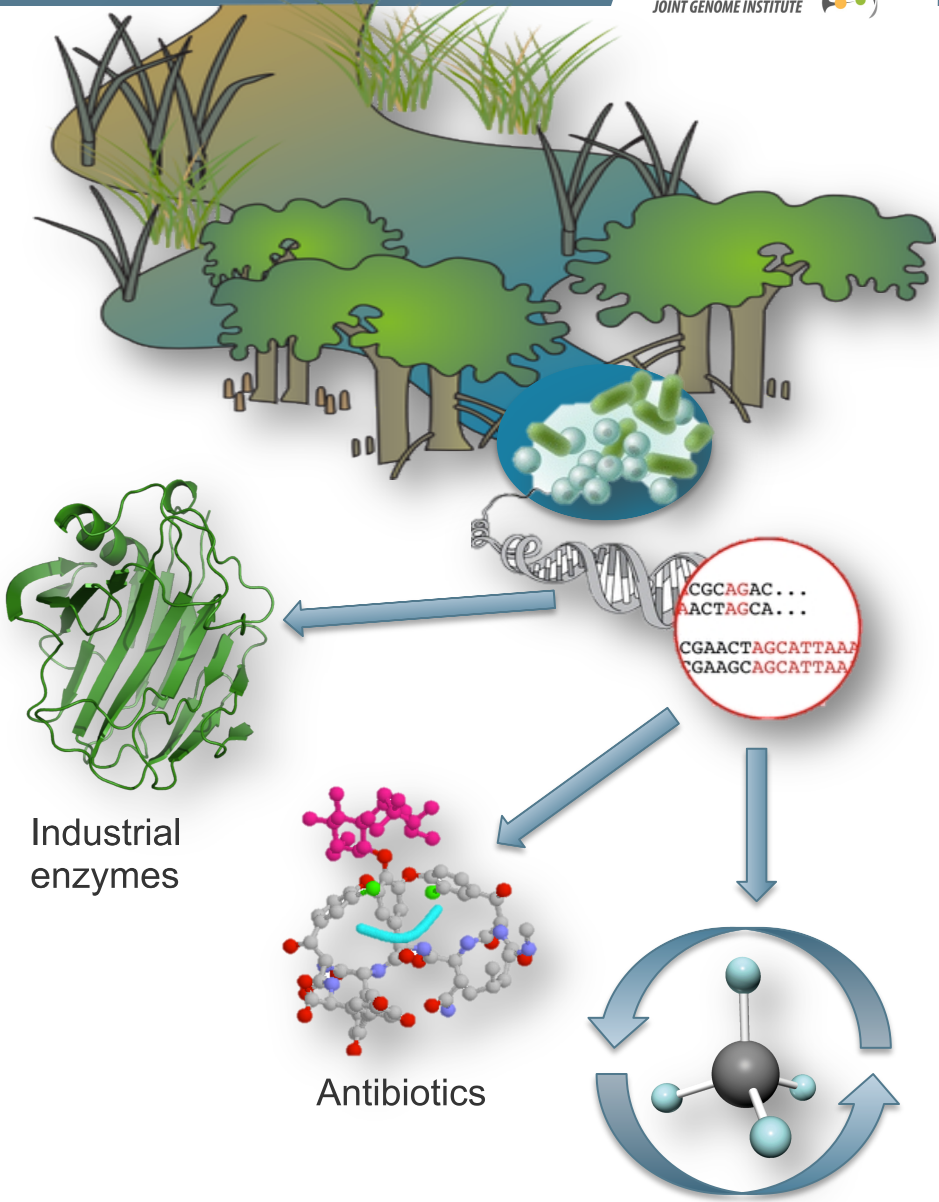


PubMed: metagenom*[Title/Abstract]

http://www.genome.gov/sequencingcosts/ ; slide modified from Mads Albertsen

Why metagenomics?

Most bacteria are uncultured!



GHG cycling

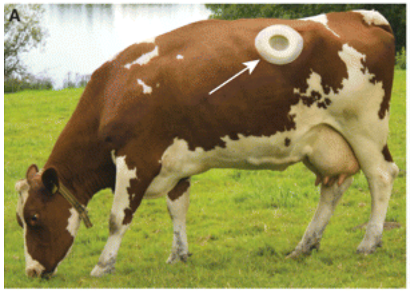
Suizenbacher et al 1997; www.chm.bris.ac.uk;

Early metagenomics

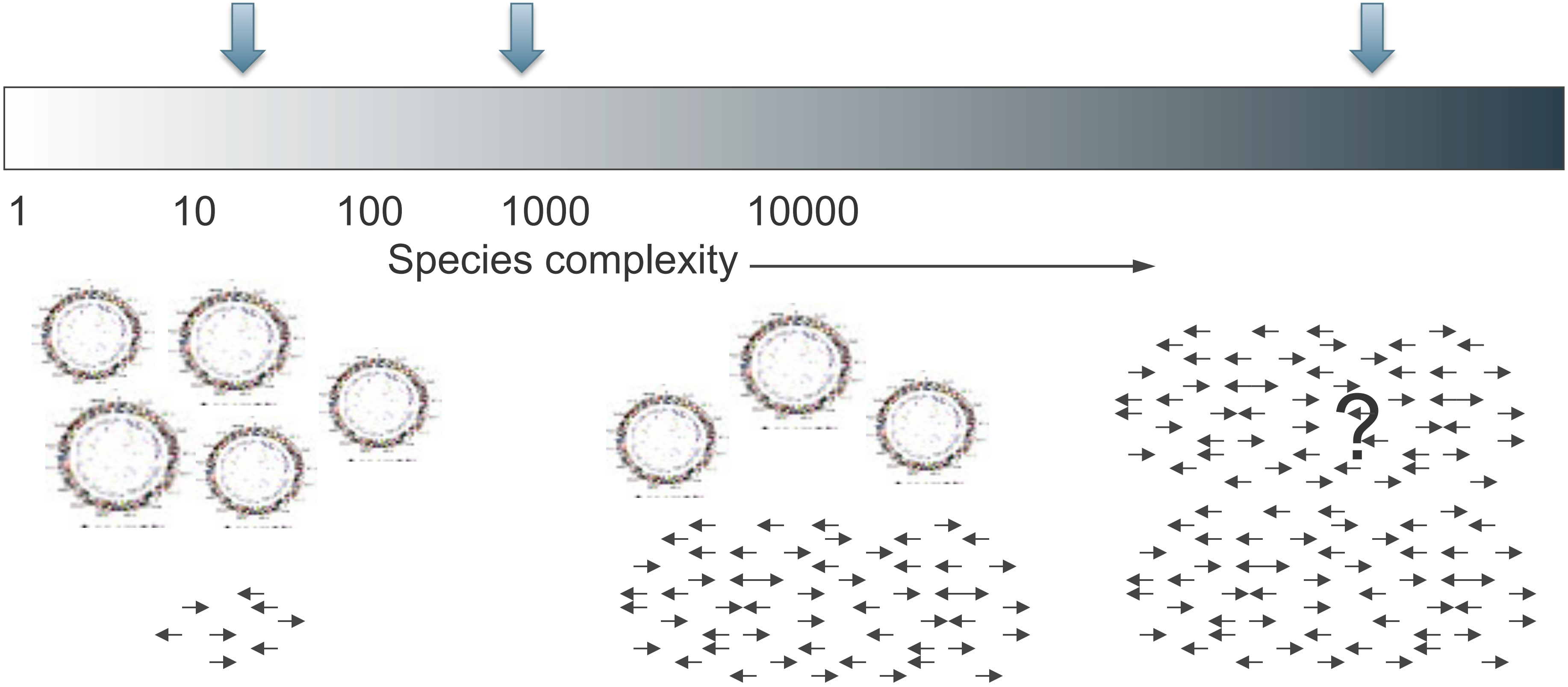
Acid mine
Tyson et al., 2004

Sargasso Sea
Venter et al., 2004

Cow Rumen
Hess et al., 2012



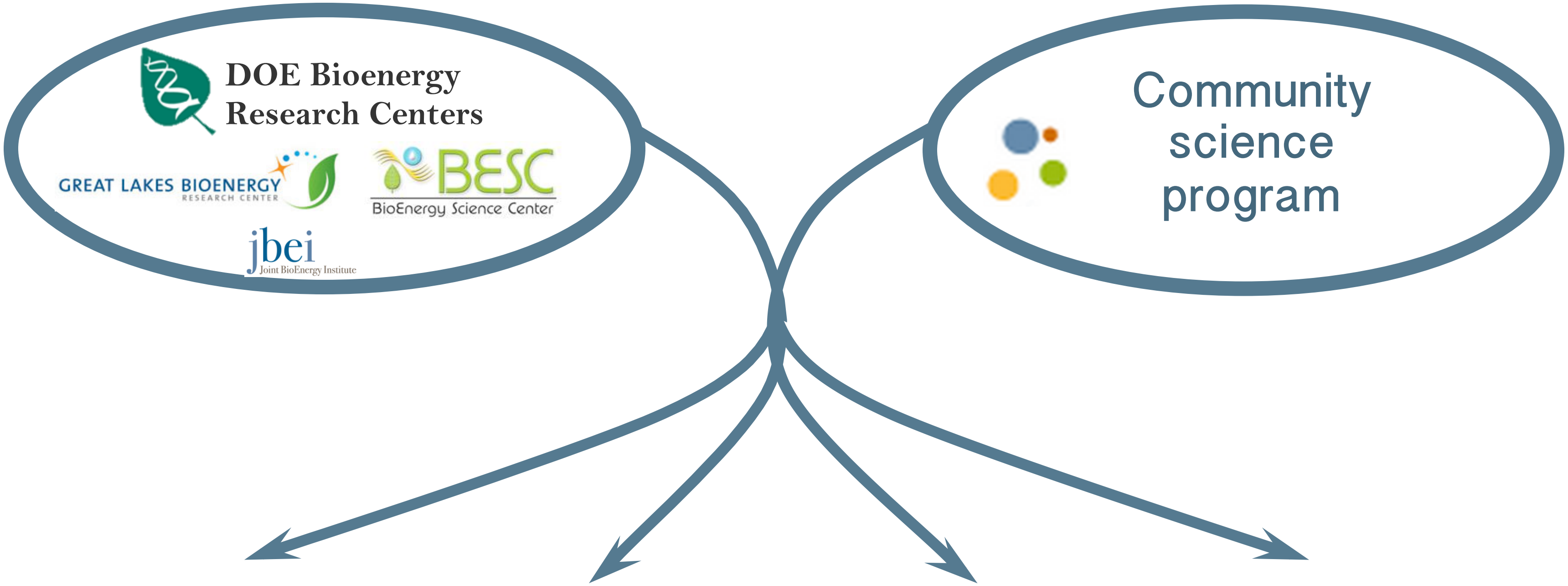
Soils



1 10 100 1000 10000

Species complexity →

User Programs



DOE Bioenergy Research Centers

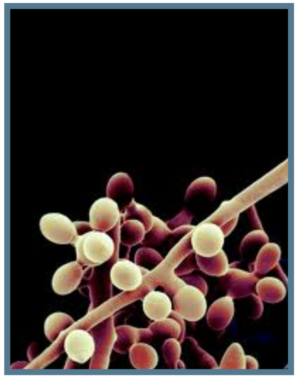
GREAT LAKES BIOENERGY
RESEARCH CENTER

BESC
BioEnergy Science Center

jbei
Joint BioEnergy Institute

Community science program

Plants



Fungi

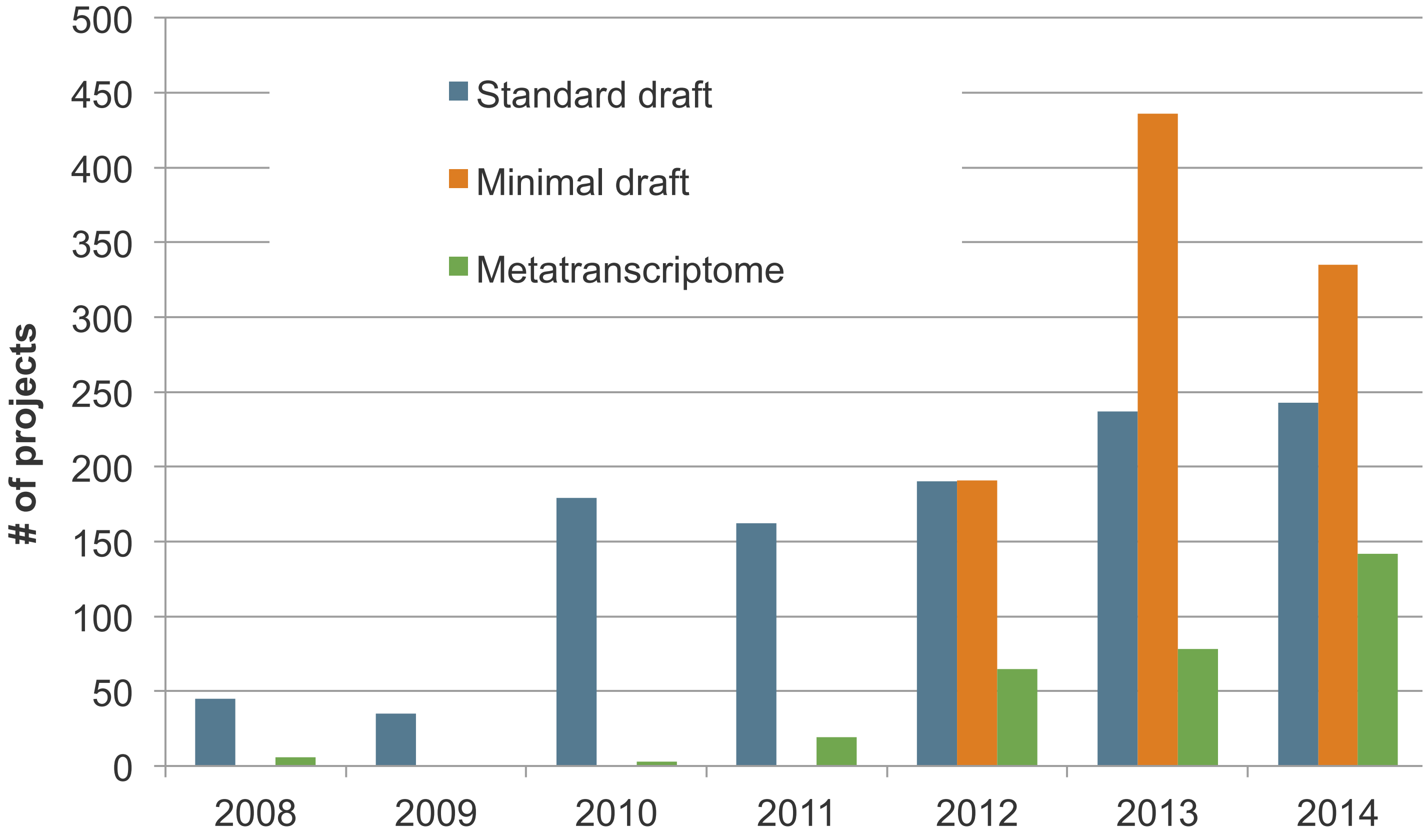
Microbes

Metagenomes

FY15 product menu

Product	Starting material	Sequencing	Analysis
16S iTags	DNA	184 samples per run, MiSeq 2X250 (~50,000 tags/sample)	OTU clustering and classification
Standard Metagenome	DNA	1-4 samples per lane, HiSeq 2X150 (15-70 Gb/sample)	Assembly, annotation in IMG/M-ER
Minimal Metagenome	DNA	6-12 samples per lane, HiSeq 2X150 (6-12 Gb/sample)	Assembly, annotation in IMG/M-ER
Metatranscriptome de novo	RNA	1-12 samples per lane, HiSeq 2X150 (5-70 Gb/sample)	Assembly, annotation in IMG/M-ER, mapping to metagenome if applicable

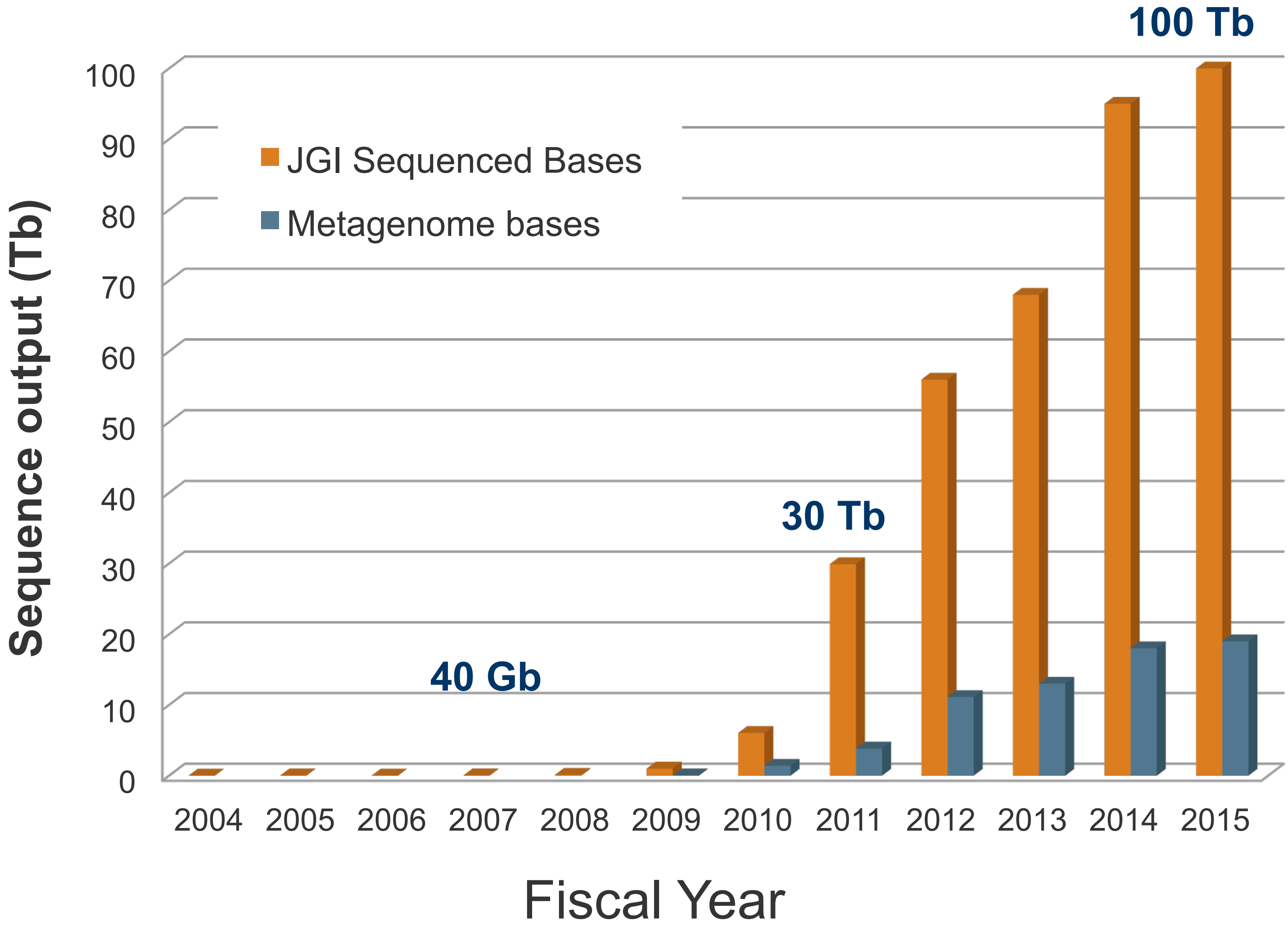
FY15 Metagenome program targets



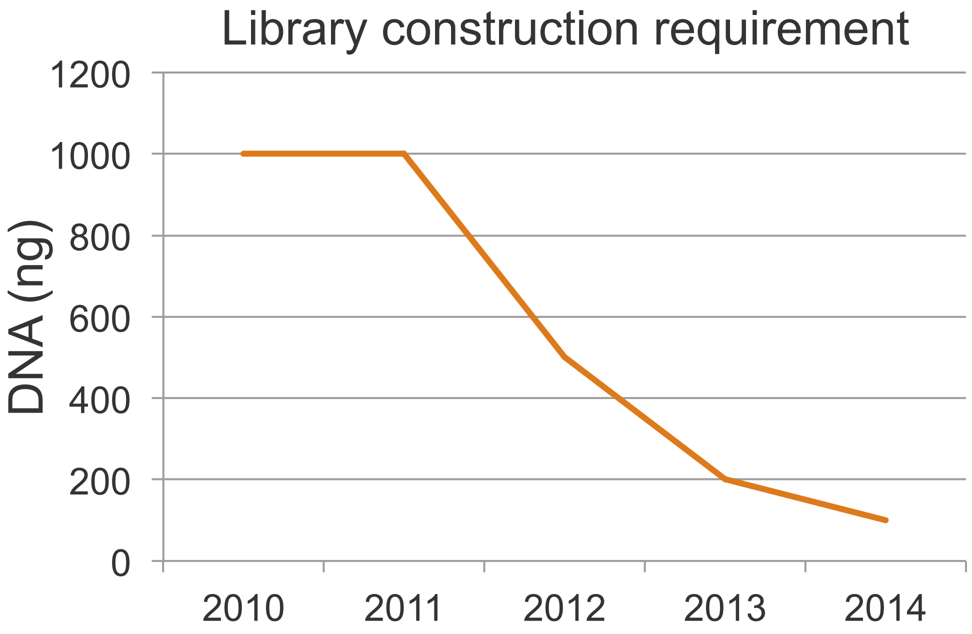
-Added higher-throughput multiplexed project type for metagenomes in 2012

-Increasing depth in FY15 using HiSeq 1T and HiSeq 2500

Metagenome sequence output



Metagenome library construction

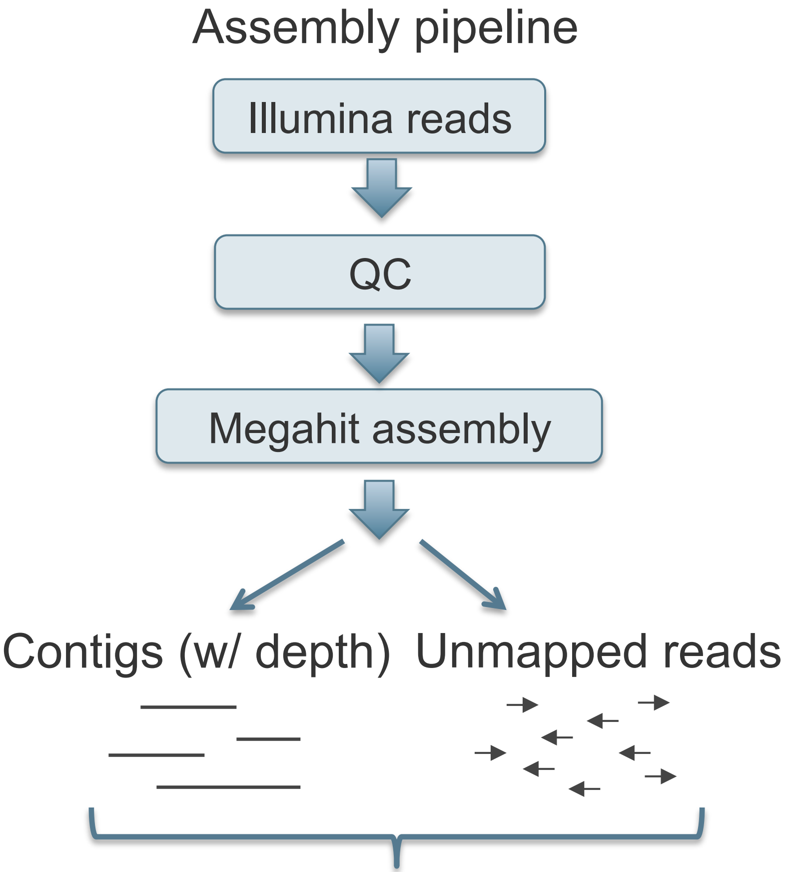


Nextera XT, Illumina TruSeq

Declining DNA quantity requirements enable sequencing of challenging samples:
-Deep ocean
-Individual soil particles
-BrdU-labeled DNA
-SIP-labeled DNA
-Upper troposphere

	16S iTag	MG minimal draft	MG standard draft
Volume	25 ul	50-100 ul	50-100 ul
Concentration	4-6 ng/ul	5-10 ng/ul	5-10 ng/ul
Mass	100-150 ng	0.5 ug	0.5 ug

Metagenome assembly



Assembly pipeline

Illumina reads

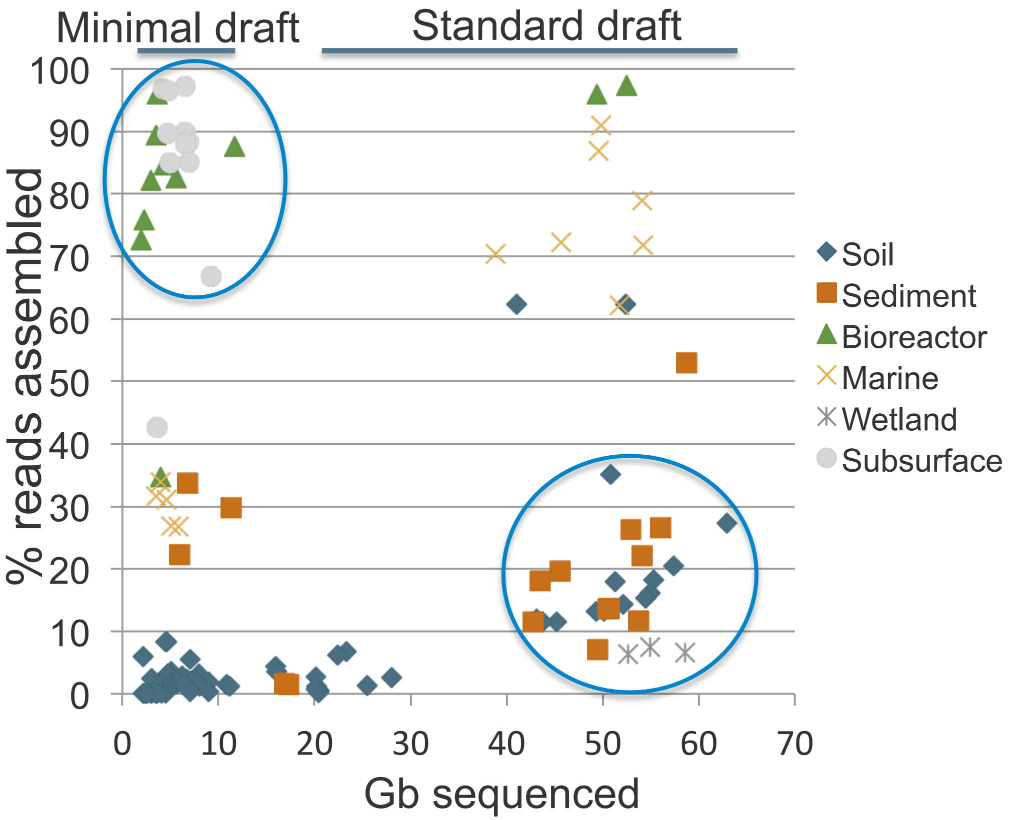
QC

Megahit assembly

Contigs (w/ depth)

Unmapped reads

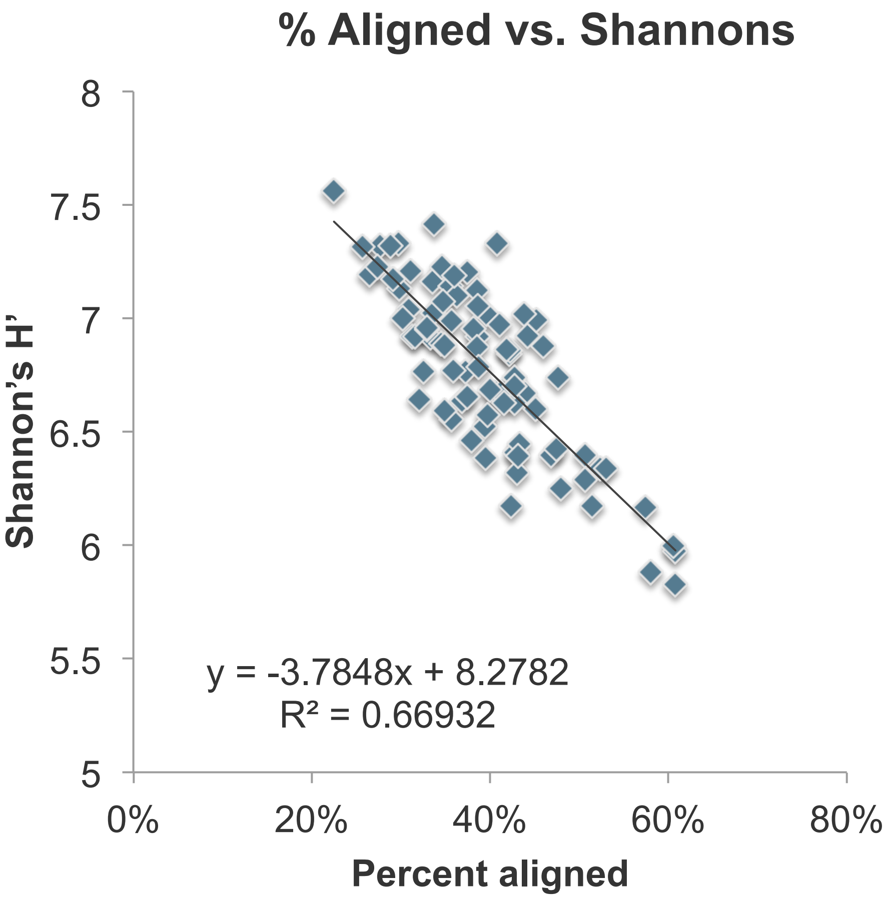
All data (assembled and unassembled) are annotated in IMG/M-ER for analysis

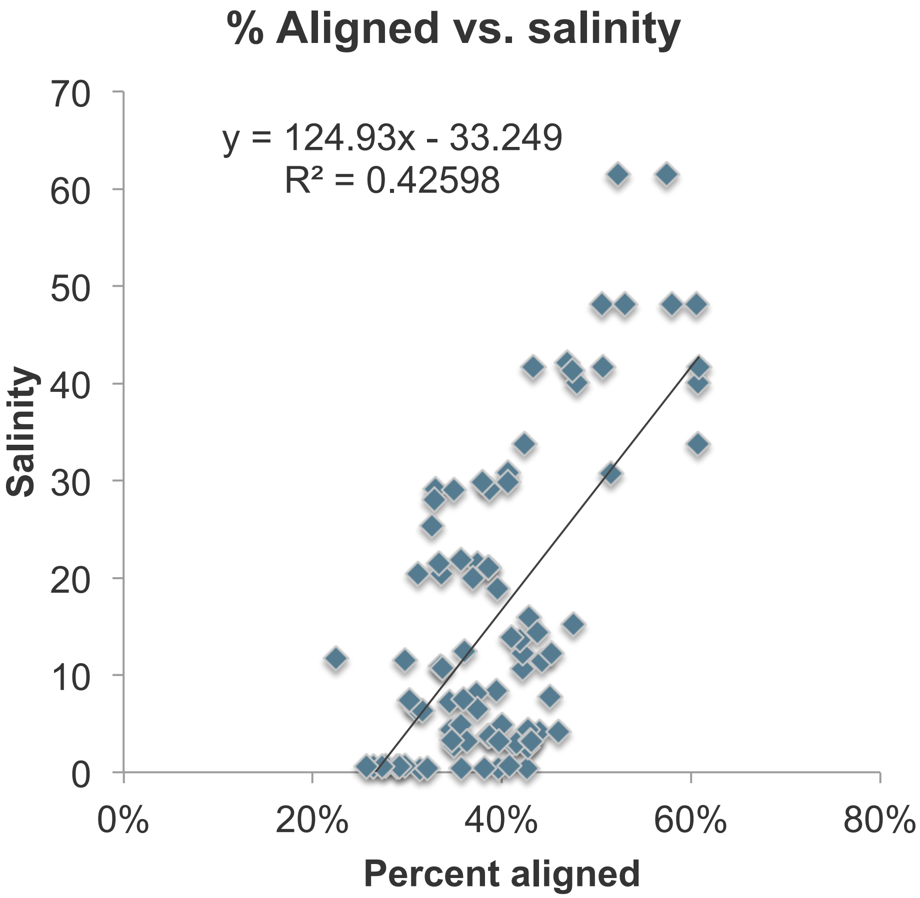


Simple communities: most data assemble but genomes are fragmented

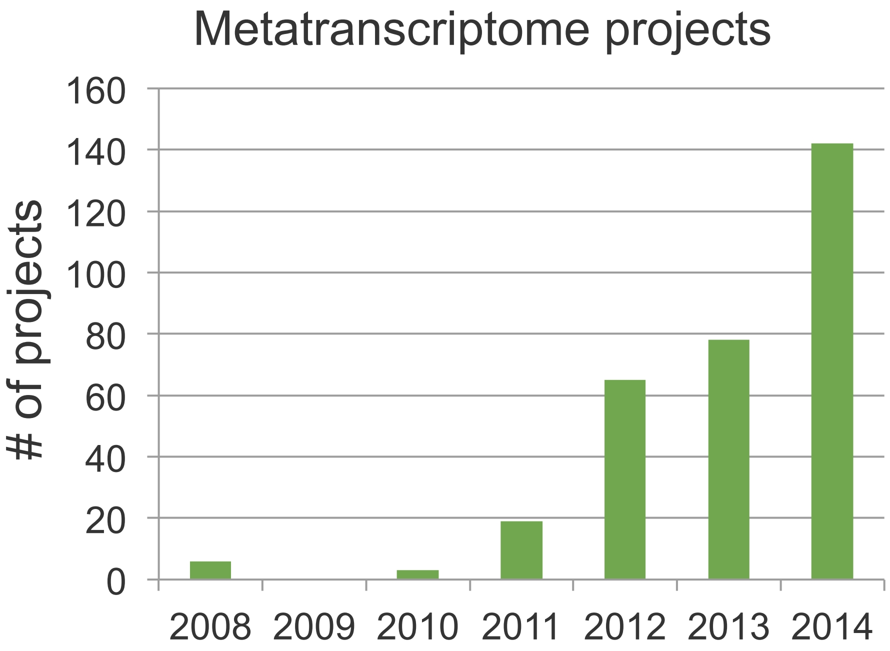
Soil and sediment: very poor assembly

Wetland metagenome assembly

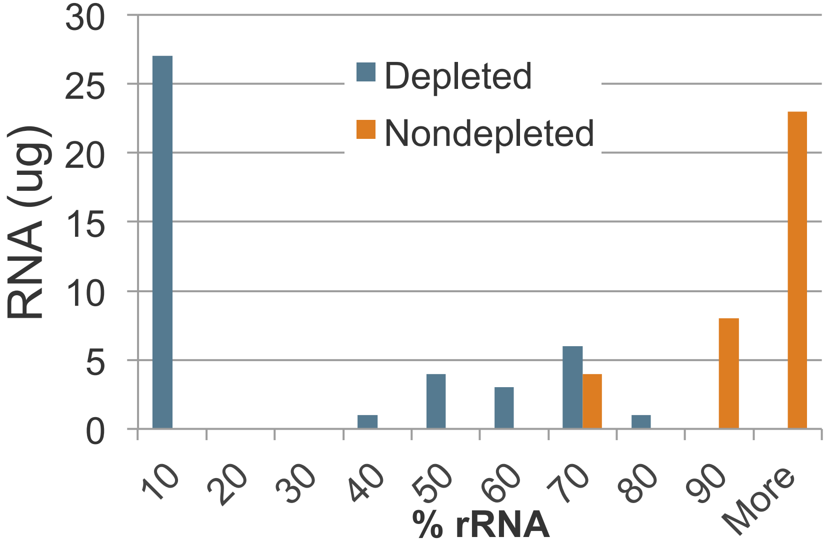


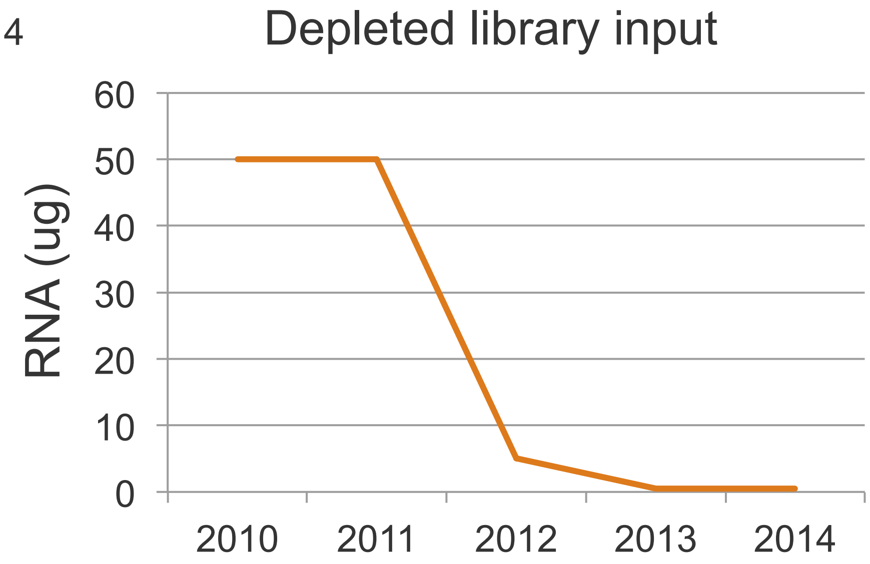


Metatranscriptome sequencing



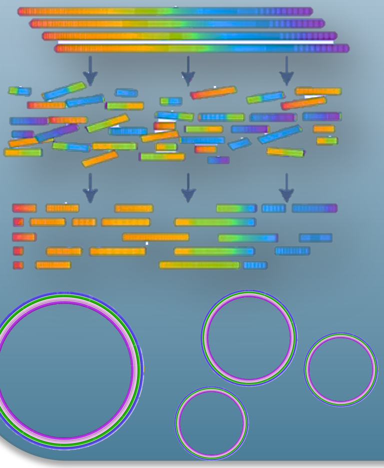
- ➢ Metatranscriptome sequencing has increased considerably but not to the level requested in CSP proposals
- ➢ Sample requirements have been limiting, especially for depleted libraries, but Ribozero kit and low-input protocols have changed the playing field





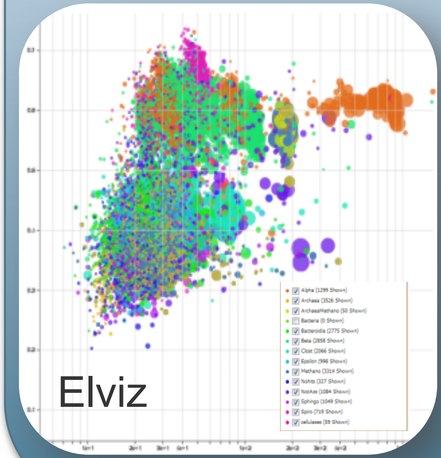
Metagenome strategic efforts

Metagenome assembly, Genome binning



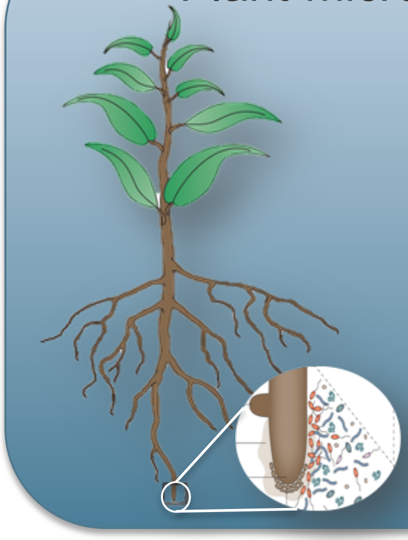
- ➢ Large-scale assembly
- ➢ Abundance-based binning

Data interpretation



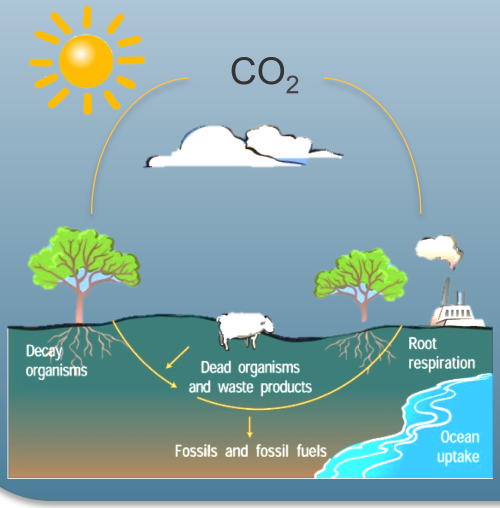
- ➢ Metagenome visualization
- ➢ Protein clusters
- ➢ OTU clusters
- ➢ PacBio full-length 16S

Plant-microbe interactions



- ➢ Arabidopsis rhizosphere,endosphere
- ➢ Poplar rhizosphere
- ➢ Maize rhizosphere
- ➢ Root-enhanced decomposition

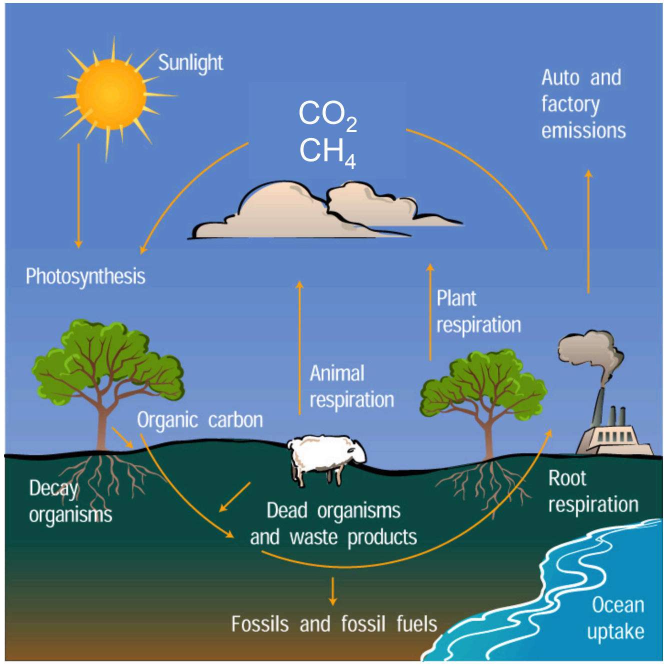
Terrestrial carbon cycling

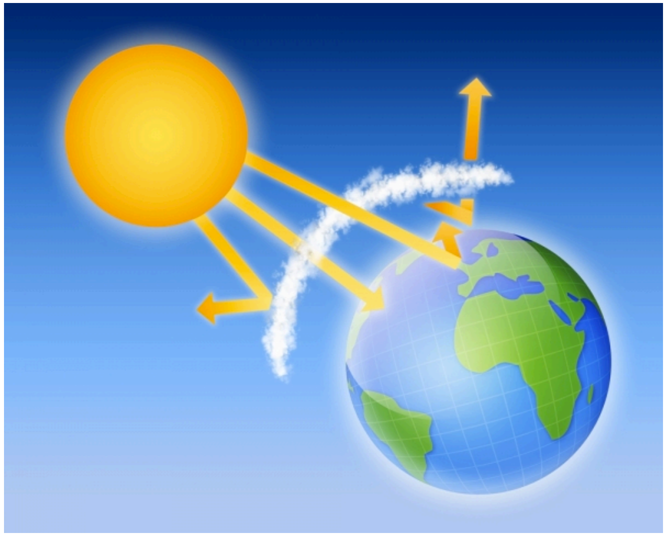


- ➢ Great Prairie Grand Challenge
- ➢ Bay/Delta wetlands
- ➢ Rangelands
- ➢ Other soils and peatlands

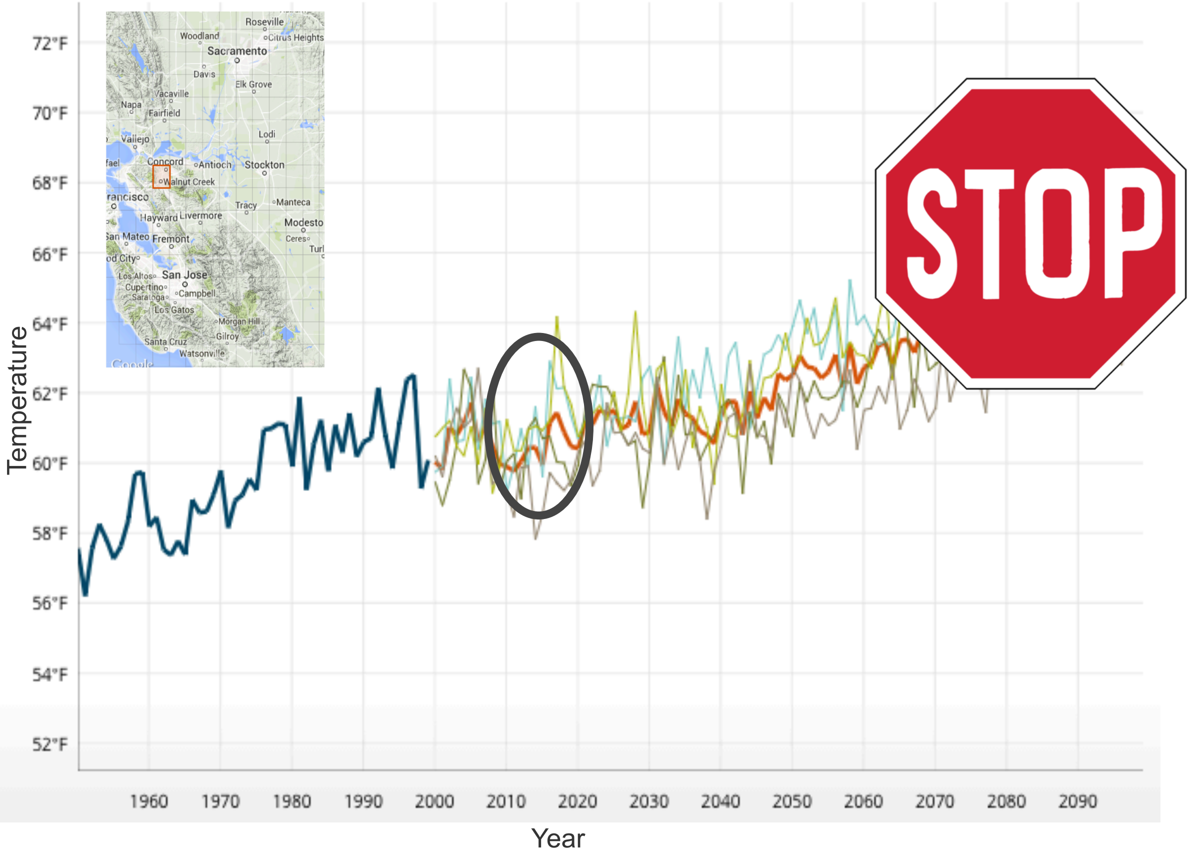
Philippot et al., 2013

Carbon cycle + global warming





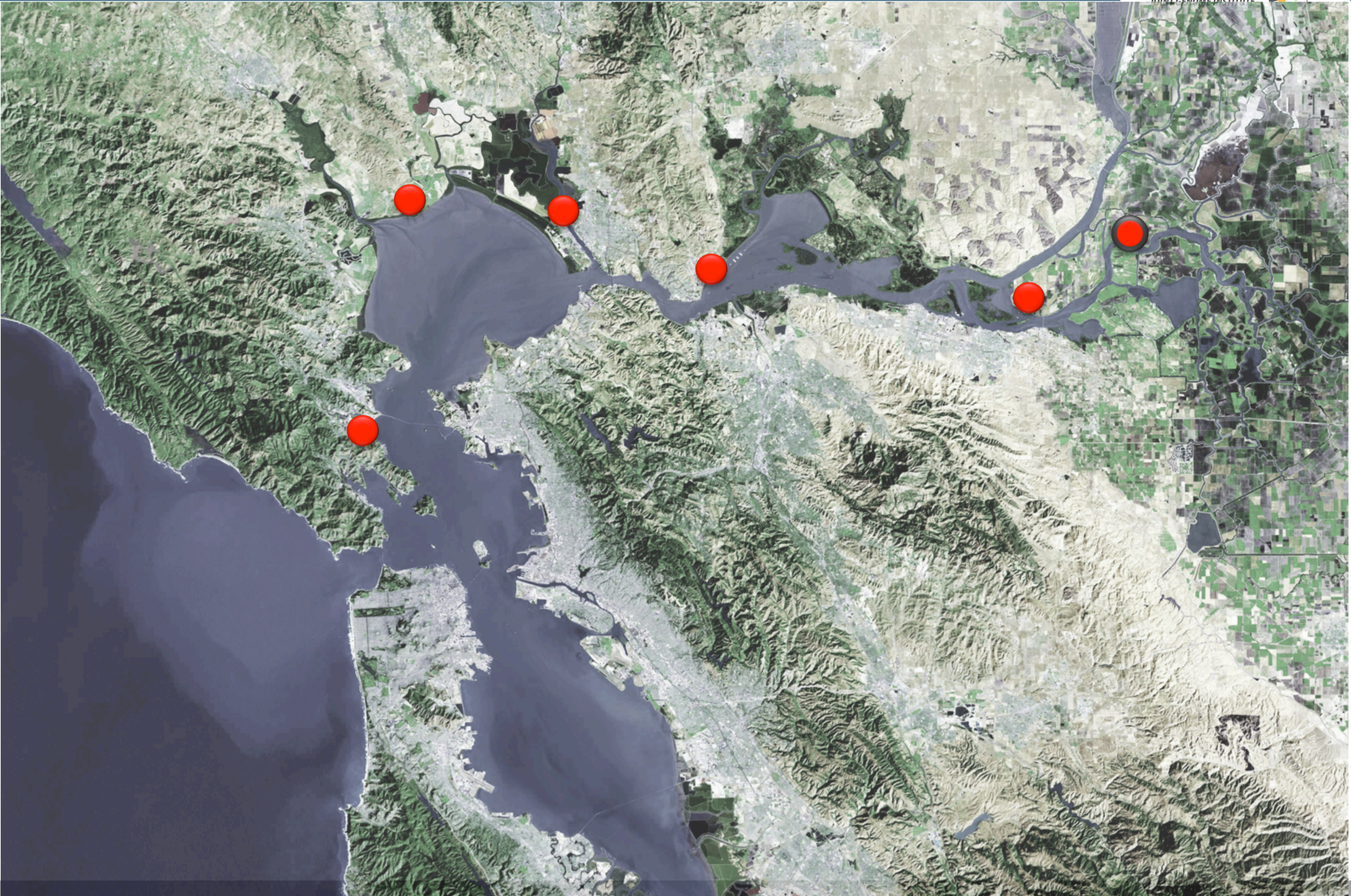
Average annual temperature Walnut Creek, CA



How can we sequester carbon and slow global warming?

cal-adapt.org

Wetland restoration



Wetland restoration

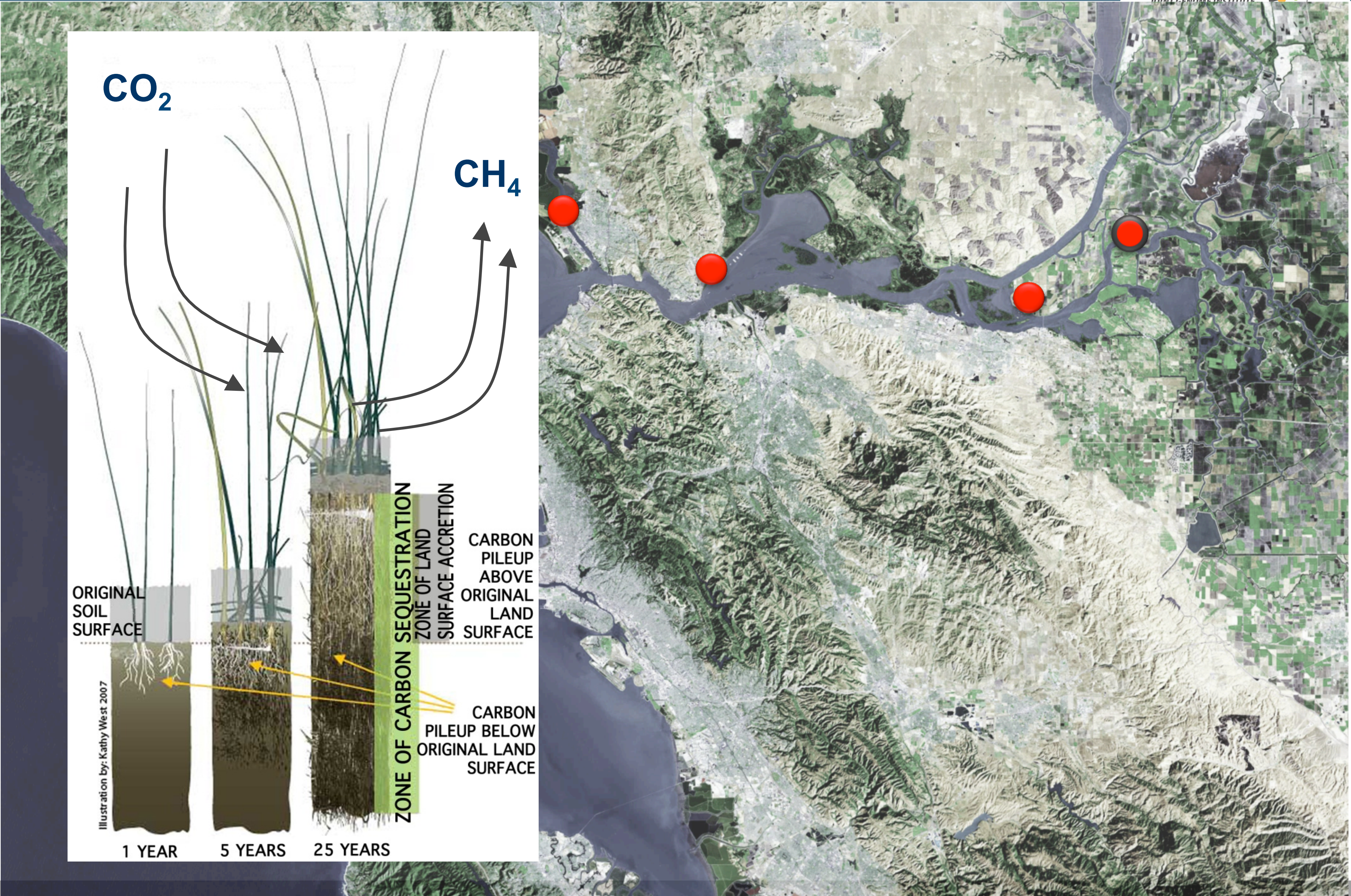
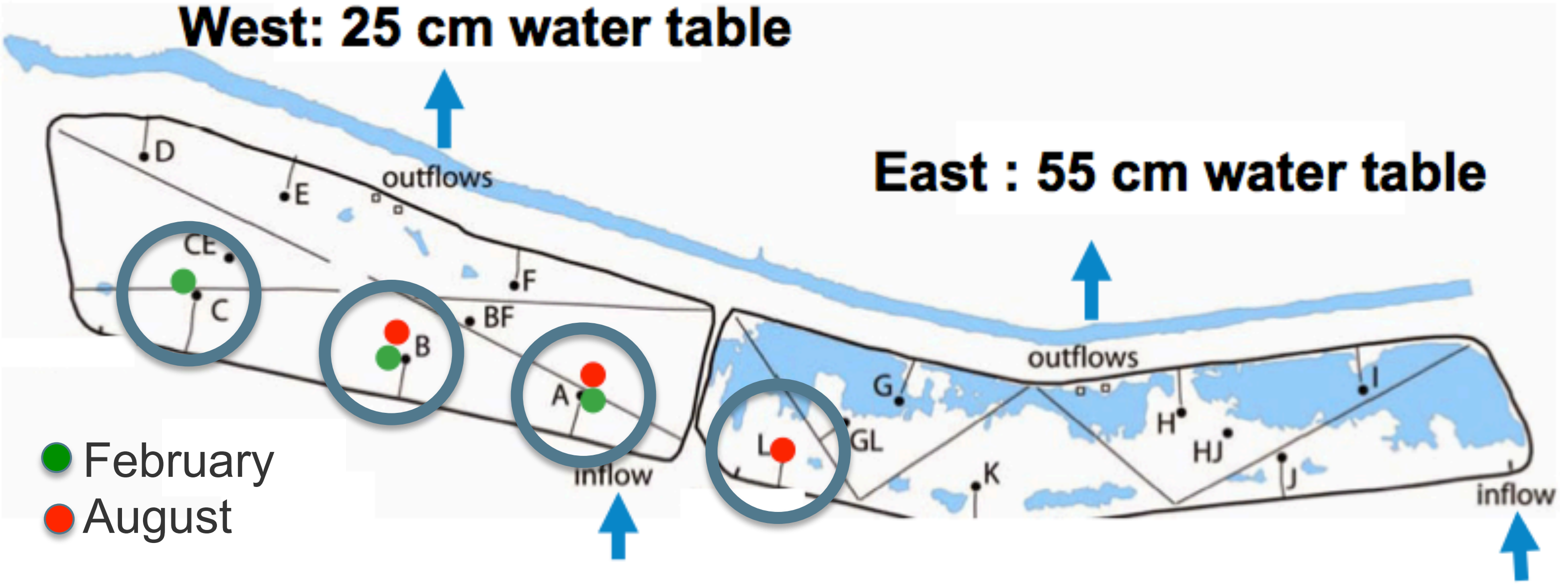


Illustration by: Kathy West 2007

Twitchell wetland

- Established 1997
- Can wetland restoration be a long term carbon storage solution?

Twitchell Wetland



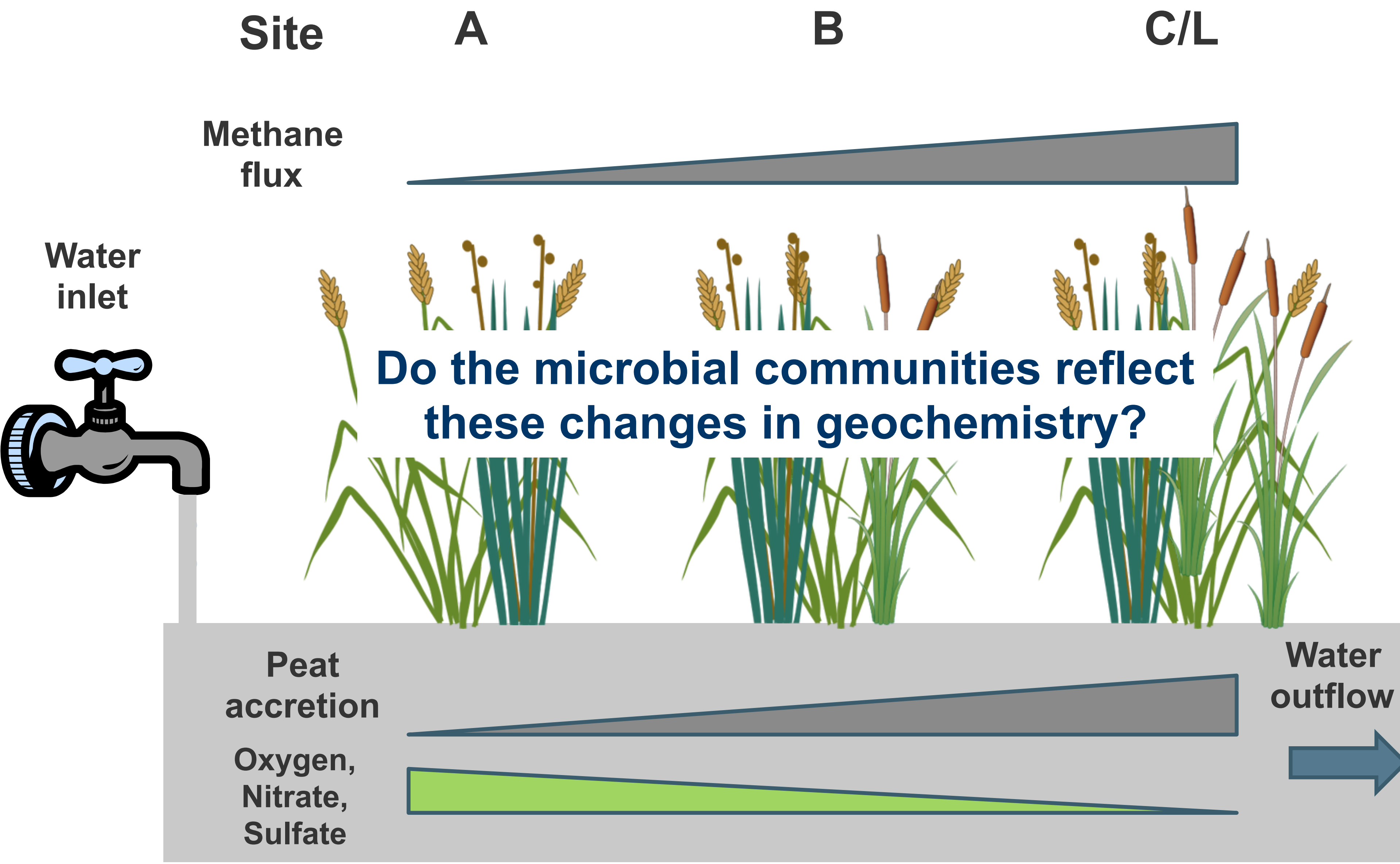
4 sites x 2 cores x 3 sample types x 2 seasons x 2 depths

= 96 samples

16S SSU rRNA gene sequencing

Shotgun Illumina HiSeq2000 libraries (2x150bp)

Sampling site gradients



Site A B C/L

Methane flux

Water inlet

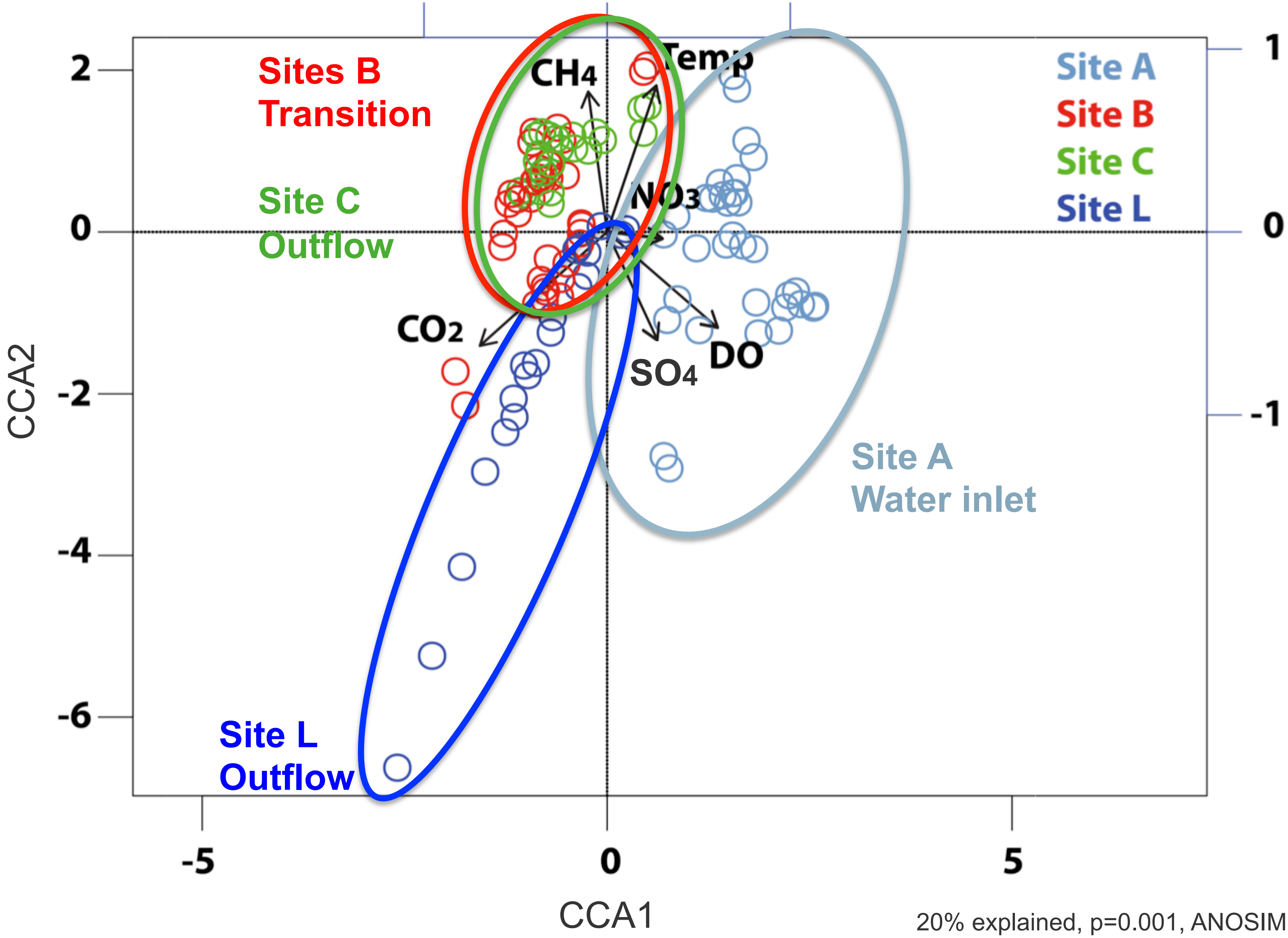
Do the microbial communities reflect these changes in geochemistry?

Peat accretion

Oxygen, Nitrate, Sulfate

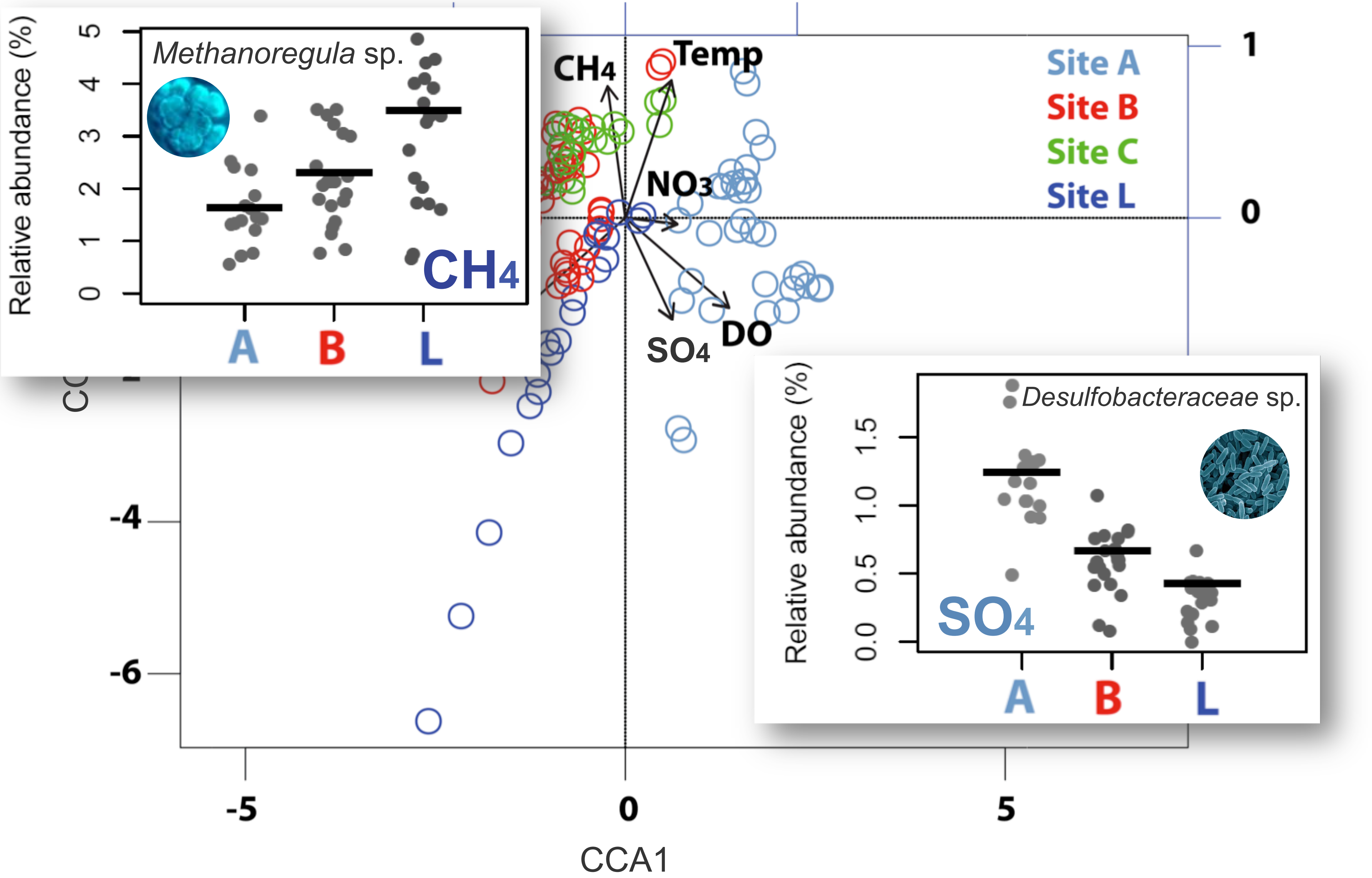
Water outflow

Bacteria + Archaea

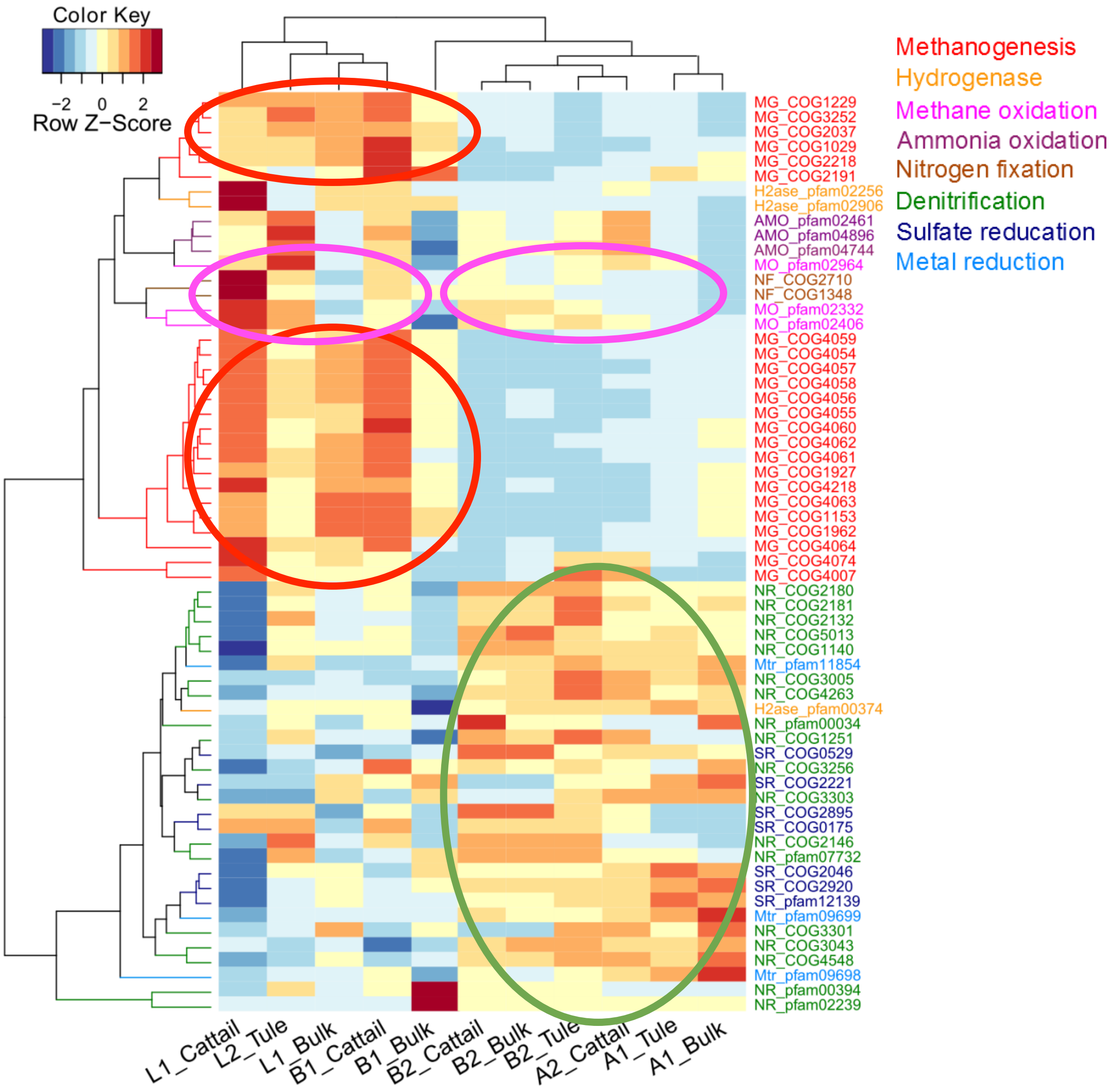


20% explained, p=0.001, ANOSIM = 0.3802

Site dictates microbial populations



Wetland gene family abundances

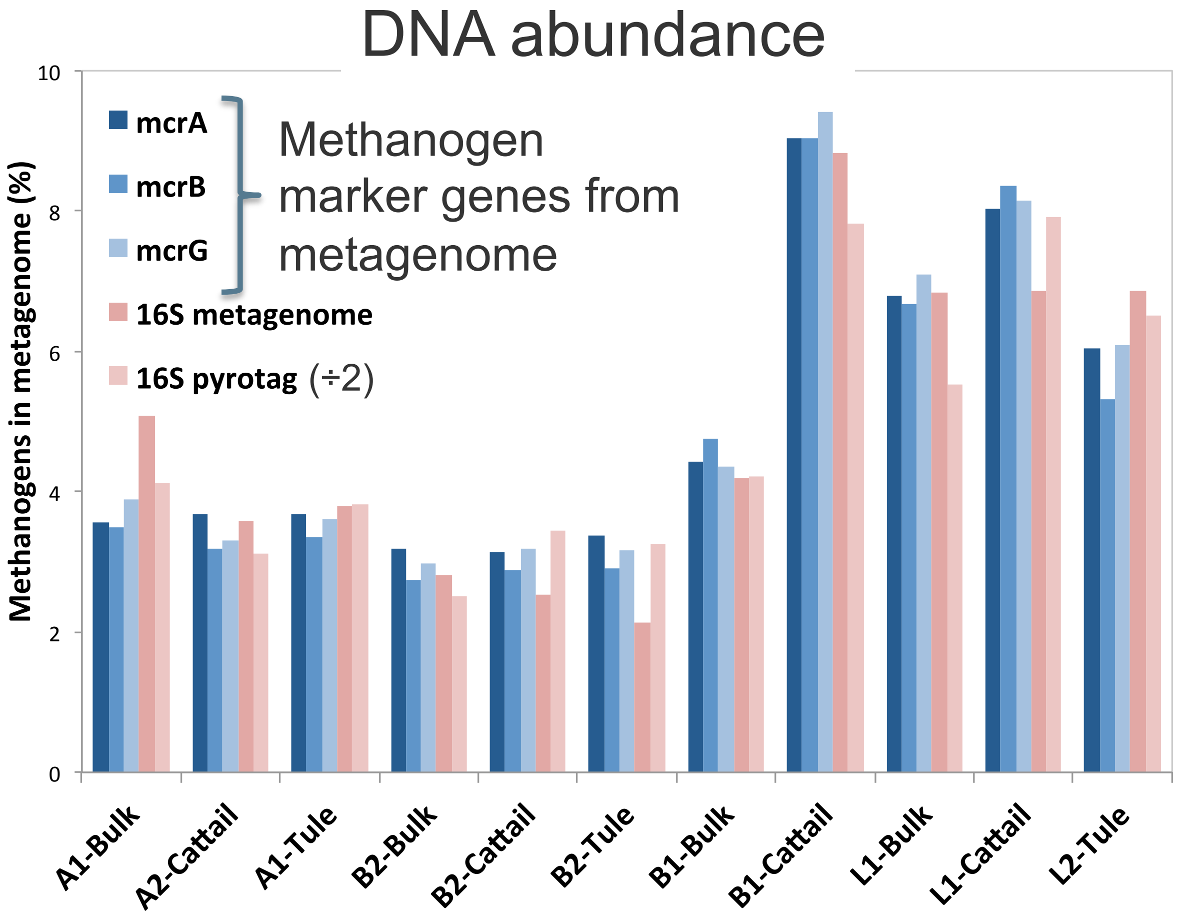


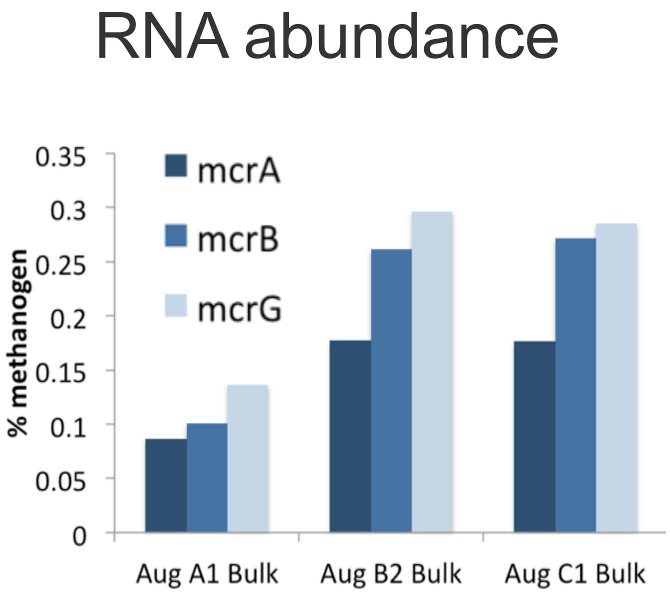
Samples with more methanogenesis genes have fewer genes in denitrification, dissimilatory sulfate reduction, and metal reduction.

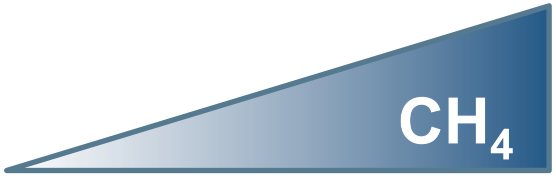
Methane oxidation genes were more abundant in rhizomes.

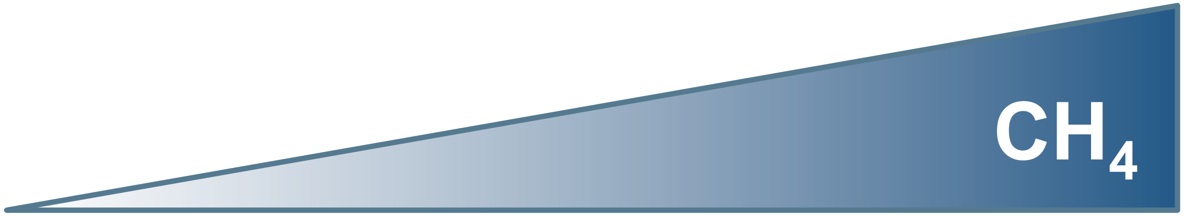
Shaomei He

Methanogen abundance



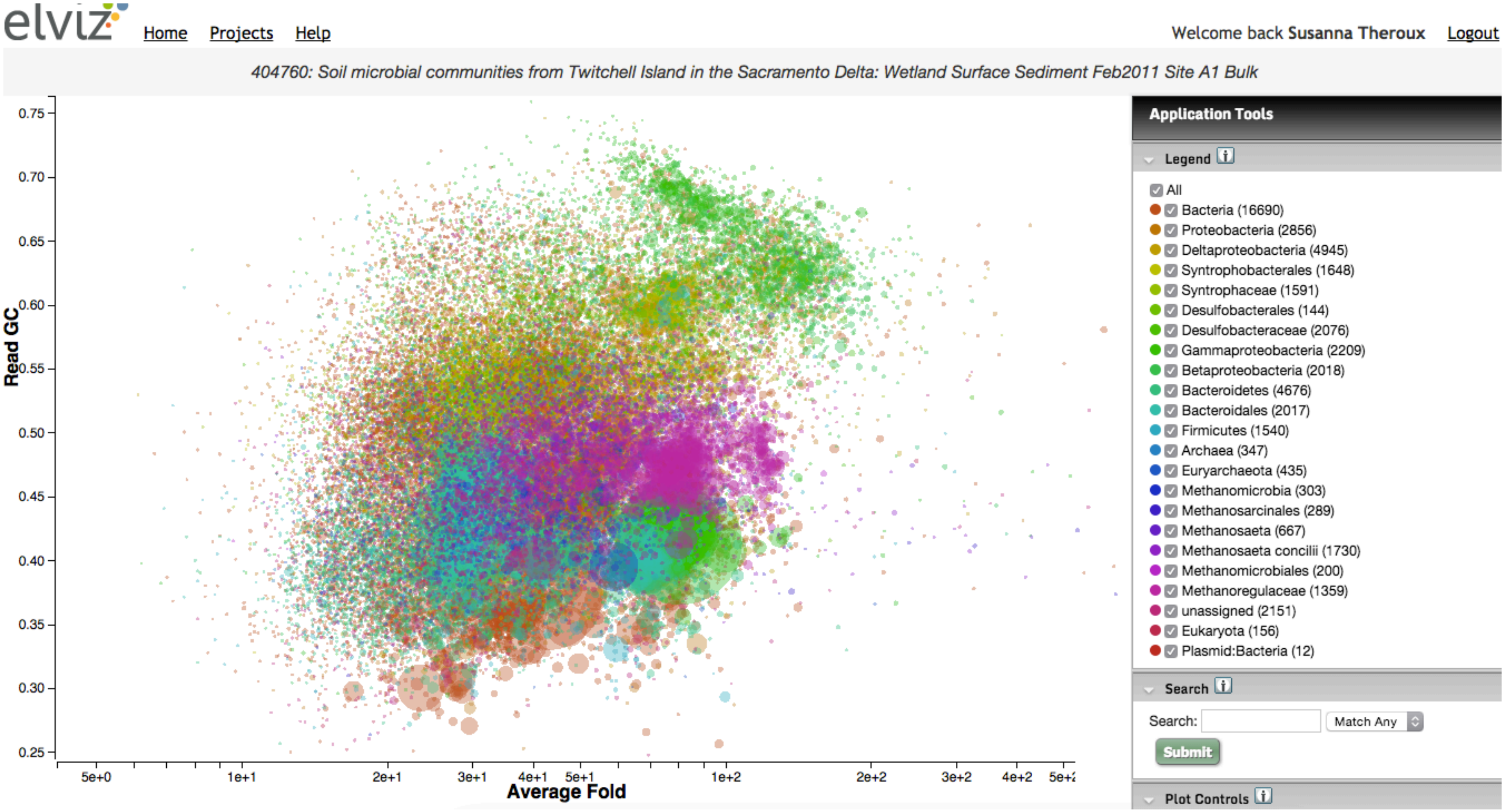






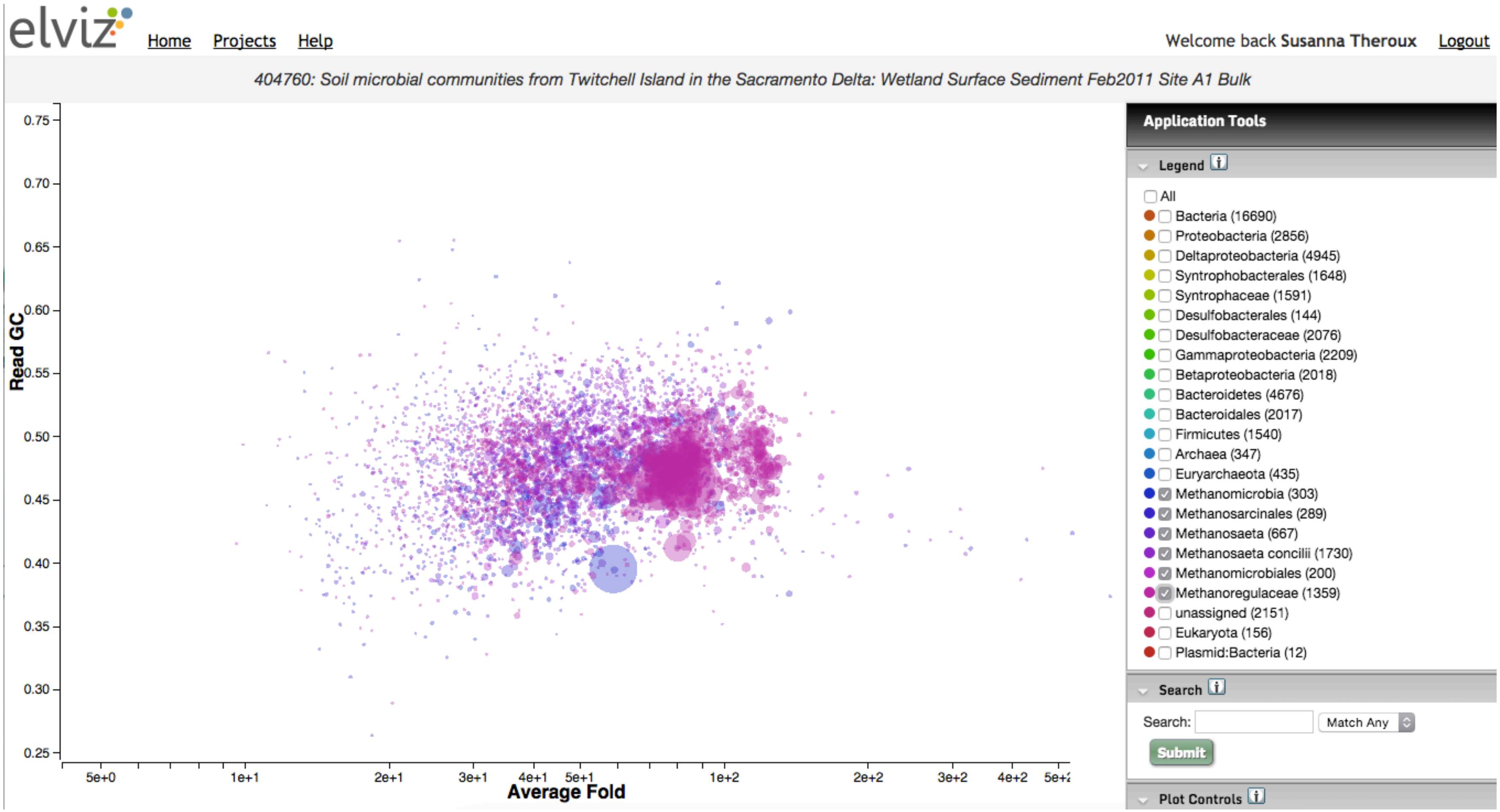
- Methane emissions correlated to methanogen ABUNDANCE and ACTIVITY

Assembly visualization



404760: Soil microbial communities from Twitchell Island in the Sacramento Delta: Wetland Surface Sediment Feb2011 Site A1 Bulk

Assembly visualization



elviz Home Projects Help

Welcome back **Susanna Theroux** Logout

404760: Soil microbial communities from Twitchell Island in the Sacramento Delta: Wetland Surface Sediment Feb2011 Site A1 Bulk

Application Tools

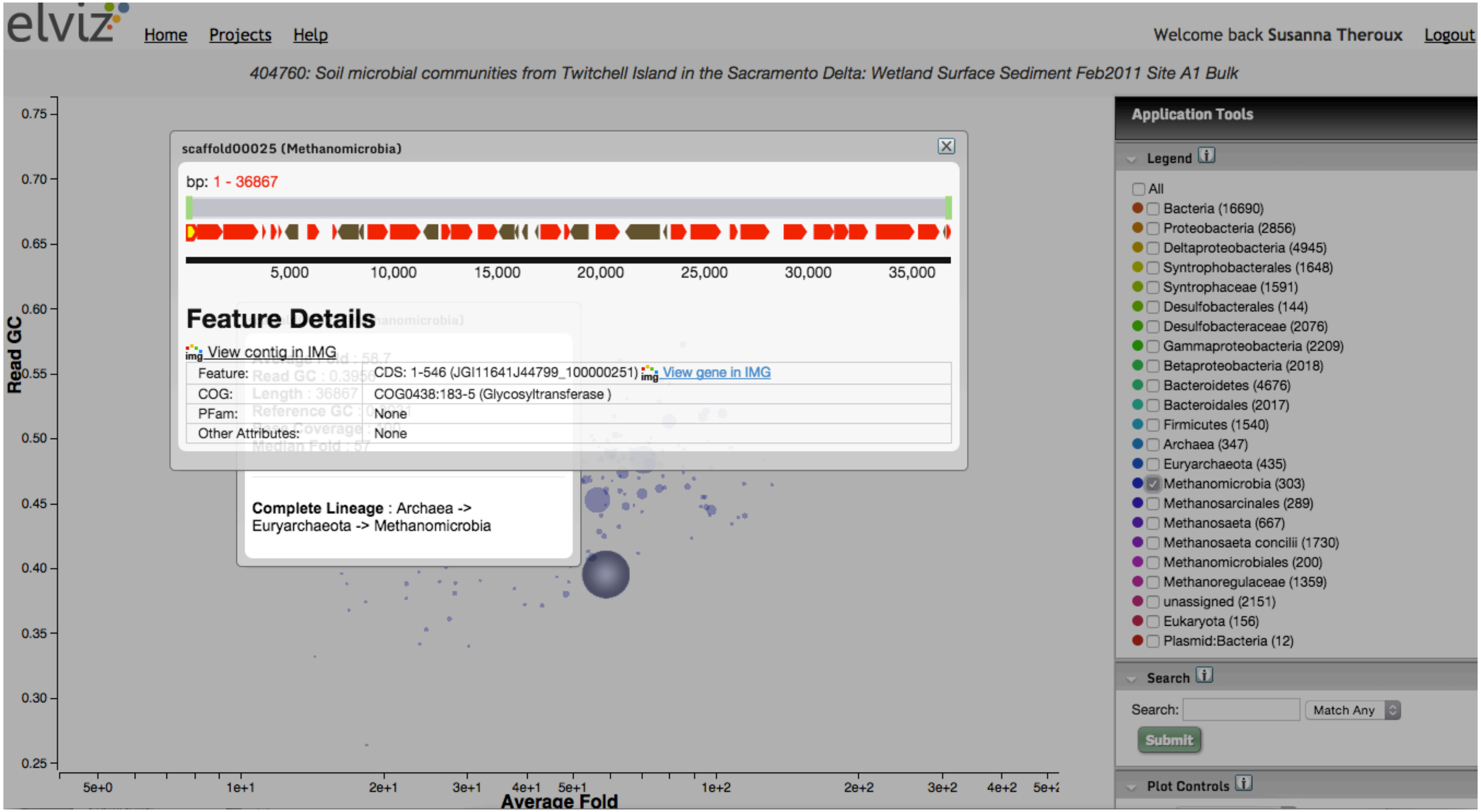
Legend

☐ All
● ☐ Bacteria (16690)
● ☐ Proteobacteria (2856)
● ☐ Deltaproteobacteria (4945)
● ☐ Syntrophobacterales (1648)
● ☐ Syntrophaceae (1591)
● ☐ Desulfobacterales (144)
● ☐ Desulfobacteraceae (2076)
● ☐ Gammaproteobacteria (2209)
● ☐ Betaproteobacteria (2018)
● ☐ Bacteroidetes (4676)
● ☐ Bacteroidales (2017)
● ☐ Firmicutes (1540)
● ☐ Archaea (347)
● ☐ Euryarchaeota (435)
● ☑ Methanomicrobia (303)
● ☑ Methanosarcinales (289)
● ☑ Methanosaeta (667)
● ☑ Methanosaeta concilii (1730)
● ☑ Methanomicrobiales (200)
● ☑ Methanoregulaceae (1359)
● ☐ unassigned (2151)
● ☐ Eukaryota (156)
● ☐ Plasmid:Bacteria (12)

Search

Search: Match Any

Assembly visualization



Welcome back **Susanna Theroux**

404760: Soil microbial communities from Twitchell Island in the Sacramento Delta: Wetland Surface Sediment Feb2011 Site A1 Bulk

scaffold00025 (Methanomicrobia)

bp: 1 - 36867

Feature Details

View contig in IMG

Feature:	CDS: 1-546 (JGI11641J44799_100000251) View gene in IMG
COG:	COG0438:183-5 (Glycosyltransferase)
PFam:	None
Other Attributes:	None

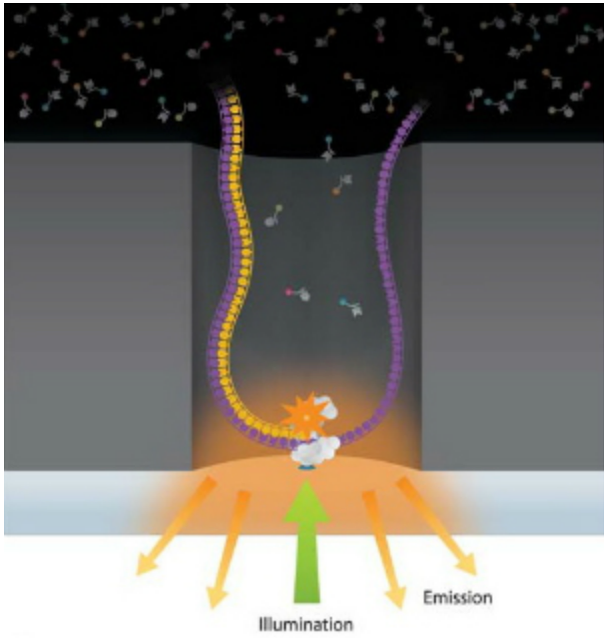
Complete Lineage : Archaea -> Euryarchaeota -> Methanomicrobia

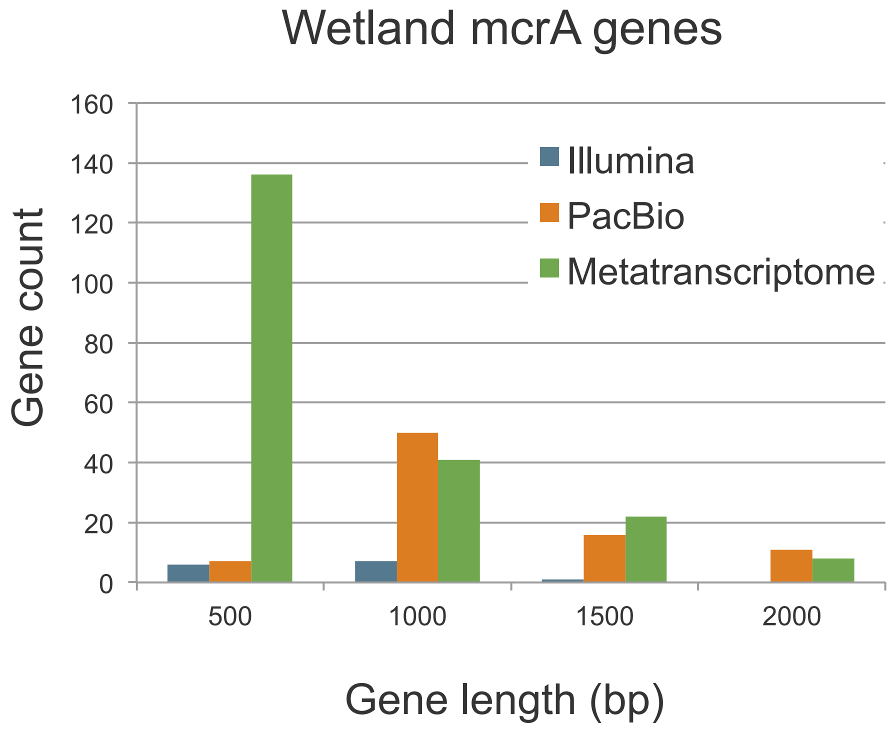
☐ All
- ● ☐ Bacteria (16690)
- ● ☐ Proteobacteria (2856)
- ● ☐ Deltaproteobacteria (4945)
- ● ☐ Syntrophobacterales (1648)
- ● ☐ Syntrophaceae (1591)
- ● ☐ Desulfobacterales (144)
- ● ☐ Desulfobacteraceae (2076)
- ● ☐ Gammaproteobacteria (2209)
- ● ☐ Betaproteobacteria (2018)
- ● ☐ Bacteroidetes (4676)
- ● ☐ Bacteroidales (2017)
- ● ☐ Firmicutes (1540)
- ● ☐ Archaea (347)
- ● ☐ Euryarchaeota (435)
- ● ☑ Methanomicrobia (303)
- ● ☐ Methanosarcinales (289)
- ● ☐ Methanosaeta (667)
- ● ☐ Methanosaeta concilii (1730)
- ● ☐ Methanomicrobiales (200)
- ● ☐ Methanoregulaceae (1359)
- ● ☐ unassigned (2151)
- ● ☐ Eukaryota (156)
- ● ☐ Plasmid:Bacteria (12)

Search: Match Any

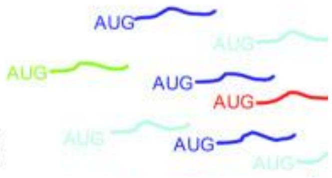
Improving gene assembly

Pacific Biosciences





Metatranscriptomics



Alternative sequencing approaches, including long-read sequencing and metatranscriptome sequencing, can improve gene (and possibly genome) contiguity

Combined metagenome assembly

24 Metagenomes
~50GB each

↓

Assembly
(Ray-META +
Megahit)

↓

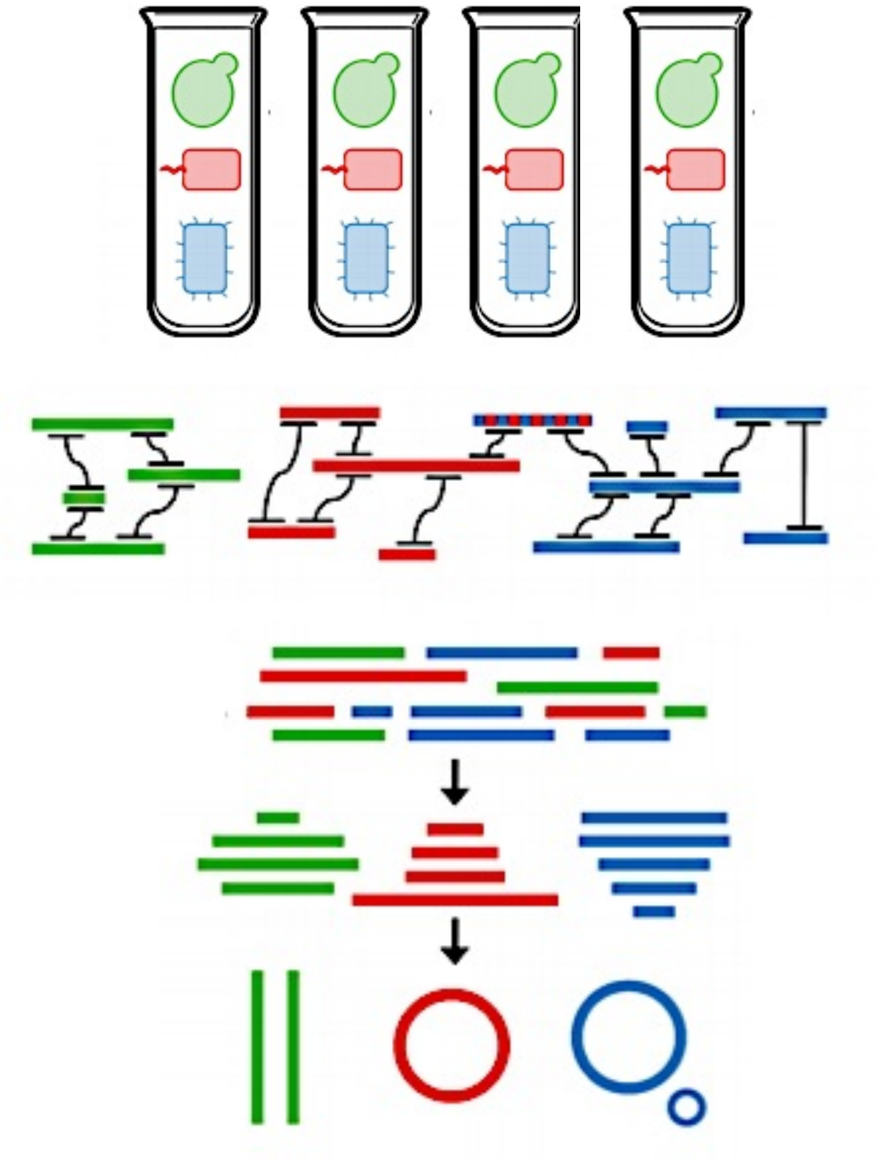
Read mapping
(BBMap)

↓

Genome binning
(MetaBAT)

↓

Bin Validation

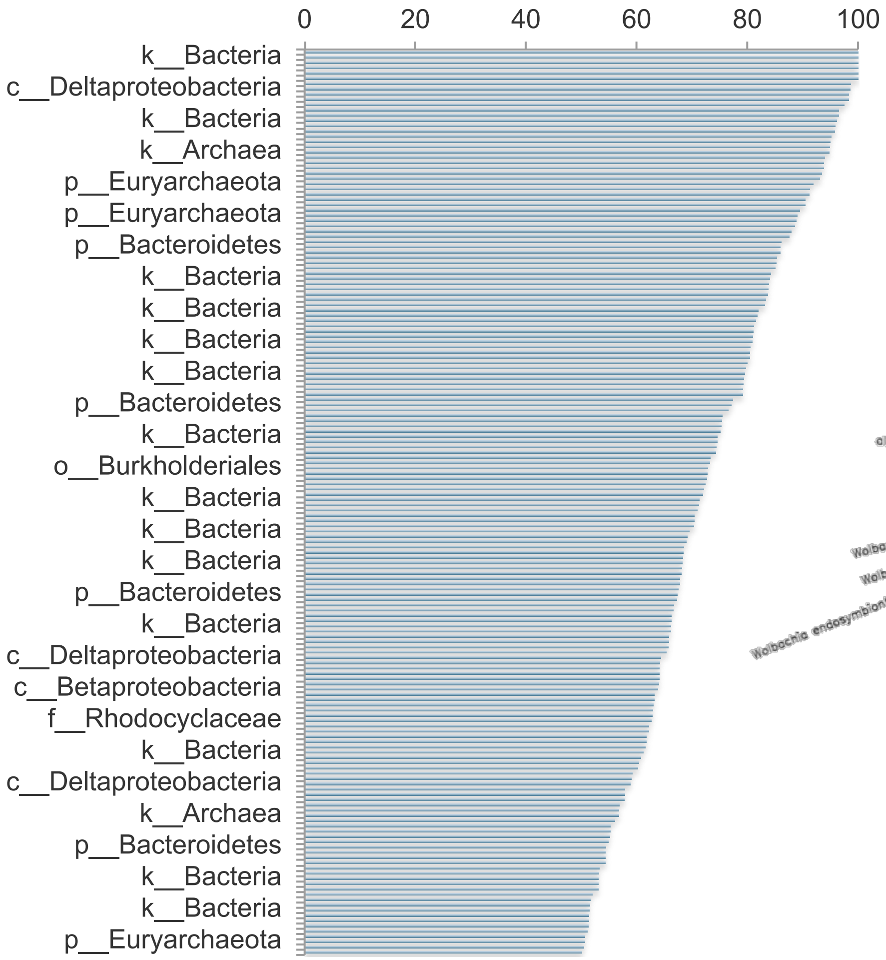


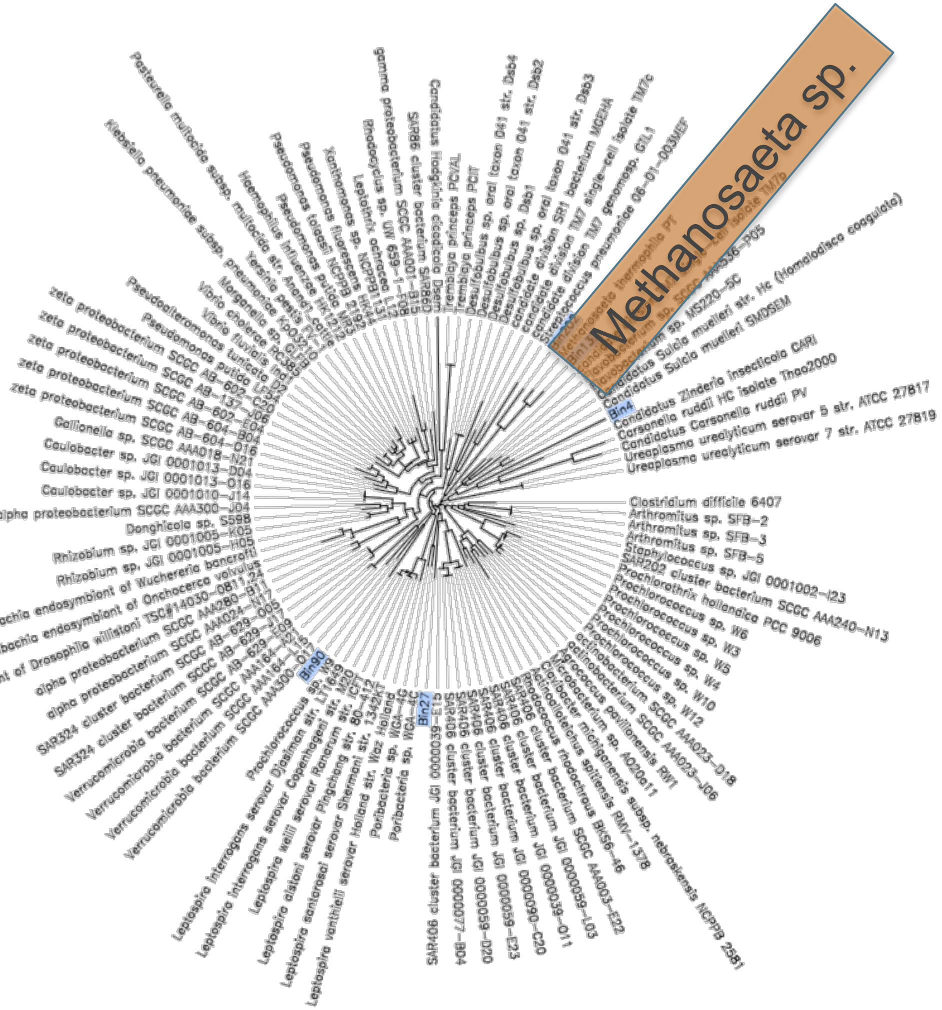
Burton et al., 2014

MetaBAT Genome Bins

400 genome bins average 2 GB in size

Bin completeness (%)



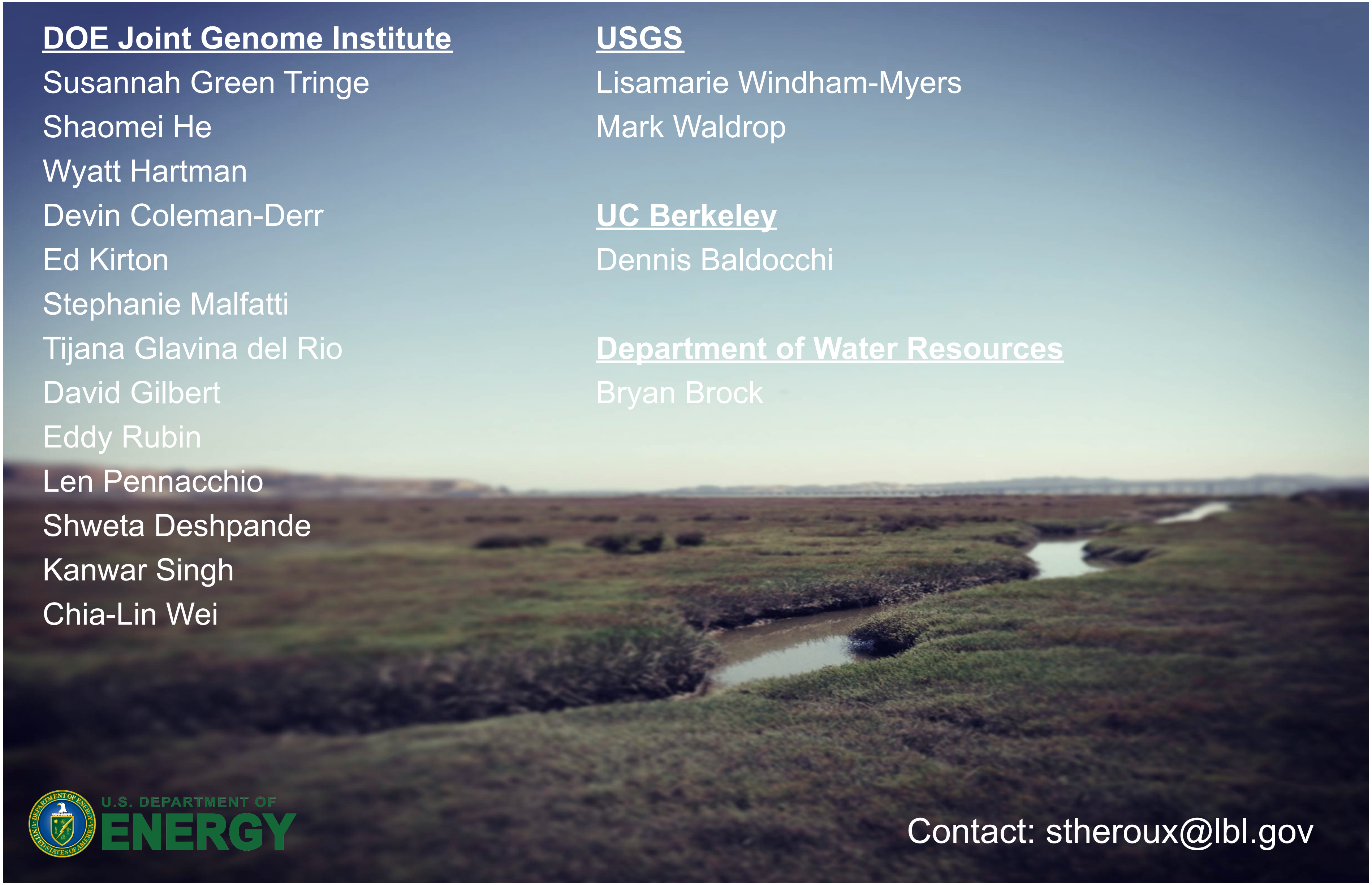


CheckM, KBase

Conclusions / Visions

- Robust, high-throughput sequencing has dramatically changed microbial community studies, enabling deeper and broader inquiry – but many studies are still sequence-limited!
- Microbial community studies are generating data at increasingly finer spatial and temporal scales, necessitating high-throughput library construction from small quantities of nucleic acid
- These studies are generating diverse data types – metagenomes, metatranscriptomes, single cell genomes, targeted metagenomes – that demand meaningful integration
- Experimental and computational tools for improving annotation of metagenome data are greatly needed

Acknowledgements



DOE Joint Genome Institute
Susannah Green Tringe
Shaomei He
Wyatt Hartman
Devin Coleman-Derr
Ed Kirton
Stephanie Malfatti
Tijana Glavina del Rio
David Gilbert
Eddy Rubin
Len Pennacchio
Shweta Deshpande
Kanwar Singh
Chia-Lin Wei

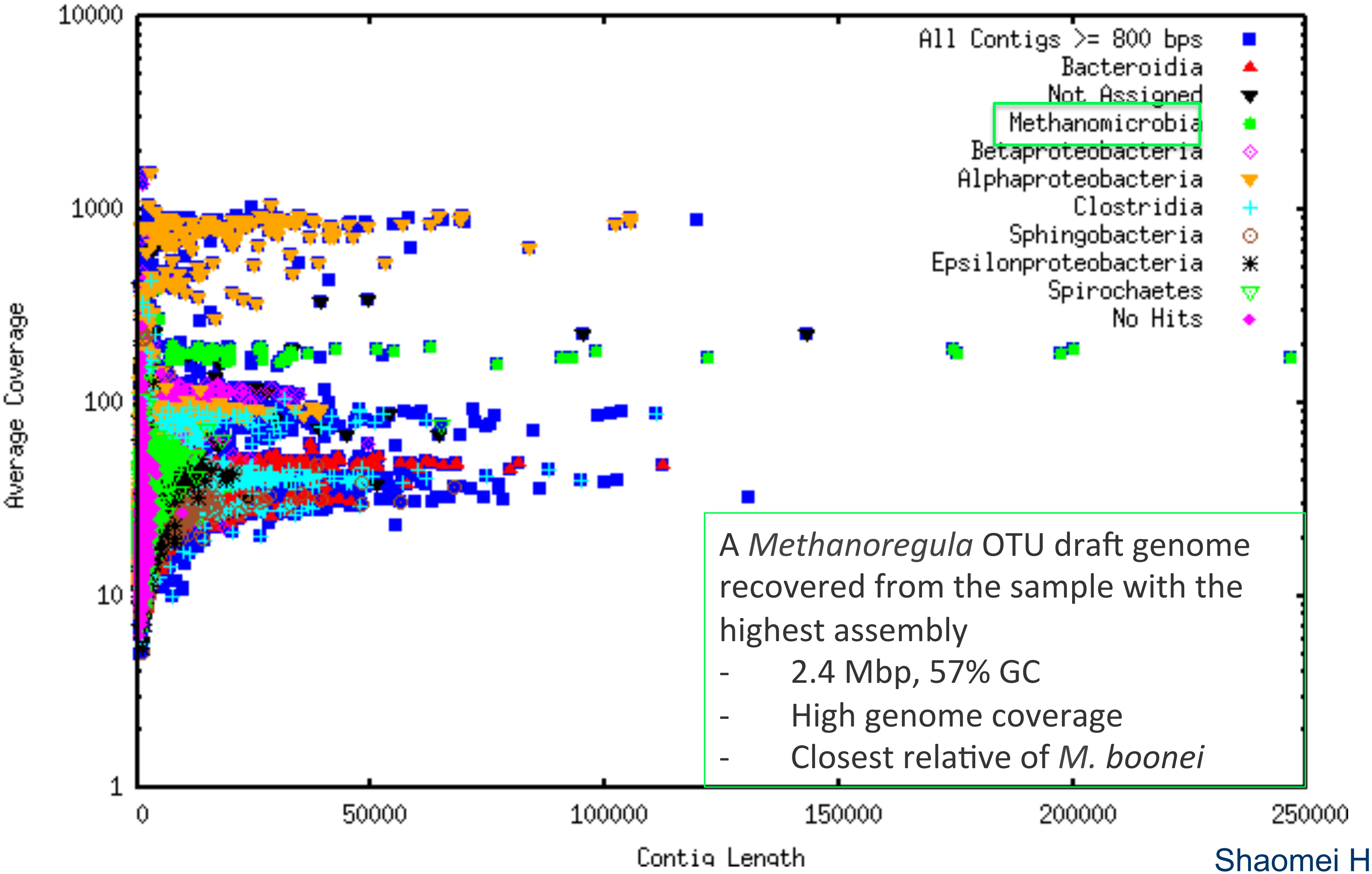
USGS
Lisamarie Windham-Myers
Mark Waldrop

UC Berkeley
Dennis Baldocchi

Department of Water Resources
Bryan Brock

Contact: stheroux@lbl.gov

Methanoregula draft genome

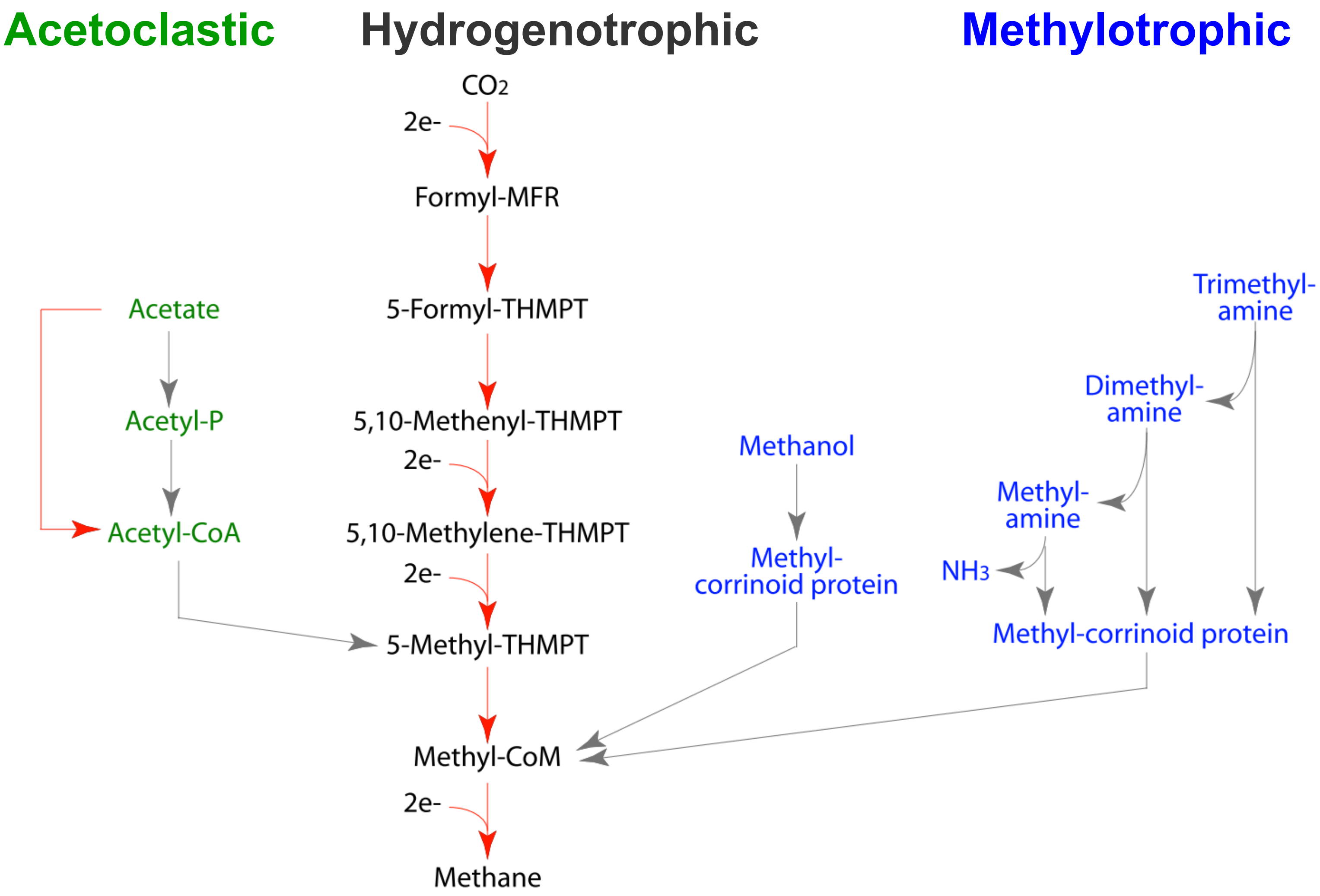


A *Methanoregula* OTU draft genome recovered from the sample with the highest assembly

- 2.4 Mbp, 57% GC
- High genome coverage
- Closest relative of *M. boonei*

Shaomei He

Methanoregula draft genome



Shaomei He

Statistics

Y 2014 Overall Sequencing Progress, Updated Quarterly

Quarter	Total Bases (Billions)			Operating Hours		
	Goal	Actual Total*	Actual % of Goal	Goal (hours)**	Actual Total	Actual % Goal
Q1 2014	15,000	18.827	126%	2,164	2208	102%
Q2 2014	17,000			2,117		
Q3 2014	18,000			2,140		
Q4 2014	18,000			2,164		
FY 2014 Total	**68,000**	**18.827**	**28%**	**8,585**	**2208**	**26%**

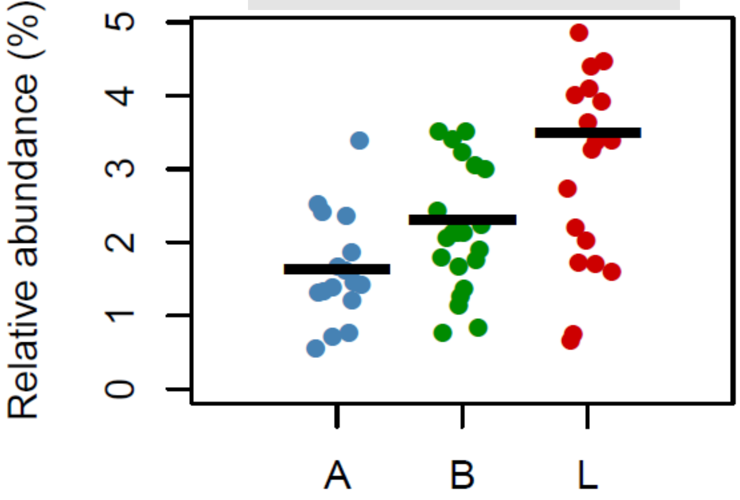
* Includes Illumina HiSeq, MiSeq and PacBio sequencing platforms.
** Operating Hour target is based on 98% of the total available hours.

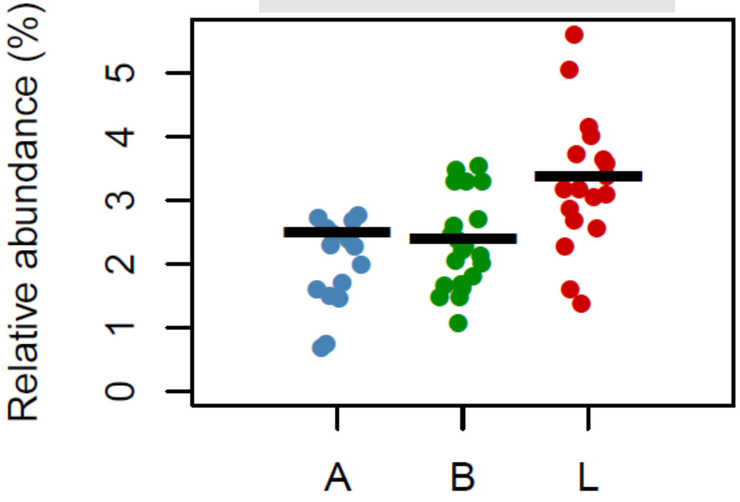
FY 2013 Overall Sequencing Progress, Updated Quarterly

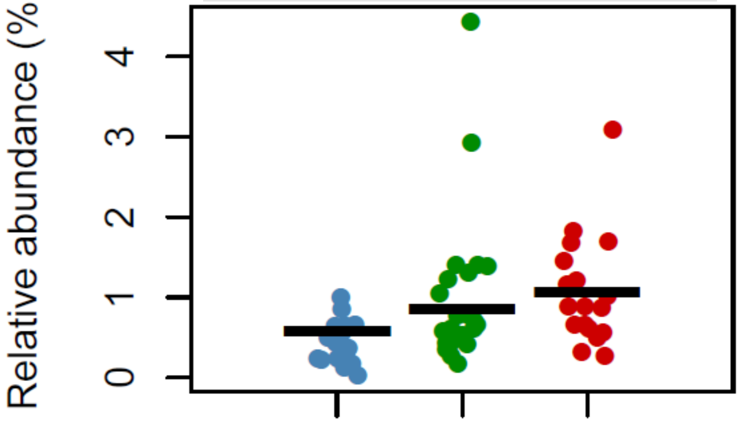
Quarter	Total Bases (Billions)			Operating Hours		
	Goal	Actual Total*	Actual % of Goal	Goal (hours)**	Actual Total	Actual % Goal
Q1 2013	15,000	20,004	133%	2,164	2,208	102%
Q2 2013	15,000	18,882	126%	2,117	2,160	102%
Q3 2013	18,000	13,904	77%	2,140	2,184	102%
Q4 2013	18,000	18,073	100%	2,164	2,208	102%
FY 2013 Total	**66,000**	**70,863**	**107%**	**8,585**	**8,760**	**102%**

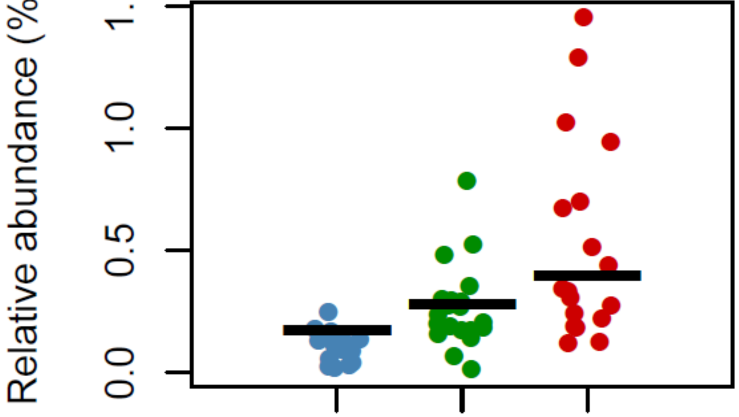
* Includes Illumina HiSeq, MiSeq and PacBio sequencing platforms.
** Operating Hour target is based on 98% of the total available hours.

Indicator OTUs for Site

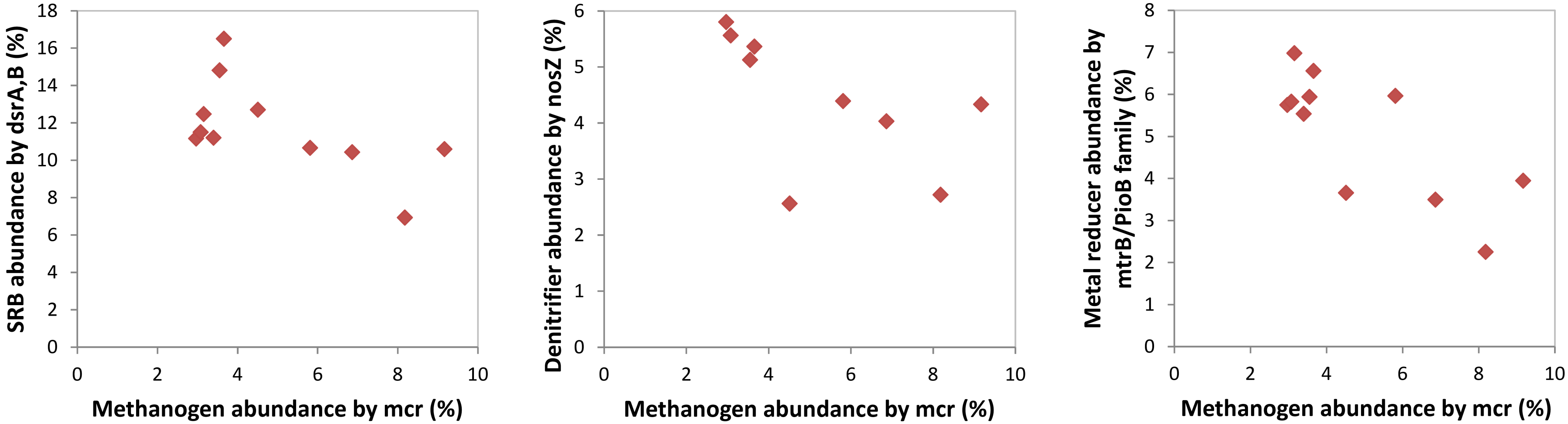

A *Methanoregula* OTU


A *Methanosaeta* OTU


A *Methanobacterium* OTU


A *Methanosarcina* OTU

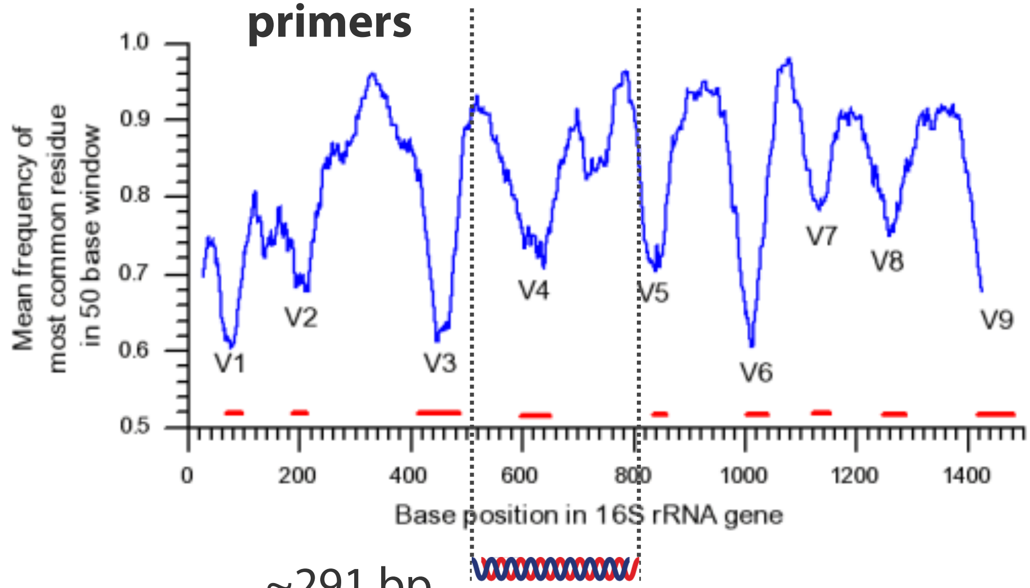
Multiple lineages of archaeal methanogens are more abundant at high-methane-production site L



Relative abundances of sulfate reducing prokaryotes estimated by dissimilatory sulfite reductase (*dsrAB*) (**a**), denitrifiers by nitrous oxide reductase (*nosZ*) (**b**) and metal reducers by multi-haem c-type cytochromes MtrB/PioB family (**c**) show negative correlations to methanogen abundance estimated by methyl-CoM reductase (*mcr*).

16S sequencing

V4 primers 515F 806R

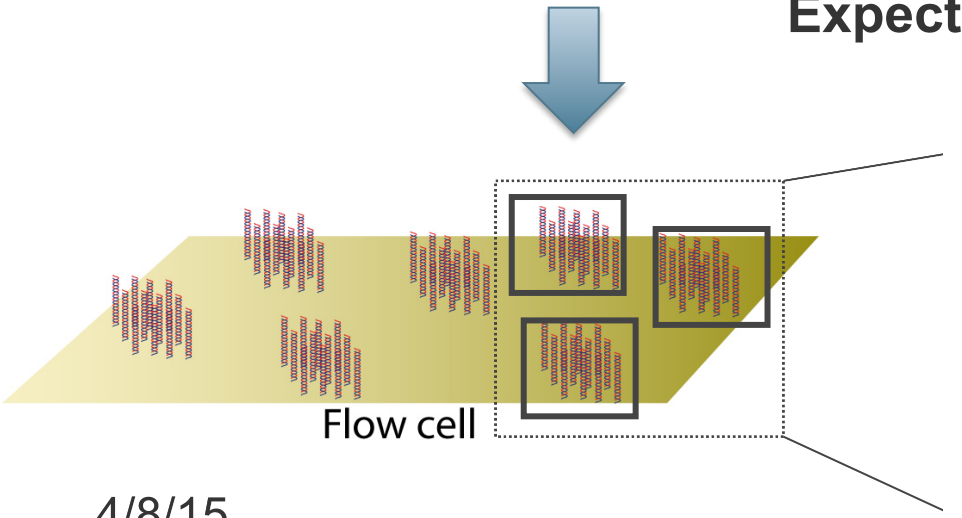


~291 bp

-Paired-end 250 bp sequencing on "staggered" amplicons produces high quality sequence – but short tags are still hard to resolve phylogenetically

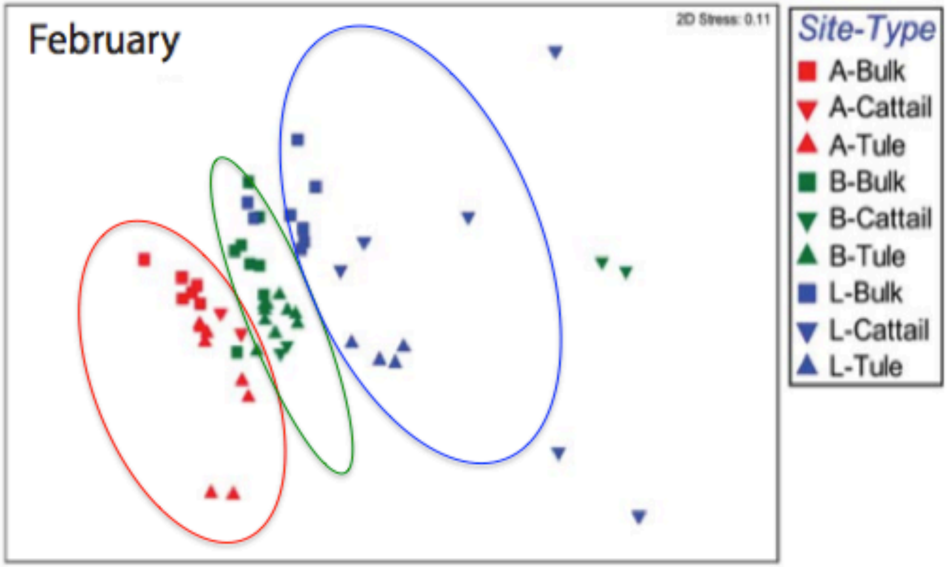
-Working on full-length 16S seqs via PacBio to accurately classify novel organisms

Expected FY14 throughput: 11,000 samples

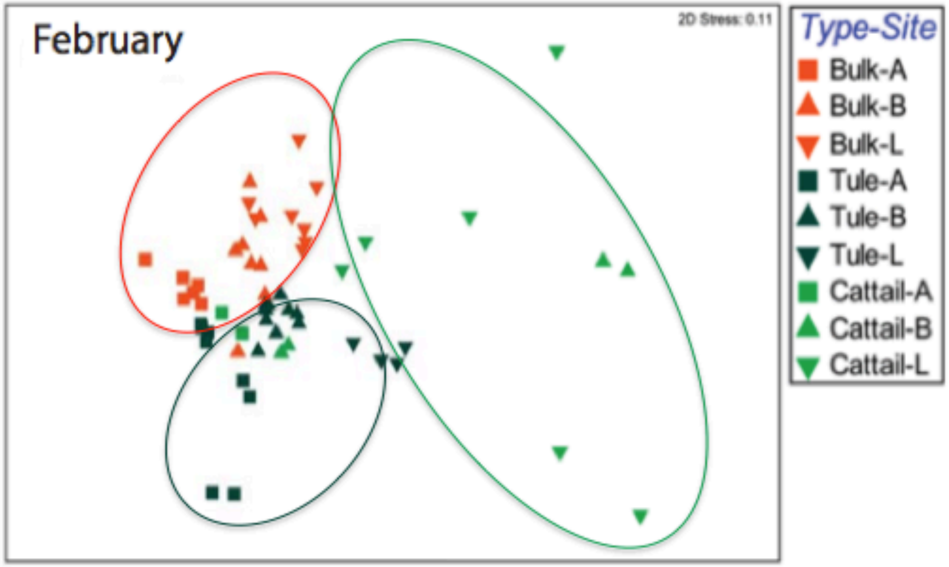


```
GTGCCAGCMGCCGCGGTAA
NGTGCCAGCMGCCGCGGTAA
NNGTGCCAGCMGCCGCGGTAA
NNNGTGCCAGCMGCCGCGGTAA
NGTGCCAGCMGCCGCGGTAA
.
.
```

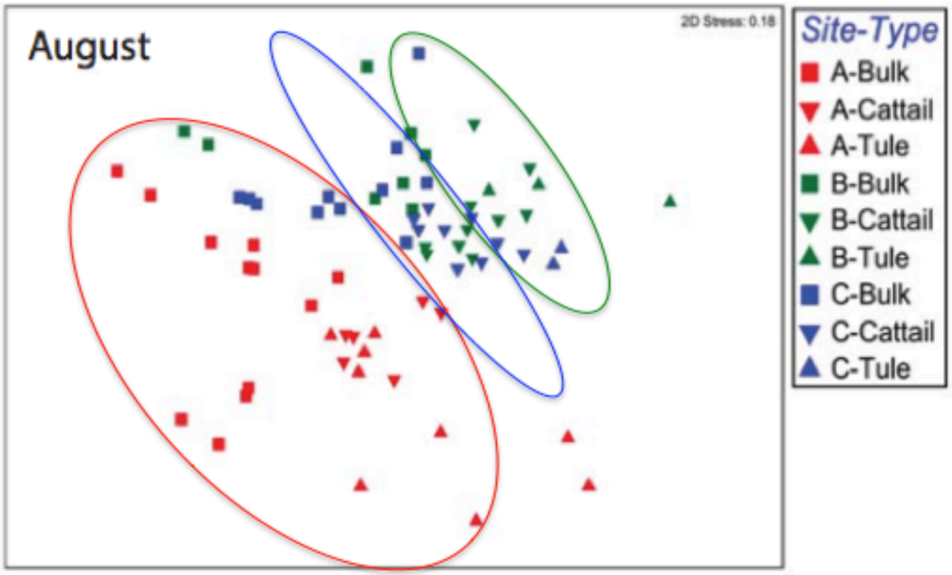
16S community composition



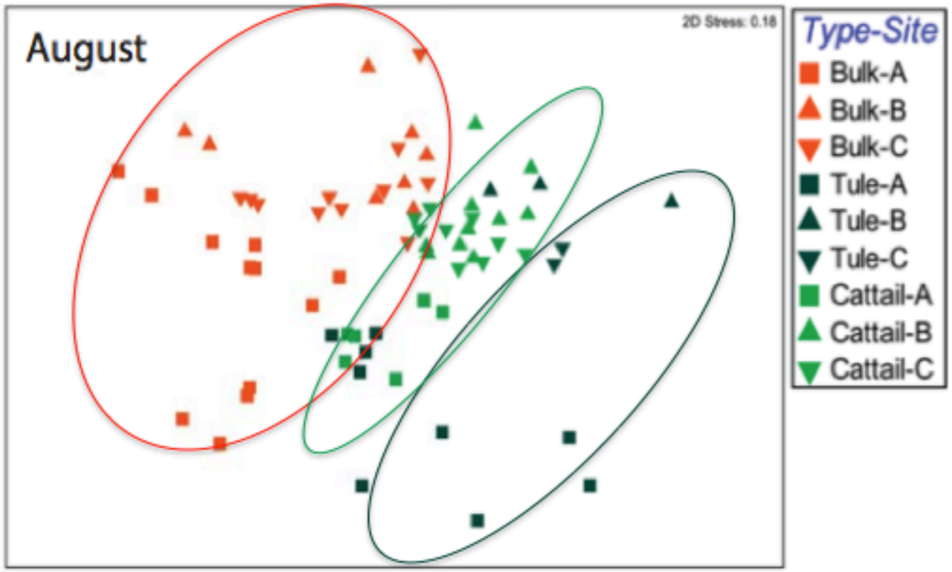
ANOSIM *R* by site = 0.411



ANOSIM *R* by type = 0.370



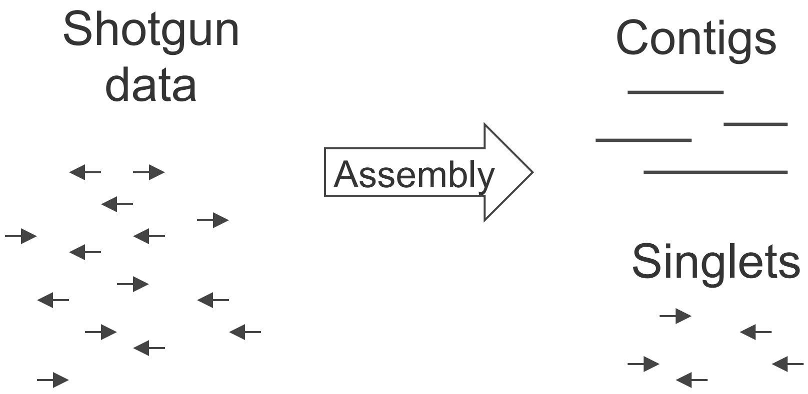
ANOSIM *R* by site = 0.294



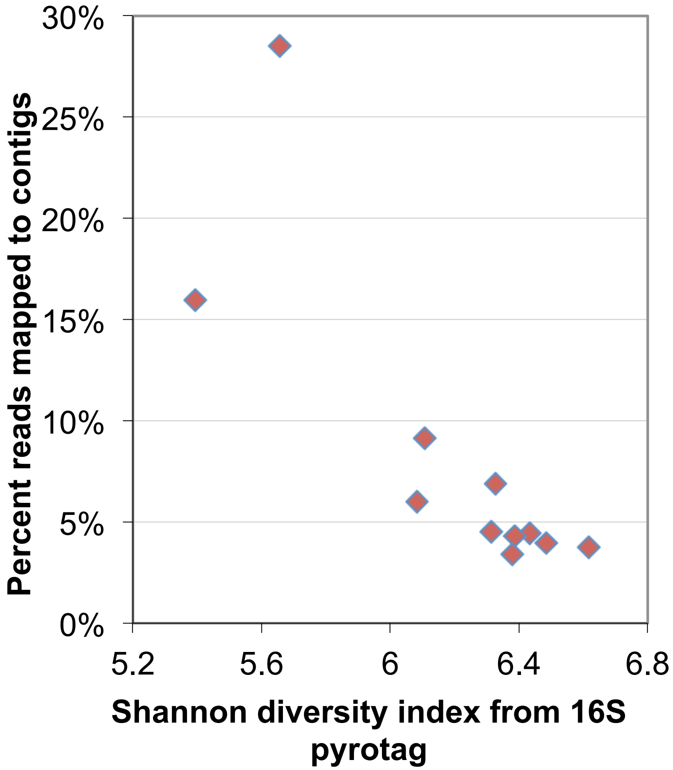
ANOSIM *R* by type = 0.395

He et al. in review

Wetland Metagenomes



Sample	Total bases generated (Gbp)	Contig N50 (bp)	Contig max length (bp)	Total base in contig (Mbp)	Percent reads mapped to contig (%)
A1_Bulk	59.9	456	55,504	208	9.12
A2_Cattail	58.8	389	85,839	84	3.73
A1_Tule	57.0	432	33,581	129	5.99
B1_Bulk	53.7	437	128,213	153	6.89
B2_Bulk	54.6	375	12,499	61	3.41
B1_Cattail	60.2	541	144,255	291	15.95
B2_Cattail	61.5	364	42,624	91	3.98
B2_Tule	53.9	373	37,498	106	4.42
L1_Bulk	60.3	439	89,191	70	4.54
L1_Cattail	59.5	779	246,650	297	28.51
L2_Tule	52.2	389	221,244	112	4.29



- ➢ More complex community, less assembly

Shaomei He

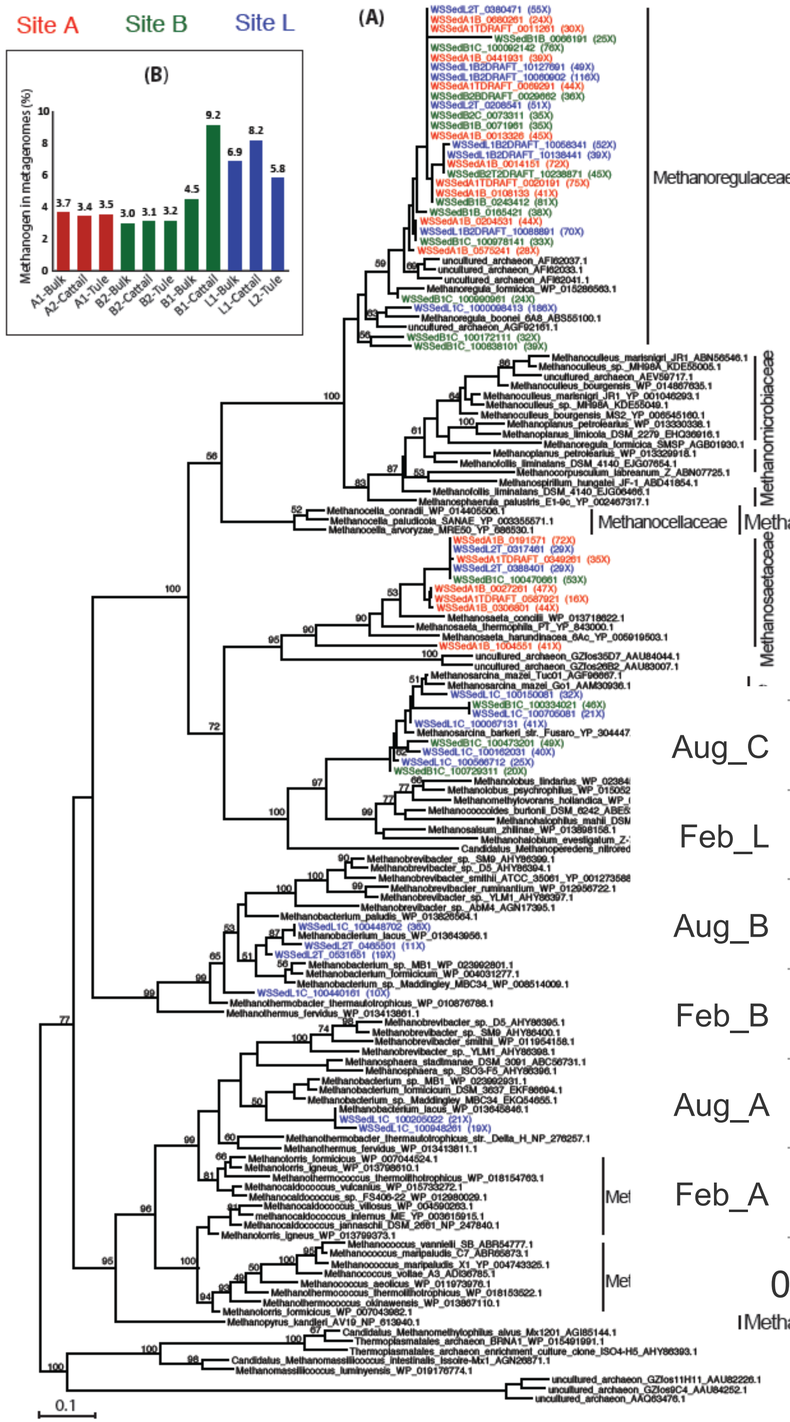
Methanogen diversity

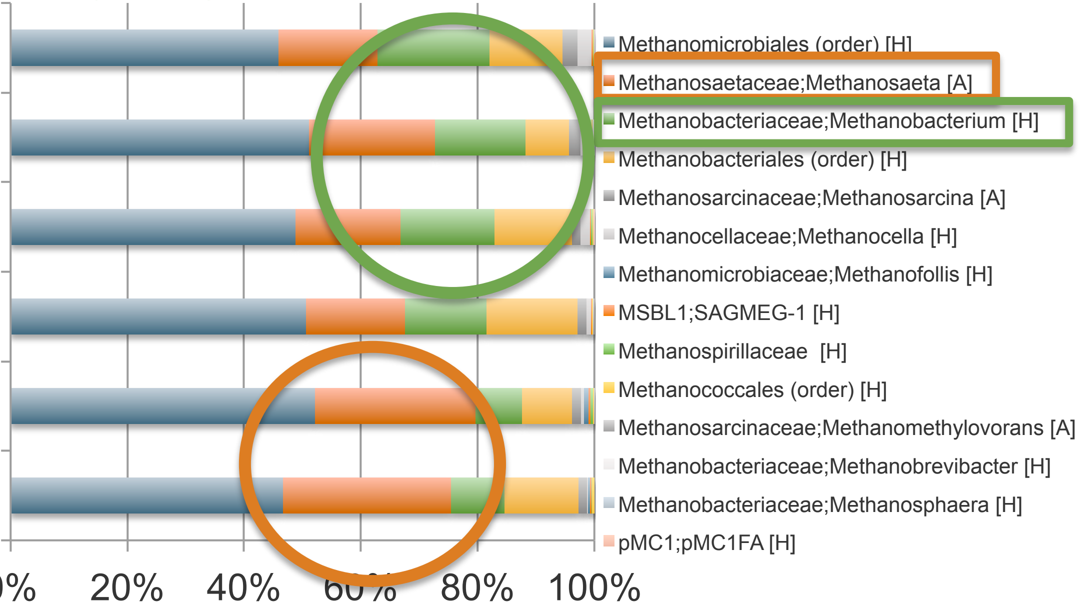
Hydrogenotrophic Methanogenesis:

$$CO_2 + 4\,H_2 \rightarrow CH_4 + 2\,H_2O$$

Acetoclastic Methanogenesis:

$$CH_3COOH \rightarrow CH_4 + CO_2$$





Hydrogenotrophic

Acetoclastic