Microbial Ecology and Metagenomics for the Environment

James M. Tiedje, Center for Microbial Ecology
Departments of Microbiology and Molecular Genetics and of Crop and Soil Sciences

Michigan State University and Yonsei University

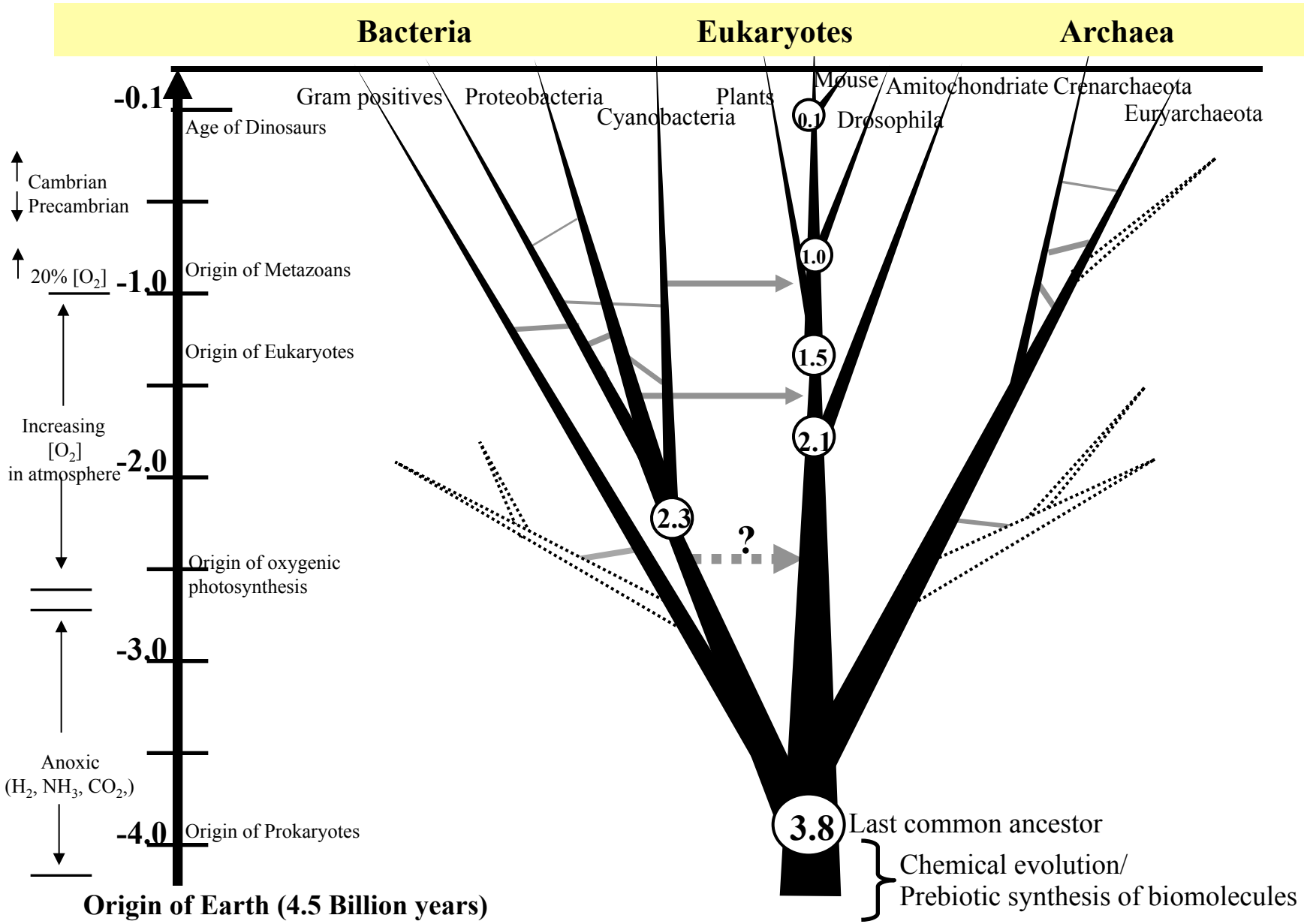
2012 Metagenomics Workshop
March 9, 2012
Yonsei University, Seoul

MICHIGAN STATE UNIVERSITY

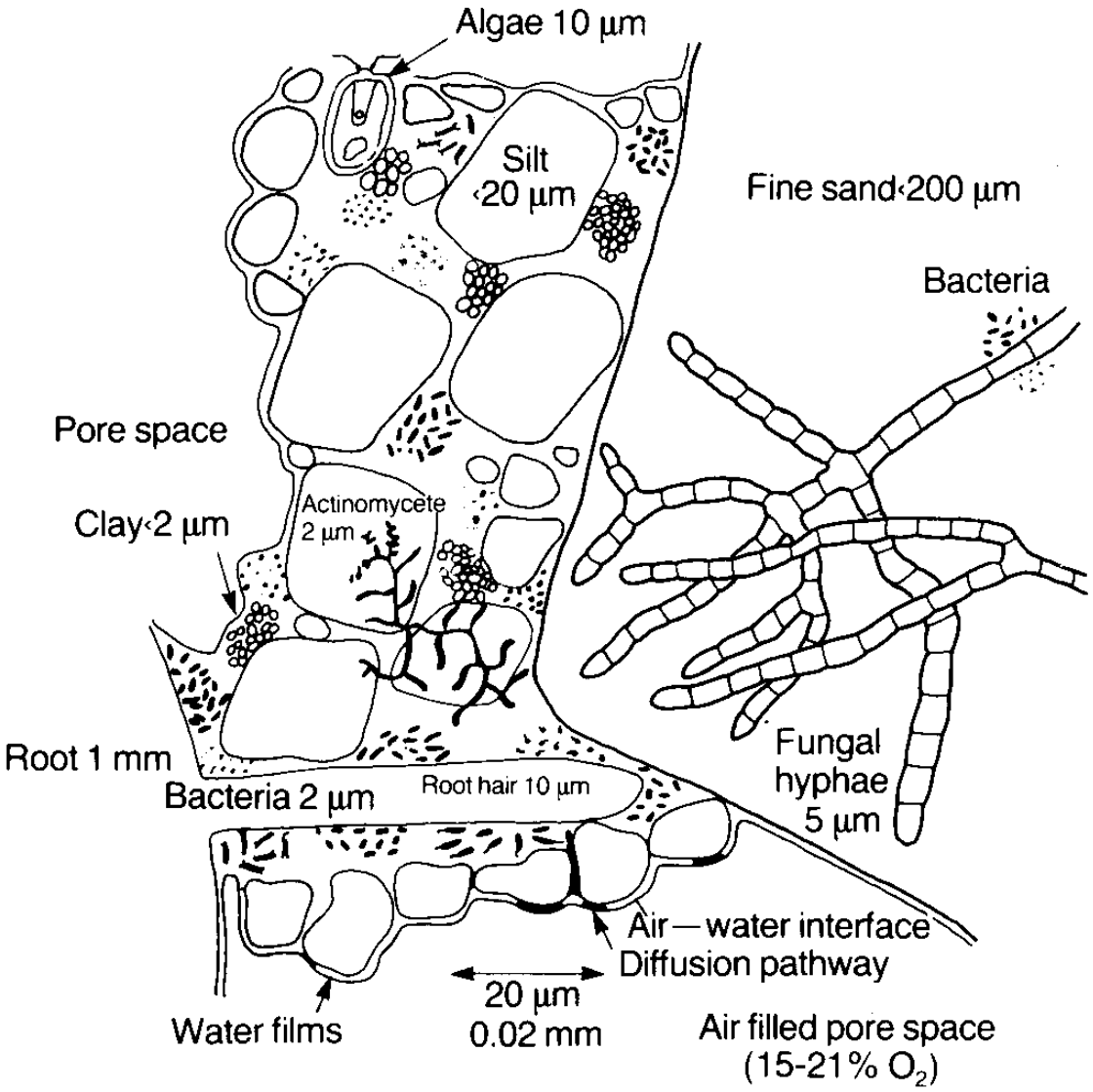
Why is microbial diversity so high in soil?

- Current microbes are the product of 3.5 billion years of evolution, many remaining and adapting to the soil habitat.
- The soil environment has an almost infinite range of conditions, complex gradients, multiple resources and is very protective.
- The pangenome is the rule for microbial “species” and providing in huge gene diversity within each “species”.

The Tree of Life: the Microbial World is OLD



The soil habitat is complex, many niches



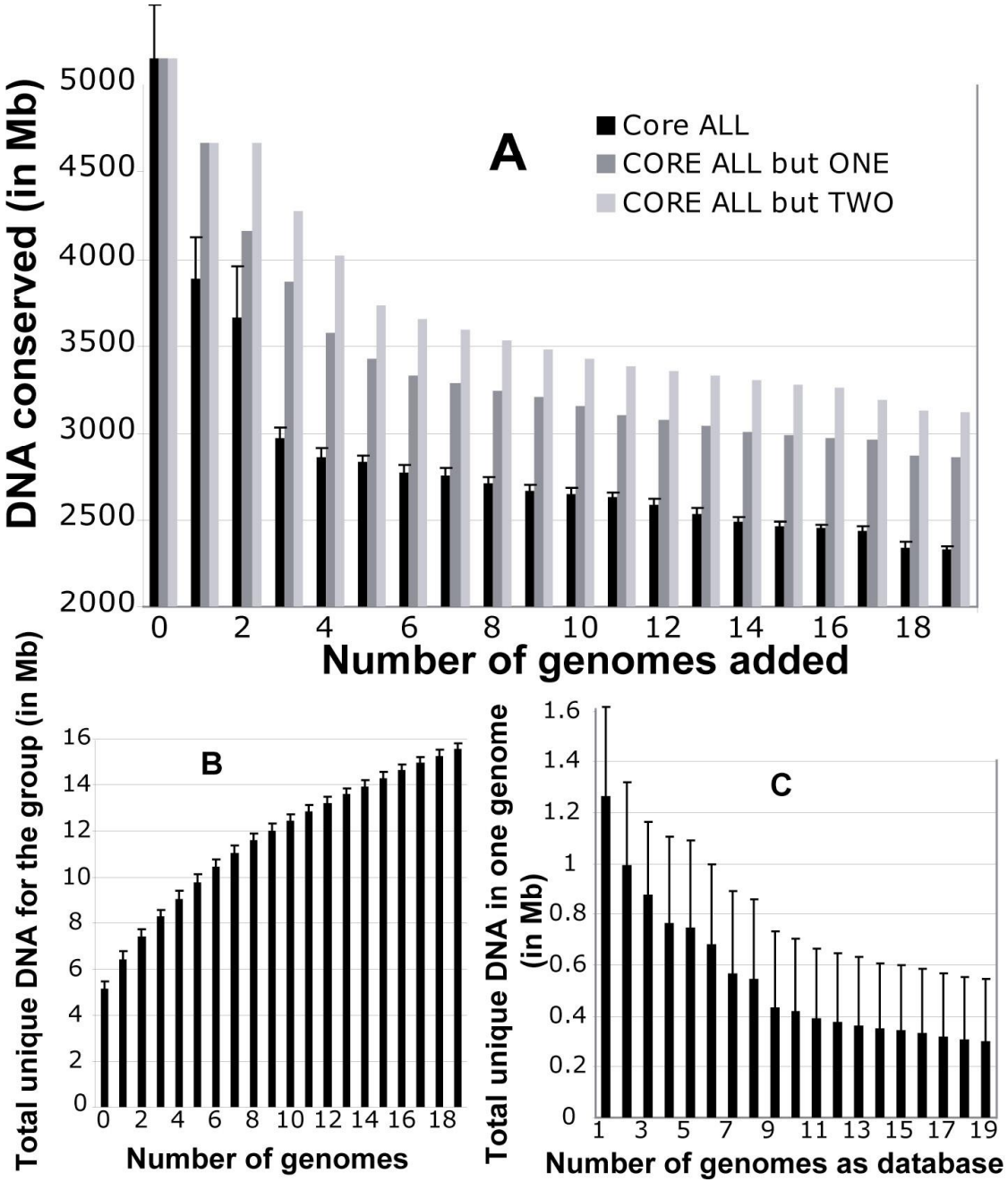
Soil is a composite of communities, a complex of subhabitats.

- microaggregates
- rhizosphere
- mycosphere
- fauna
- pore surfaces
- OM coatings

Its relatively

- stable,
- extensive,
- ancient,
- protective

One example: *E. coli*

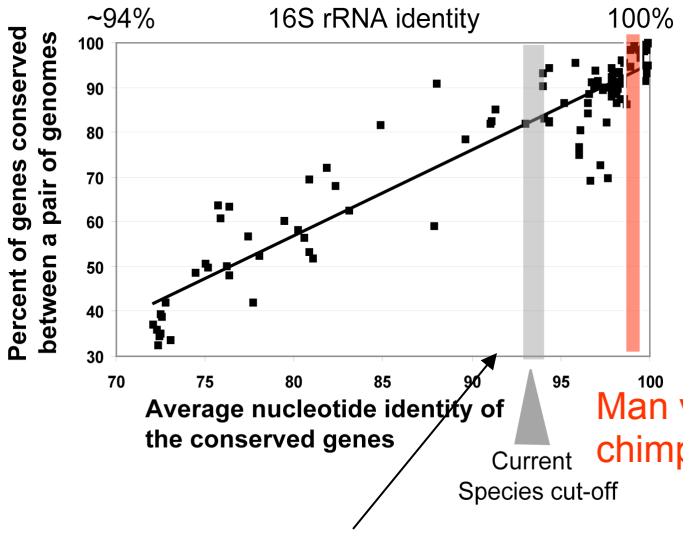


Size, 5.1Mb
Core, 3,000
Each new genome,
300 genes
16,000 genes explored
(the pan genome)

Konstantinidis et. al
Trans Royal Soc.,
London: B, 2006

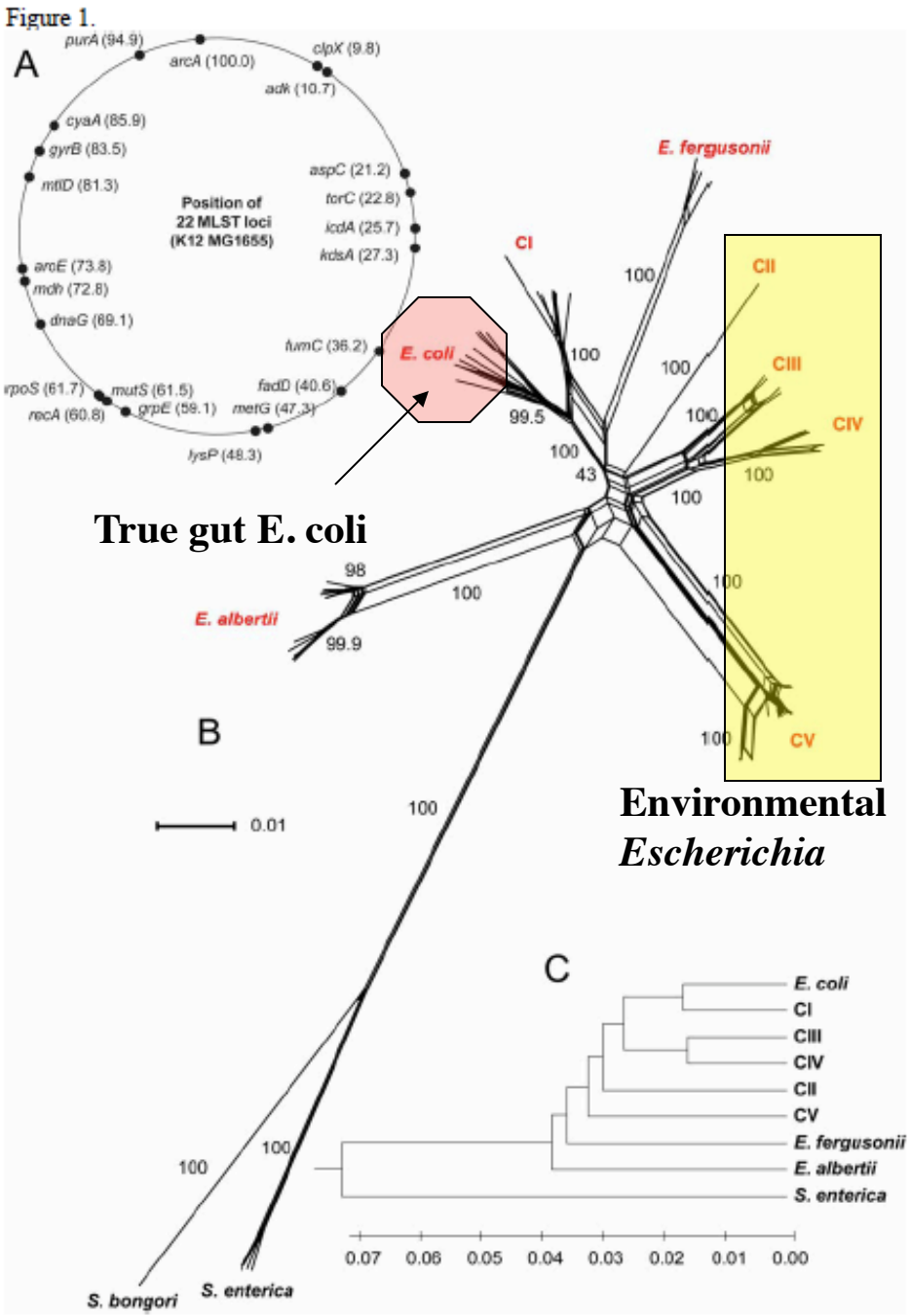
Escherichia is a broad genus

The phenotypic traits for *E. coli* do not distinguish the environmental *Escherichia*

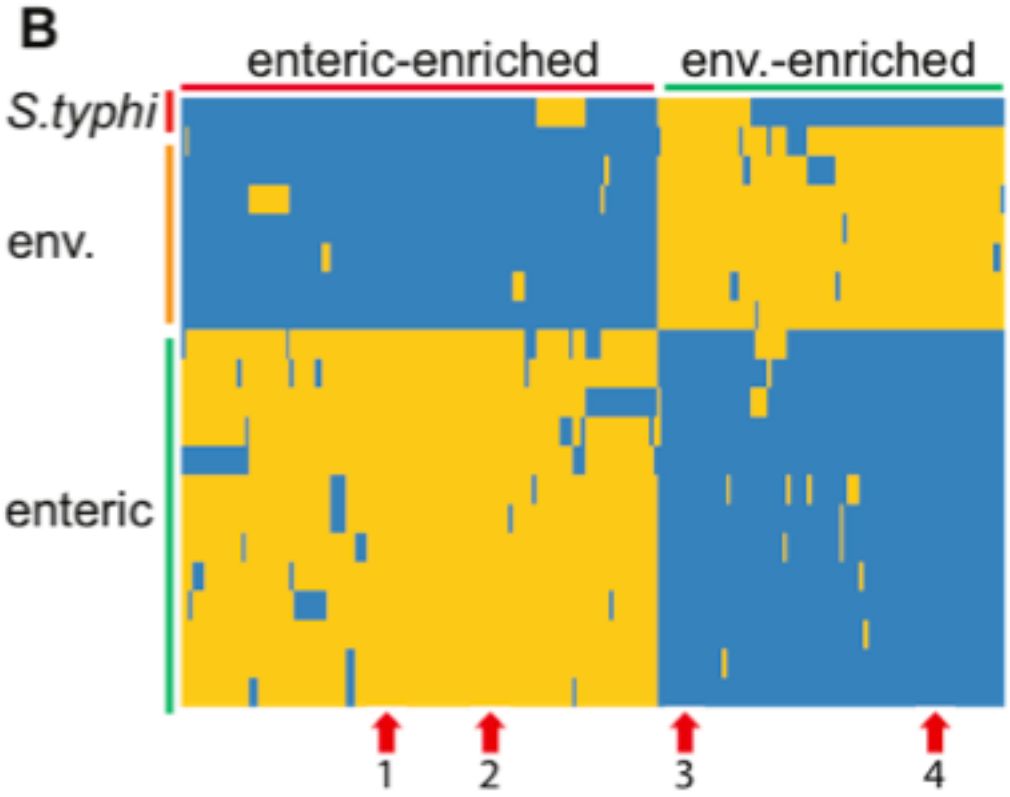


Environmental *Escherichia*, 91-93% ANI

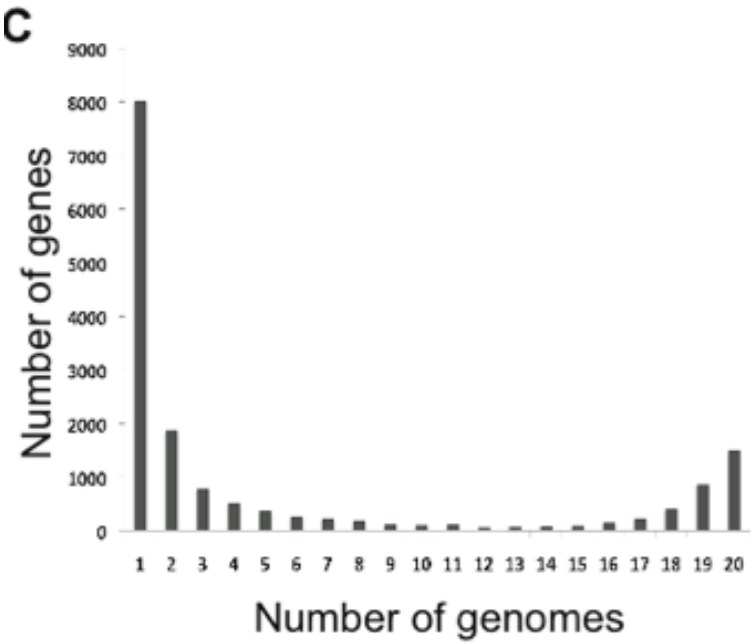
Cryptic lineages of *Escherichia*
Walk, et. al., AEM, 2009



Genes more common in environmental vs enteric strains



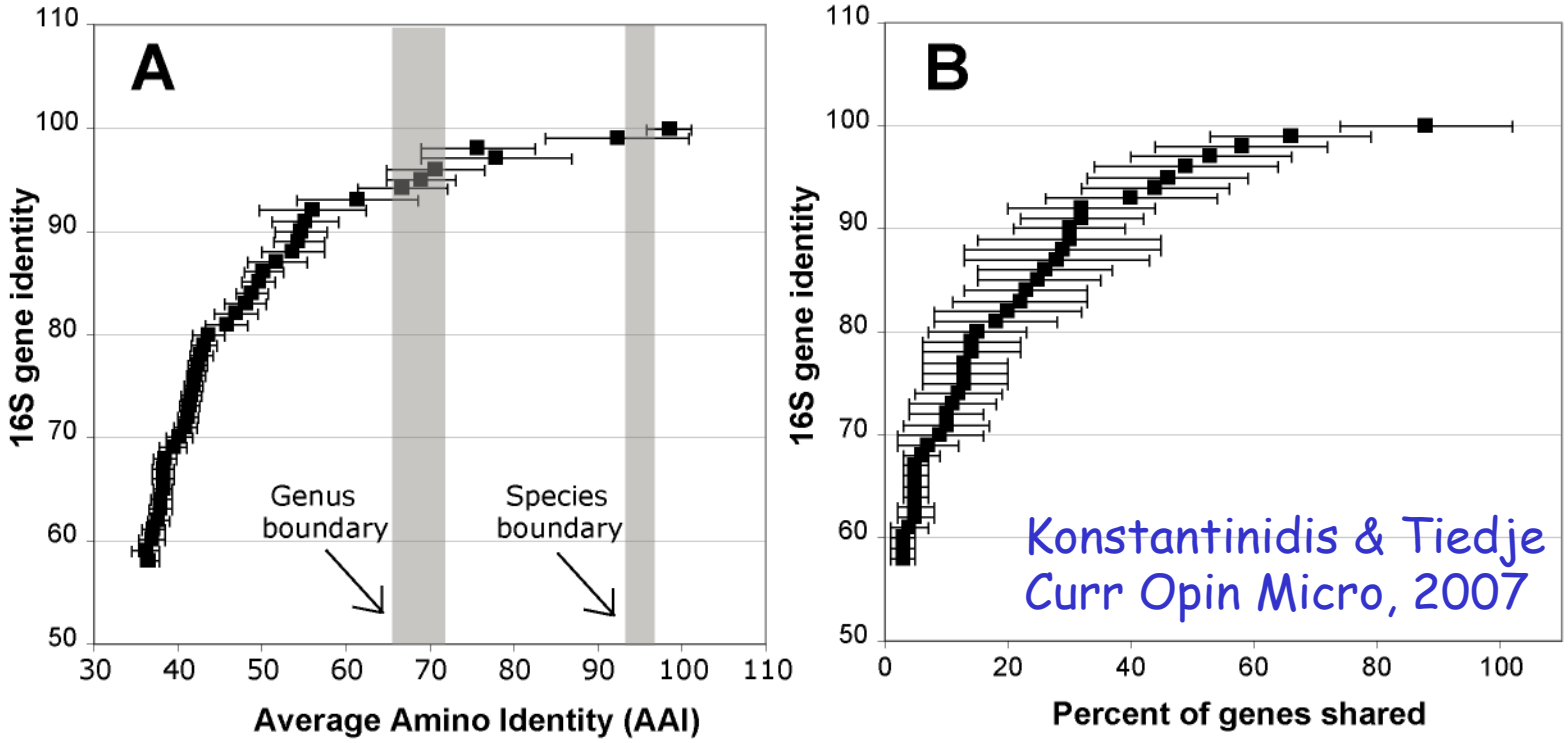
Genome-specific ⇨ common (core)



Luo, et. al, Tiedje and Konstantinidis, PNAS 108:7200-7205 (2011)

What is a species & what is a genus?

In terms of common gene content



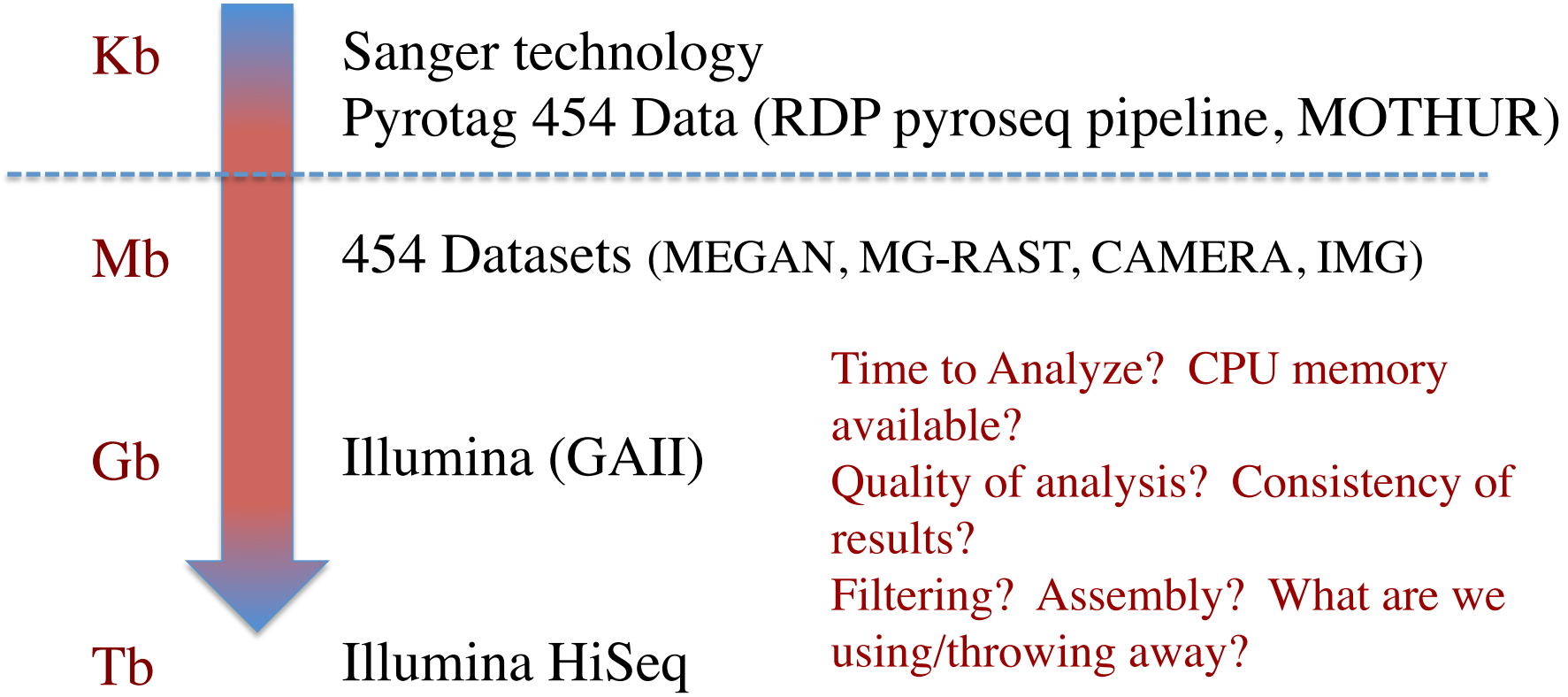
Species > 70%; Genus 20 - 40% common genes

Humans and sea urchins have 70% of their genes in common

Why is microbial diversity so high in soil?

- Current microbes are the product of 3.5 billion years of evolution, many remaining and adapting to the soil habitat.
- The soil environment has an almost infinite range of conditions, complex gradients, multiple resources and is very protective.
- The pangenome is the rule for microbial “species” and providing in huge gene diversity within each “species”.

Growth of Sequencing Output



Tackling Metagenomes' Largest Challenge: *The GREAT PRAIRIE*

Now at 1.9Tbp

Coordinators:

James Tiedje (MSU), Janet Jansson (LBL)

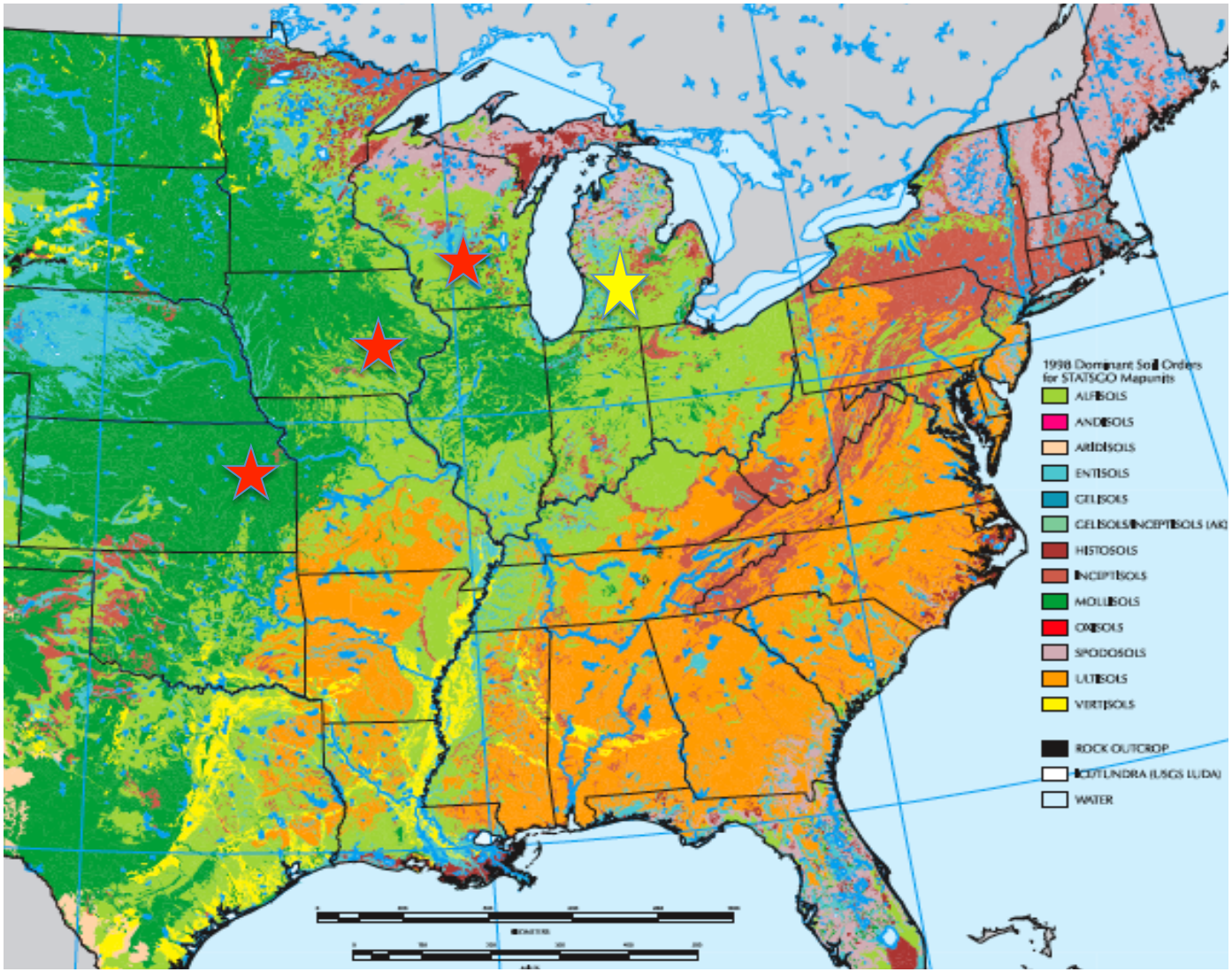
Plus many collaborators

at JGI, LBNL, MSU and the sites

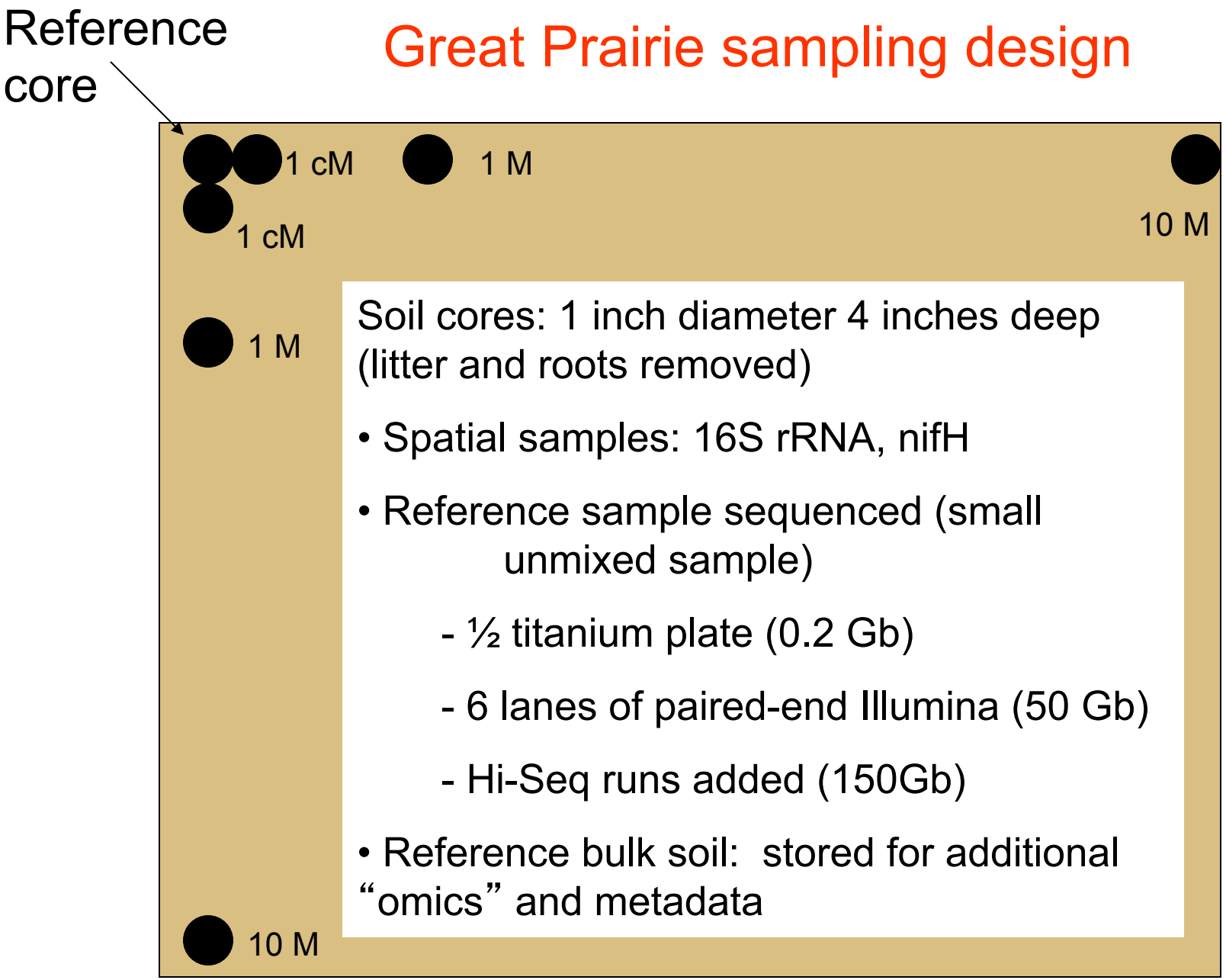
Paired native prairie and long-term agriculture sites:
- Wisconsin (Goose Pond)
- Iowa (Morris Prairie)
- Kansas (Konza Prairie area)

The US Great (Midwest) Prairie

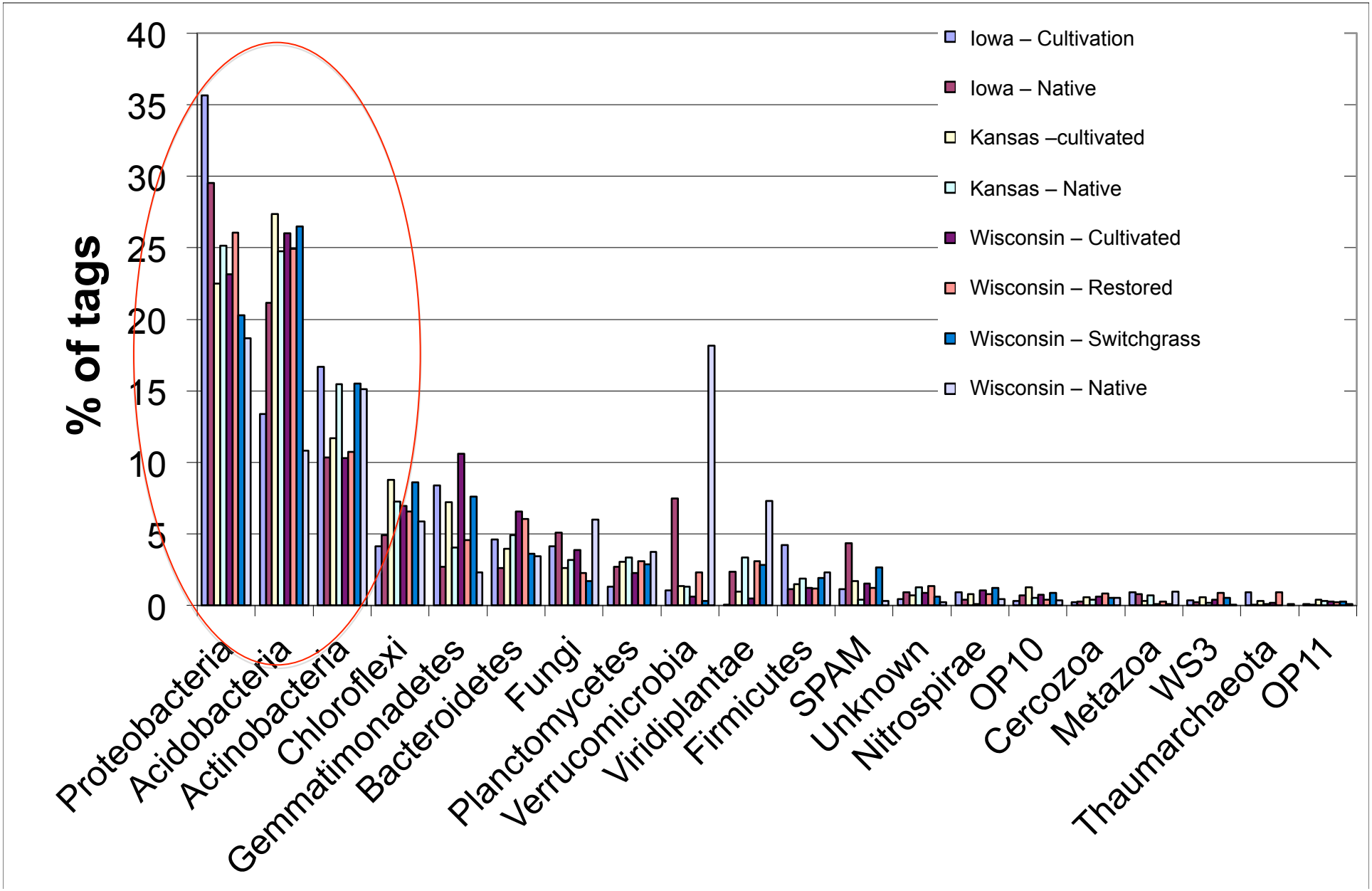
Holds about 1/3 of US soil carbon stocks & vital to food security



Great Prairie sampling design

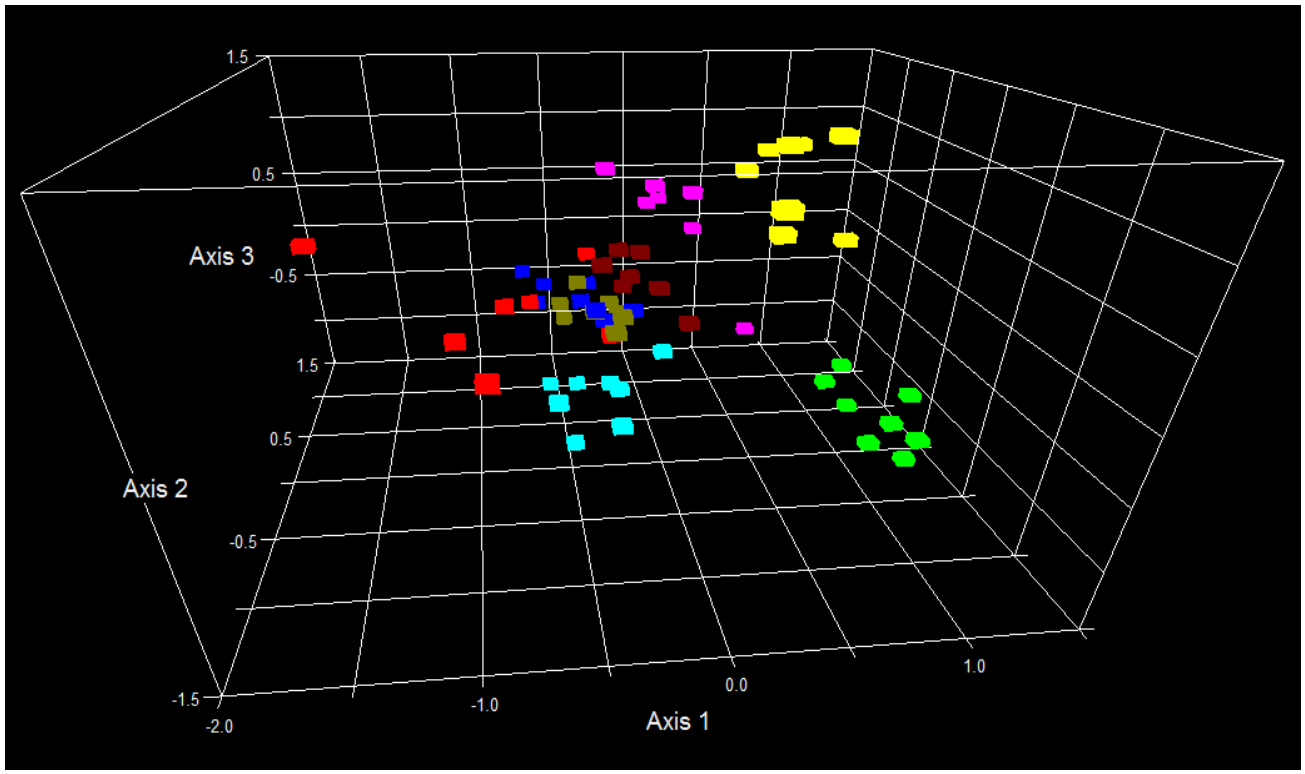


Who lives in the Great Prairie's soil?



Do the microbial communities differ among sites?

16S rRNA gene (Pyrotag analysis)

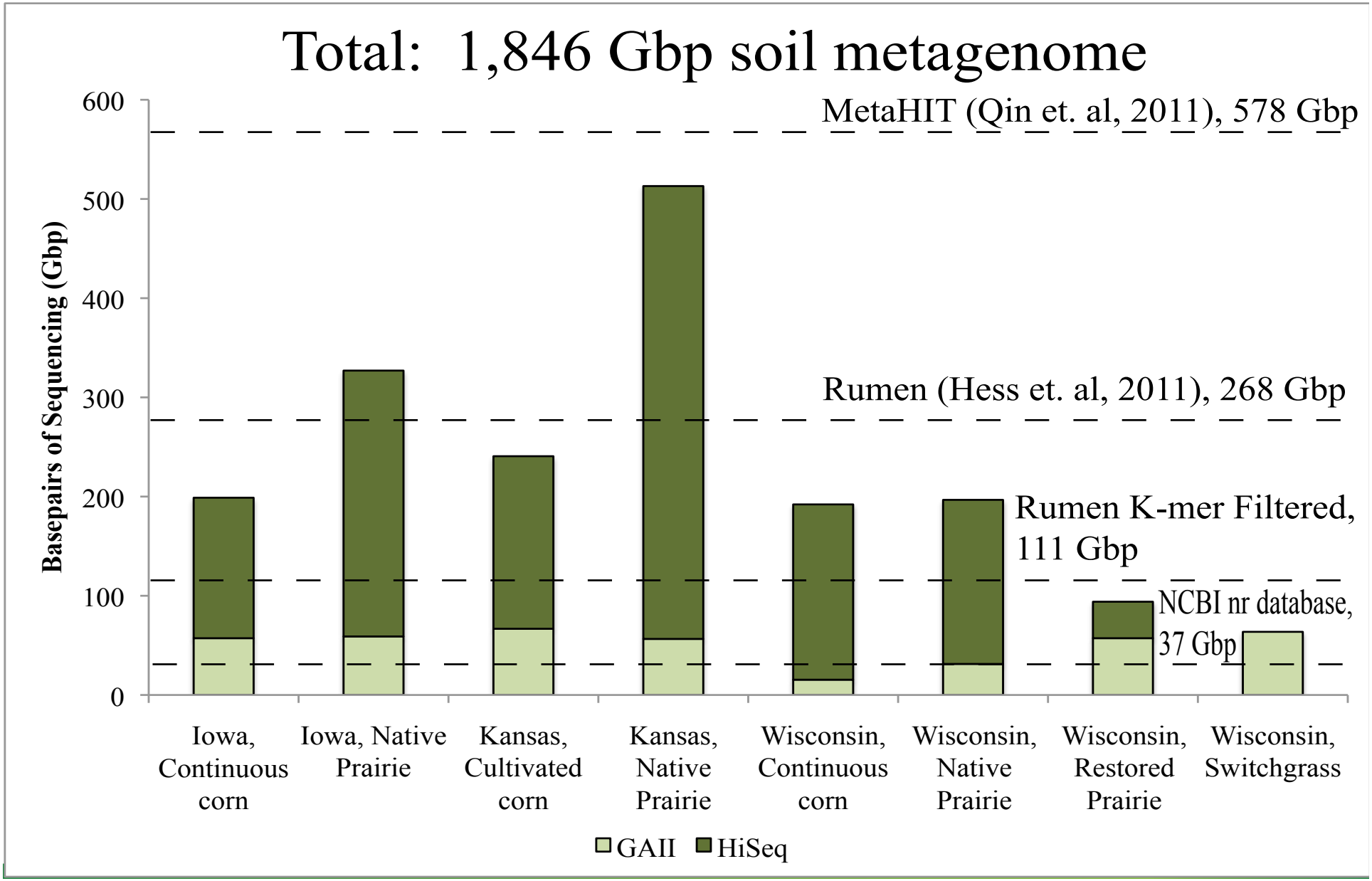


- Iowa Corn
- Iowa Prairie
- Kansas Corn
- Kansas Prairie
- Wisconsin Corn
- Wisconsin Prairie
- Wisconsin Restored Prairie
- Wisconsin Switchgrass

Non-metric Multi Dimensional Scaling (nMDS) using PCORD v5 software
Bray Curtis distance measure
Final stress for 3-D solution =13.04

Regina Lamendella,
LBNL

Great Prairie Sequencing Summary



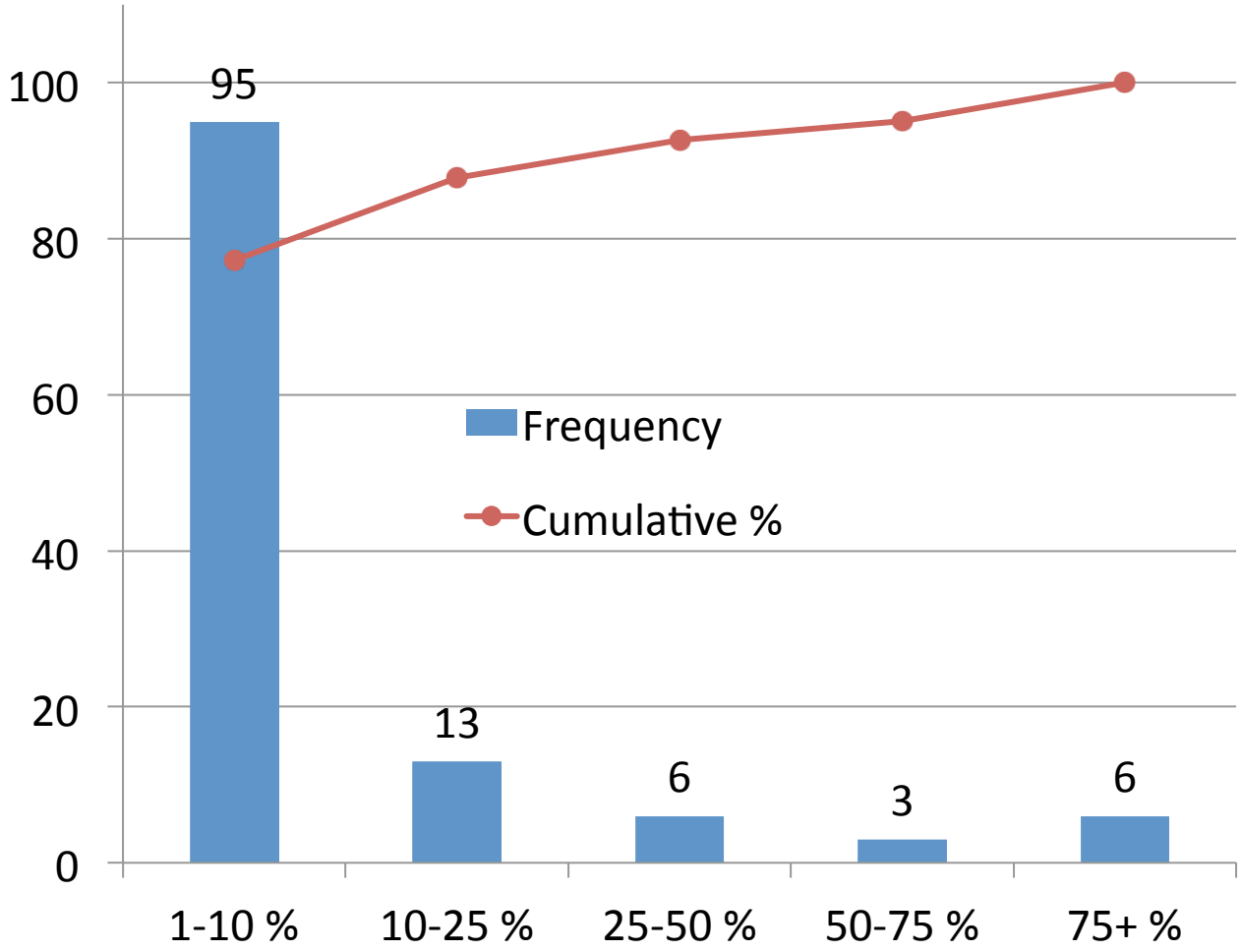
What can we do with Illumina short reads?

- Search for rRNA and ecogenes (will replace tag pyrosequencing)
- Map reads to sequenced genomes
- Target ecofunctional genes
  - By amplicon targeting (Gene-targeted metagenomics)
  - By computational walking to assemble genes of interest
- Assemble shotgun reads into contigs and more
  - New approaches being developed for the massive data

Each can be used to link to taxonomy

- *but is dependent on database, which is deficient for soil*
- *assumes horizontal gene transfer is minor, but not always the case*

Distribution of detected organisms by their genome recovery



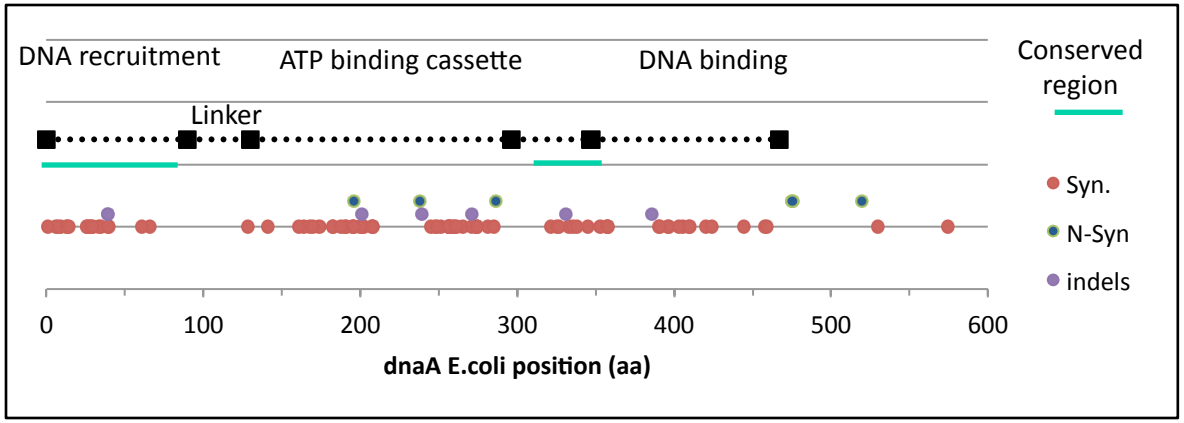
Best Genome Matches for One Soil

- *Metagenome sequences map to which sequenced strains?*
- *How much of the genome is present (%), what 'x' coverage (%)*
- *How many and where are SNPs and indels*

		MSR1		
Reference	Genome	SNP count	Recovery	Coverage (X)
AE007869	Agrobacterium_tumefaciens_C58 circular chromosome	203	93.79%	4.4
AE007870	Agrobacterium_tumefaciens_C58 linear chromosome	159	92.70%	5.9
CP000712	Pseudomonas_putida_F1	141	88.44%	5.0
CP001635	Variovorax_paradoxus_S110	546	75.41%	7.1
CP001636	Variovorax_paradoxus_S110 plasmid	51	47.17%	5.1
CP002248	Agrobacterium_sp_H13-3	68	84.74%	2.7
CP002249	Agrobacterium_sp_H13-3	77	92.00%	3.9
CP002505	Rahnella_sp_Y9602	120	97.33%	384.0
CP002506	Rahnella_sp_Y9602 plasmid	17	91.81%	327.0
CP002585	Pseudomonas_brassicacearum_NFM421	338	94.91%	24.3

What genes were most commonly affected by sequence changes?

Annotation (unique organisms affected)	Syn. SNP	Non-Syn. SNP	Indels
Hypothetical protein (61)	29	16	258
Conserved hypothetical (46)	17	11	226
DNA replication initiator *dnaA (40)*	86	7	6
gyrB (13)	20	4	0
gyrA (9)	12	1	4



What can we do with Illumina short reads?

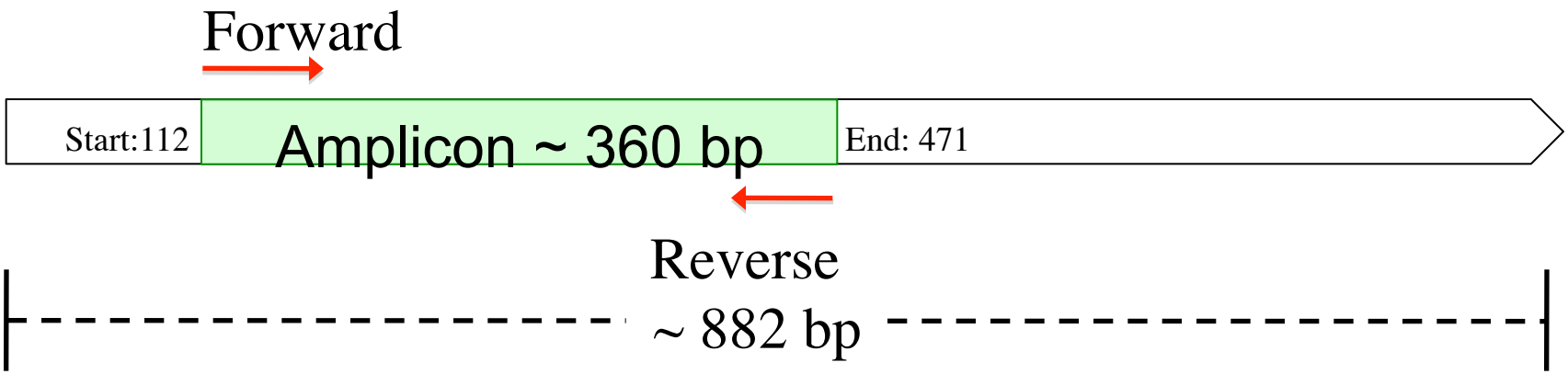
- Search for rRNA and ecogenes (will replace tag pyrosequencing)
- Map reads to sequenced genomes
- Target ecofunctional genes
  - By amplicon targeting (Gene-targeted metagenomics)
  - By computational walking to assemble genes of interest
- Assemble shotgun reads into contigs and more
  - New approaches being developed for the massive data

Each can be used to link to taxonomy

- *but is dependant on database, which is deficient for soil*
- *assumes horizontal gene transfer is minor, but not always the case*

The nifH gene is a key gene in nitrogen fixation

Example: *Trichodesmium thiebautii*

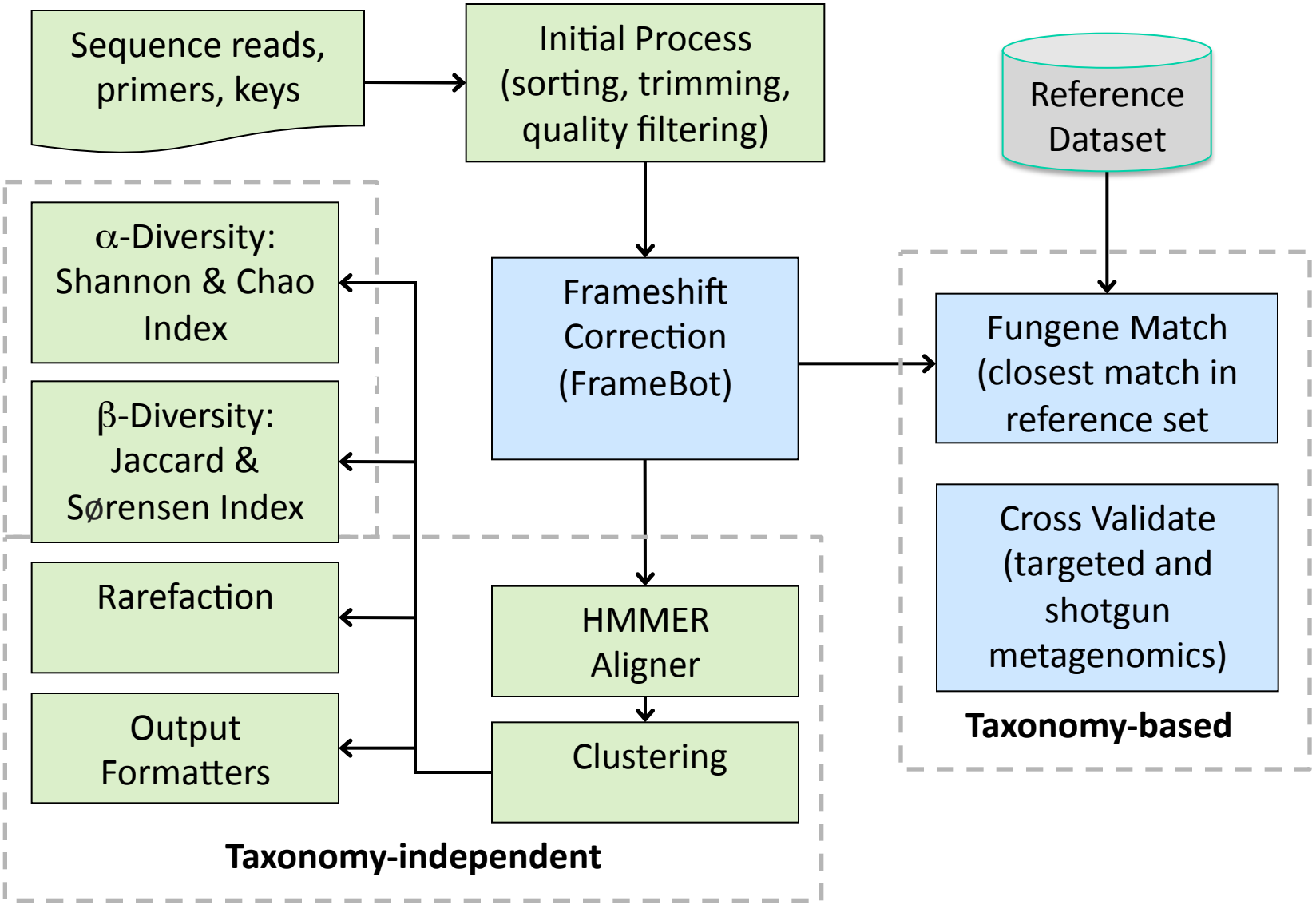


Zehr et al., 1989. Forward TGYGAYCCNAARGCNGA
Reverse ADNGCCATCATYTCNCC

Poly et al., 2001. Forward TGCGAYCCSAARGCBGACTC
Reverse ATSGCCATCATYTCRCCGGA

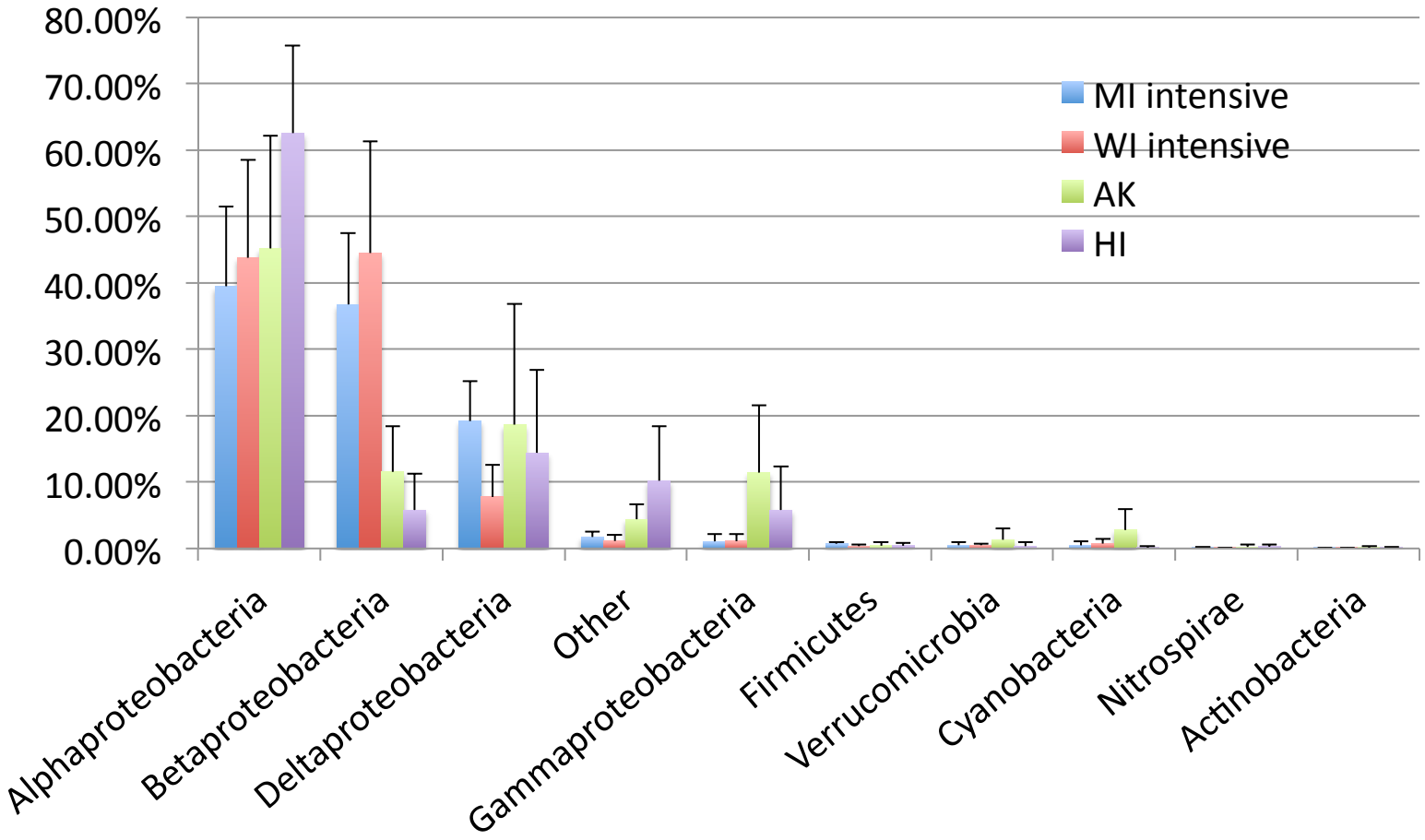
http://rdp.cme.msu.edu

Gene-Targeted Metagenomics Pipeline



http://rdp.cme.msu.edu

Phylum Level NifH FunGene Match



8 hrs to process 222 samples (1.1 million reads)
92.7% within 90% aa identity of best reference sequence

How to find genes from short (Illumina) reads?

Two versions of the same theme: assemble:

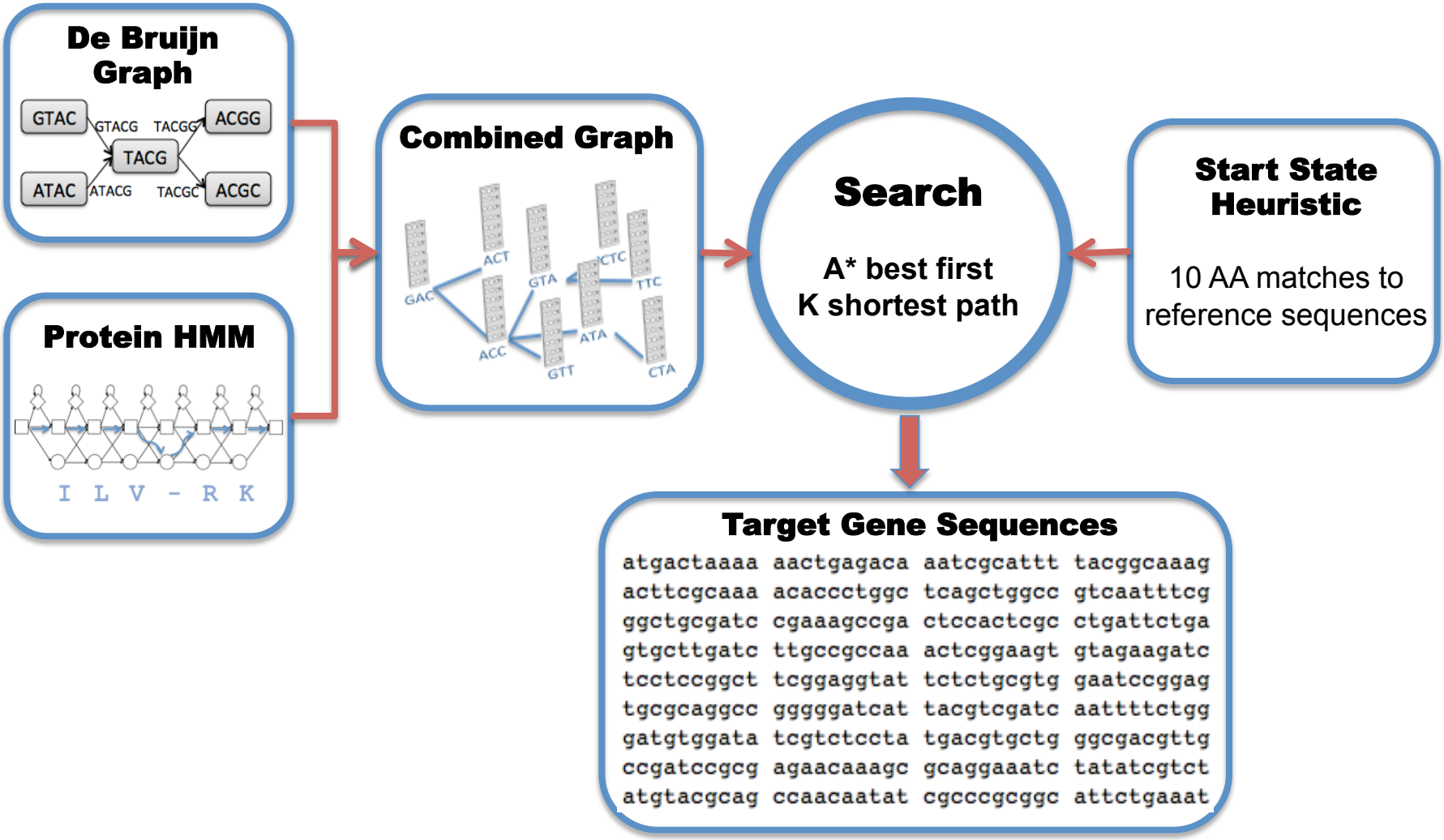
1. **Target genes of interest (ecofunctional genes) & assemble, e.g. Xander**
   - **Focus resources early on genes of interest**
   - **Use more of the data**
2. **Assemble and then search for genes of interest**

Xander developers

Jordan Fish

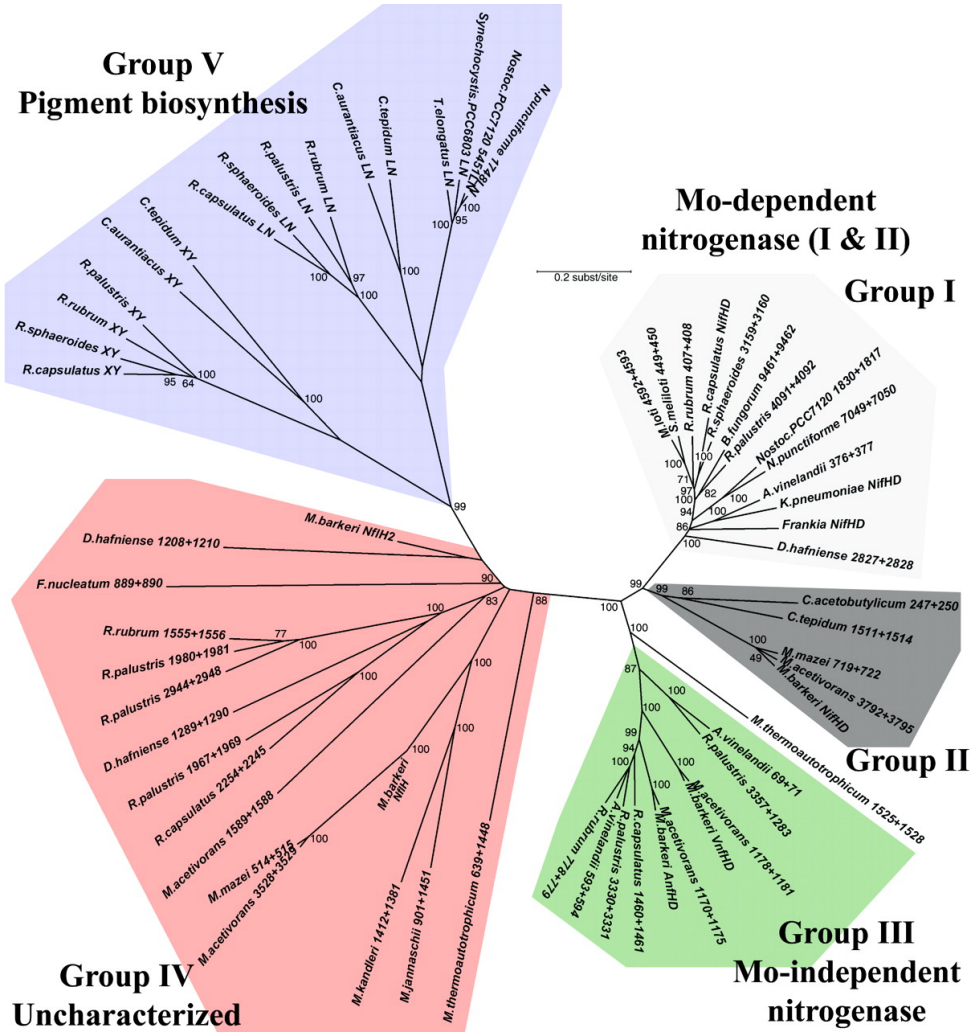
Jim Cole

Xander: Gene-Targeted Metagenome Assembler



Targeting the N-Cycle; a Key Process in Most Ecosystems

Example; nifH of Nitrogenase



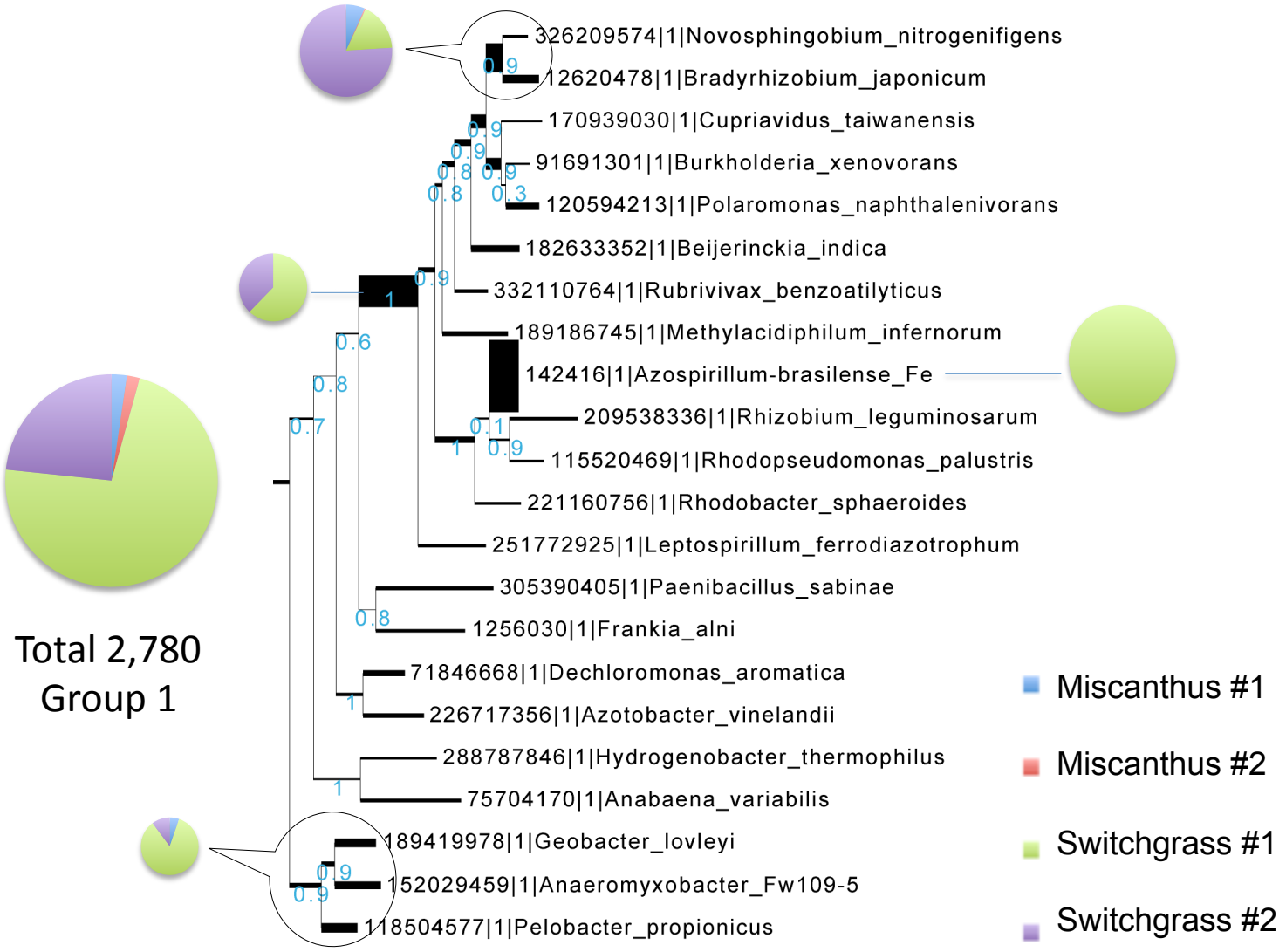
A key standards issue:
Claims based on unproven reference data

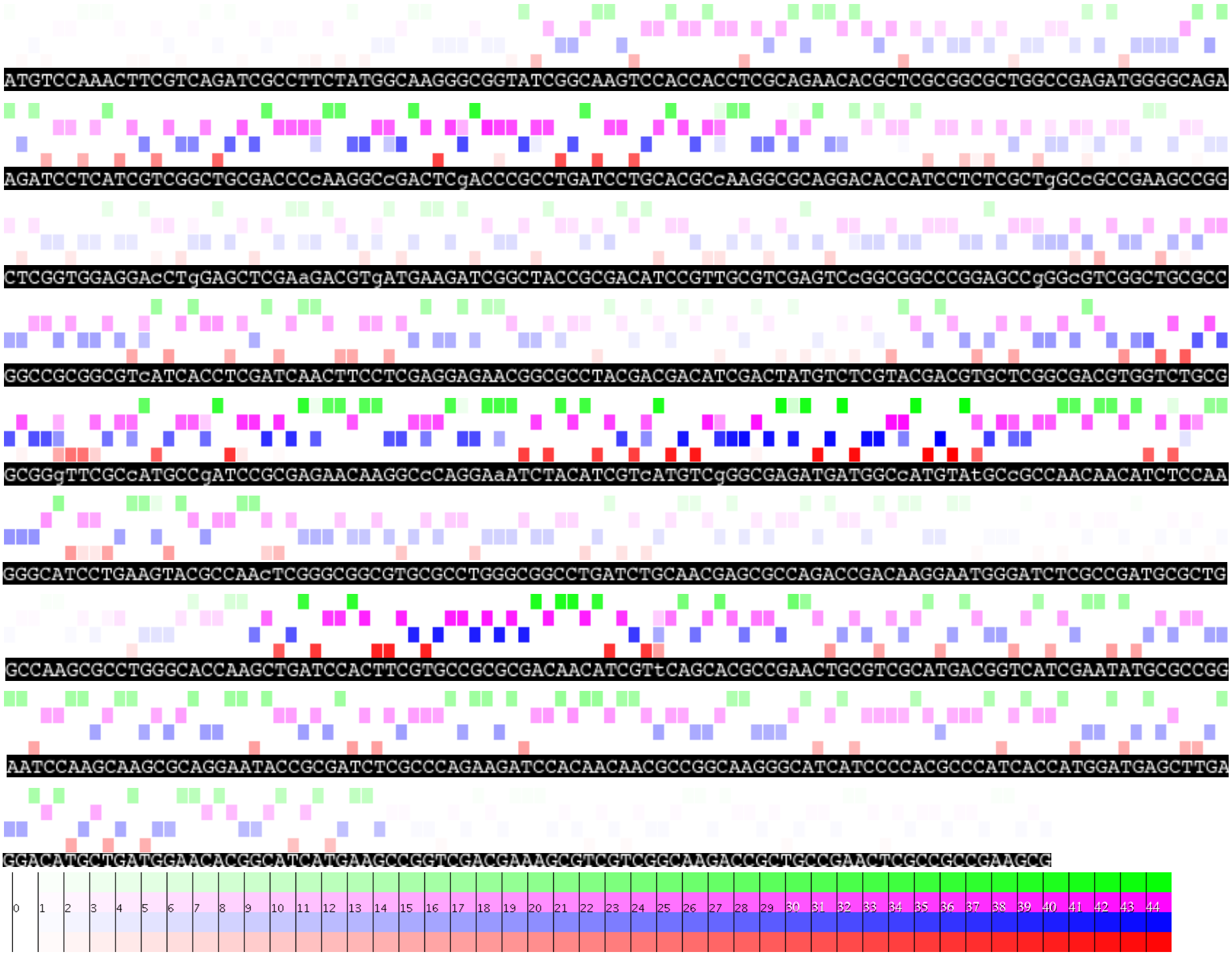
Five phylogenetic groups from concatenated phylogenetic tree composed of NifH and NifD homologs found in complete genomes.

Raymond J et al. Mol Biol Evol 2004;21:541-554

MBE MOLECULAR BIOLOGY AND EVOLUTION

Group-1 nifH genes found using Xander (single-path search) in four soil metagenomes




ATGTCCAAACTTCGTCAGATCGCCTTCTATGGCAAGGGCGGTATCGGCAAGTCCACCACCTCGCAGAACACGCTCGCGGCGCTGGCCGAGATGGGGCAGA
AGATCCTCATCGTCGGCTGCGACCCcAAGGCcGACTCgACCCGCCTGATCCTGCACGCcAAGGCGCAGGACACCATCCTCTCGCTgGCcGCCGAAGCCGG
CTCGGTGGAGGAcCTgGAGCTCGAaGACGTgATGAAGATCGGCTACCGCGACATCCGTTGCGTCGAGTCcGGCGGCCCGGAGCCgGGcGTCGGCTGCGCC
GGCCGCGGCGTcATCACCTCGATCAACTTCCTCGAGGAGAACGGCGCCTACGACGACATCGACTATGTCTCGTACGACGTGCTCGGCGACGTGGTCTGCG
GCGGgTTCGCcATGCCgATCCGCGAGAACAAGGCcCAGGAaATCTACATCGTcATGTCgGGCGAGATGATGGCcATGTAtGCcGCCAACAACATCTCCAA
GGGCATCCTGAAGTACGCCAAcTCGGGCGGCGTGCGCCTGGGCGGCCTGATCTGCAACGAGCGCCAGACCGACAAGGAATGGGATCTCGCCGATGCGCTG
GCCAAGCGCCTGGGCACCAAGCTGATCCACTTCGTGCCGCGCGACAACATCGTtCAGCACGCCGAACTGCGTCGCATGACGGTCATCGAATATGCGCCGG
AATCCAAGCAAGCGCAGGAATACCGCGATCTCGCCCAGAAGATCCACAACAACGCCGGCAAGGGCATCATCCCCACGCCCATCACCATGGATGAGCTTGA
GGACATGCTGATGGAACACGGCATCATGAAGCCGGTCGACGAAAGCGTCGTCGGCAAGACCGCTGCCGAACTCGCCGCCGAAGCG
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44

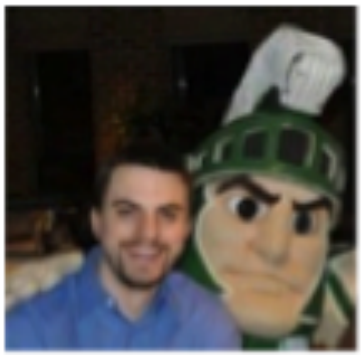
Soil Assembly Obstacles ⟶ New solutions

- Scalability
- Diversity of soil
- Non uniform coverage
- Availability of computational resources
- Lack of reference genomes

Dragon slayers

New assembly approach

- *Memory efficient*
- *Digital normalization*
- *Partitioning*
- *Artifact removal*
- *Assembly*

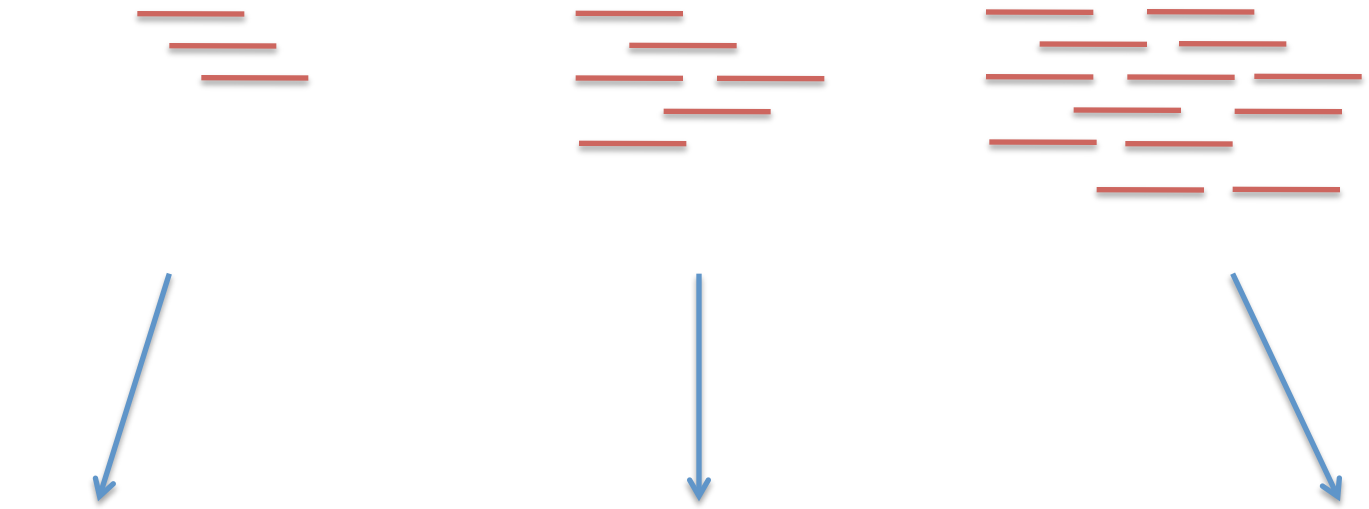


C. Titus Brown,
Asst Prof of Computer Sci,
MSU

Adina Howe,
Brown/Tiedje Labs
MSU

The Basic Idea of the Assembly Solution

Separate disconnected (non-overlapping) sets of reads into different bins

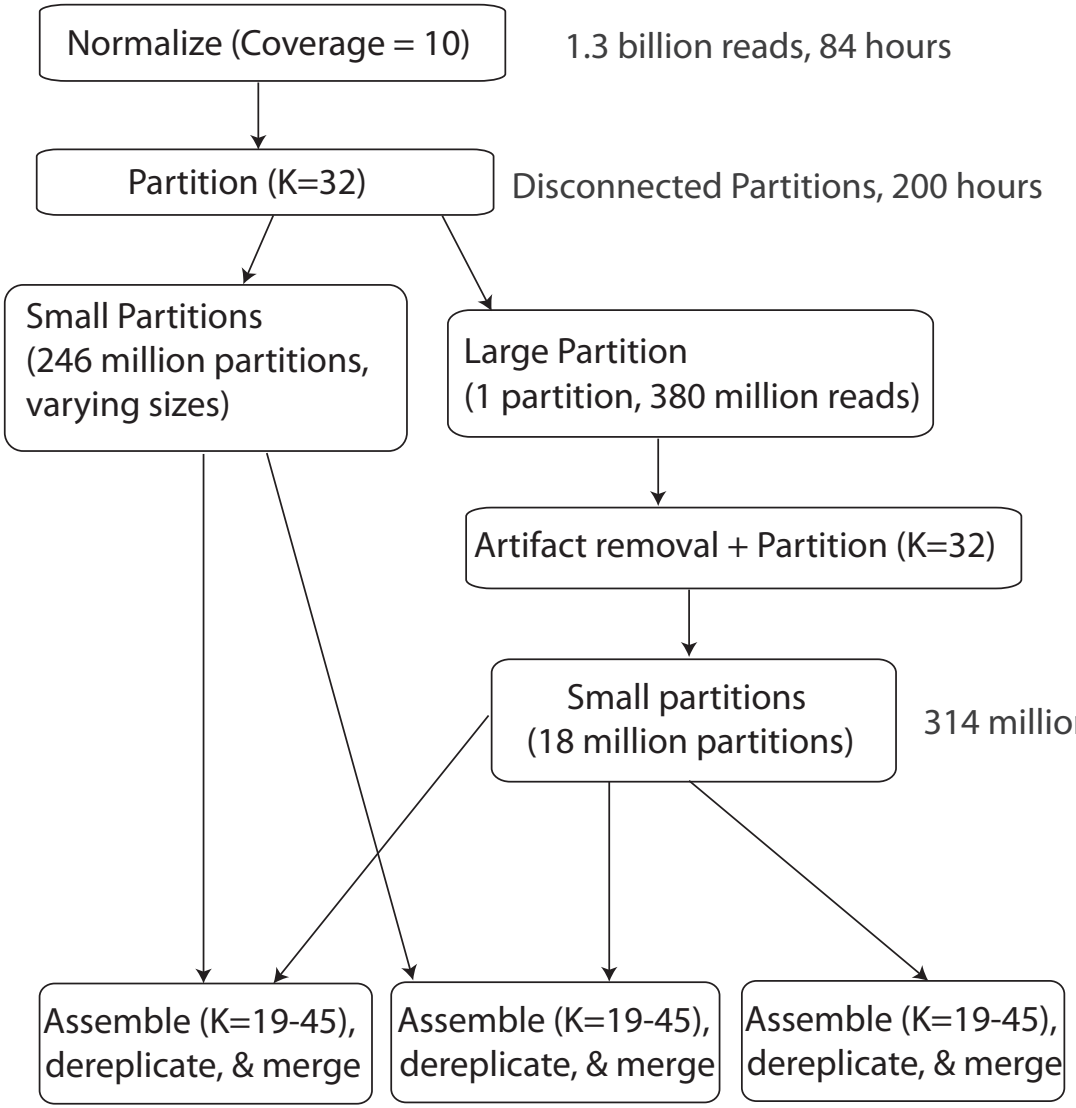


Bin 1 Bin 10 Bin 50

Assembling these bins independently => identical to global assembly

No kmer filtering to discard less common sequences

Summary of New Assembly Method Applied to Corn Soil Metagenome (1.8 billion reads)



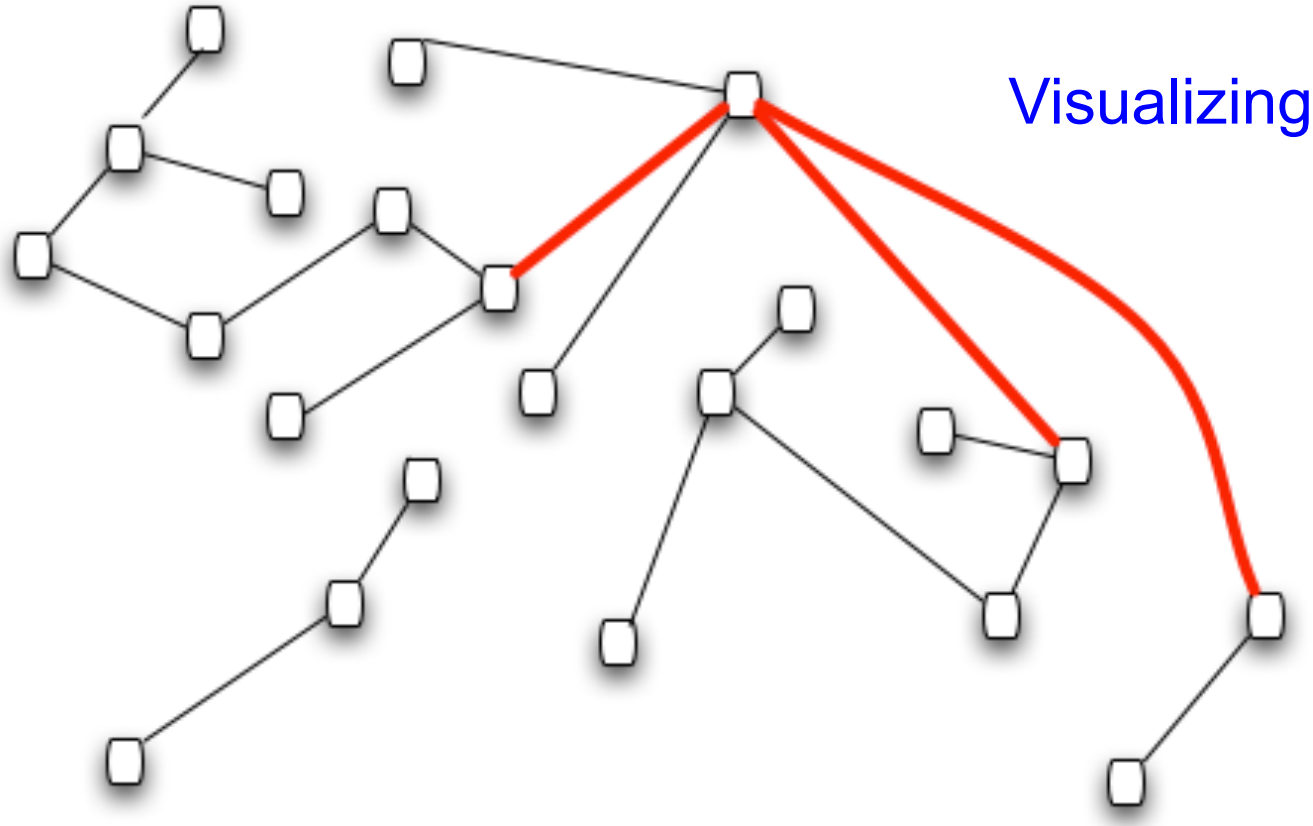
Final Assembly, 25 days, 274 Gb Memory

The Major Advances

1. Memory efficient Bloom filter
2. Digital normalization
3. Partitioning
4. Artifact removal
5. Combine assemblies

By Titus Brown, Adina Howe and more of Titus' group

Problem: more sequence, less assembly
Why?



Illumina sequencing error causes more reads to link

Successful assembly of Iowa corn and prairie metagenomes

For ~$10,000 of sequencing:

	454 Titanium (300 bp raw reads)	Illumina HiSeq (>300 bp contigs assembled)
	166 Mbp	2,545 Mbp
	179 Mbp	3,522 Mbp

Putting it in perspective: Assembly equivalent of:

- ***~1200 bacterial genomes***
- ***Human genome (~3 billion bp)***

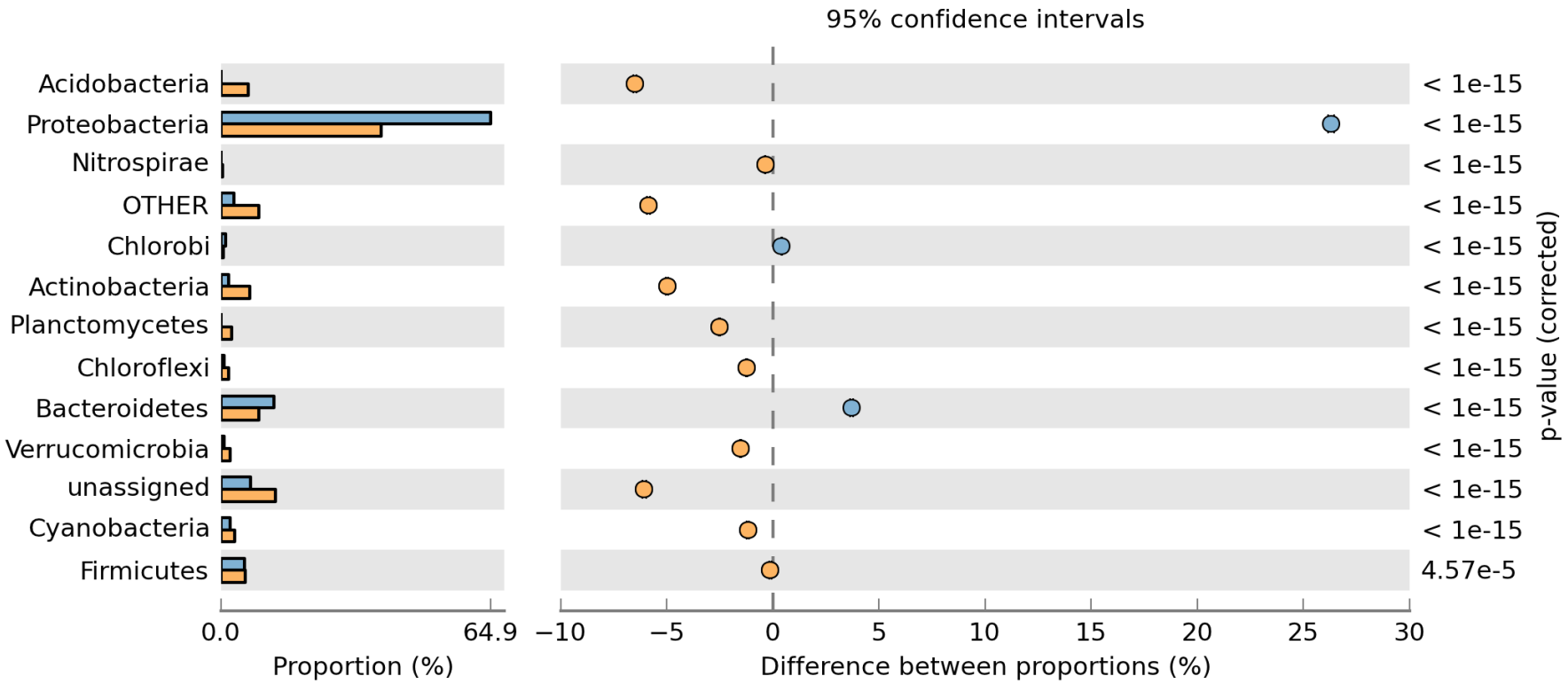
Largest Soil Metagenomes Assembled To Date: Iowa Corn and Prairie

	Assembly Length (bp)*	Reads Used for Assembly	*rplB* genes	% of Lump remaining after artifact removal	% GC
Iowa Corn	2.5 bill	19%	391	73%	62%
Iowa Prairie	3.5 bill	23%	466	84%	59%

- Putatively identified *rplB* genes cover diversity of existing *rplB* references and indicate several novel clades.
- Longest contigs 69,000 and 104,000 bp, respectively

*Contigs > 300 bp

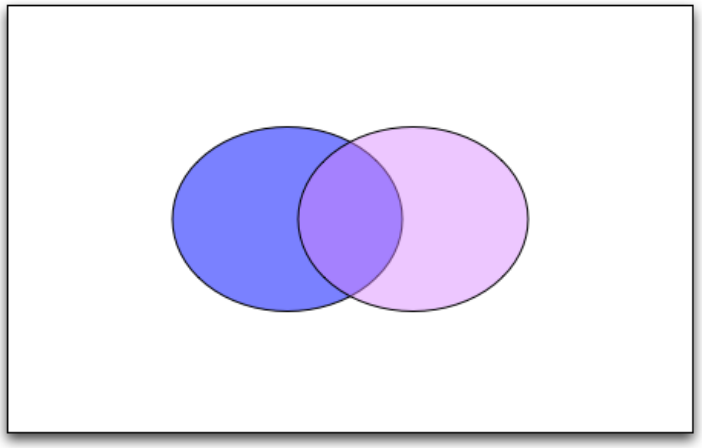
Rhizosphere enrichment



All subsystems = yellow Only N metabolism = blue

Soil is *very* diverse, so how much sequence is needed for "good" assembly

Currently we see < 10x coverage for *deepest* contig. That is insufficient for good short-read assembly!



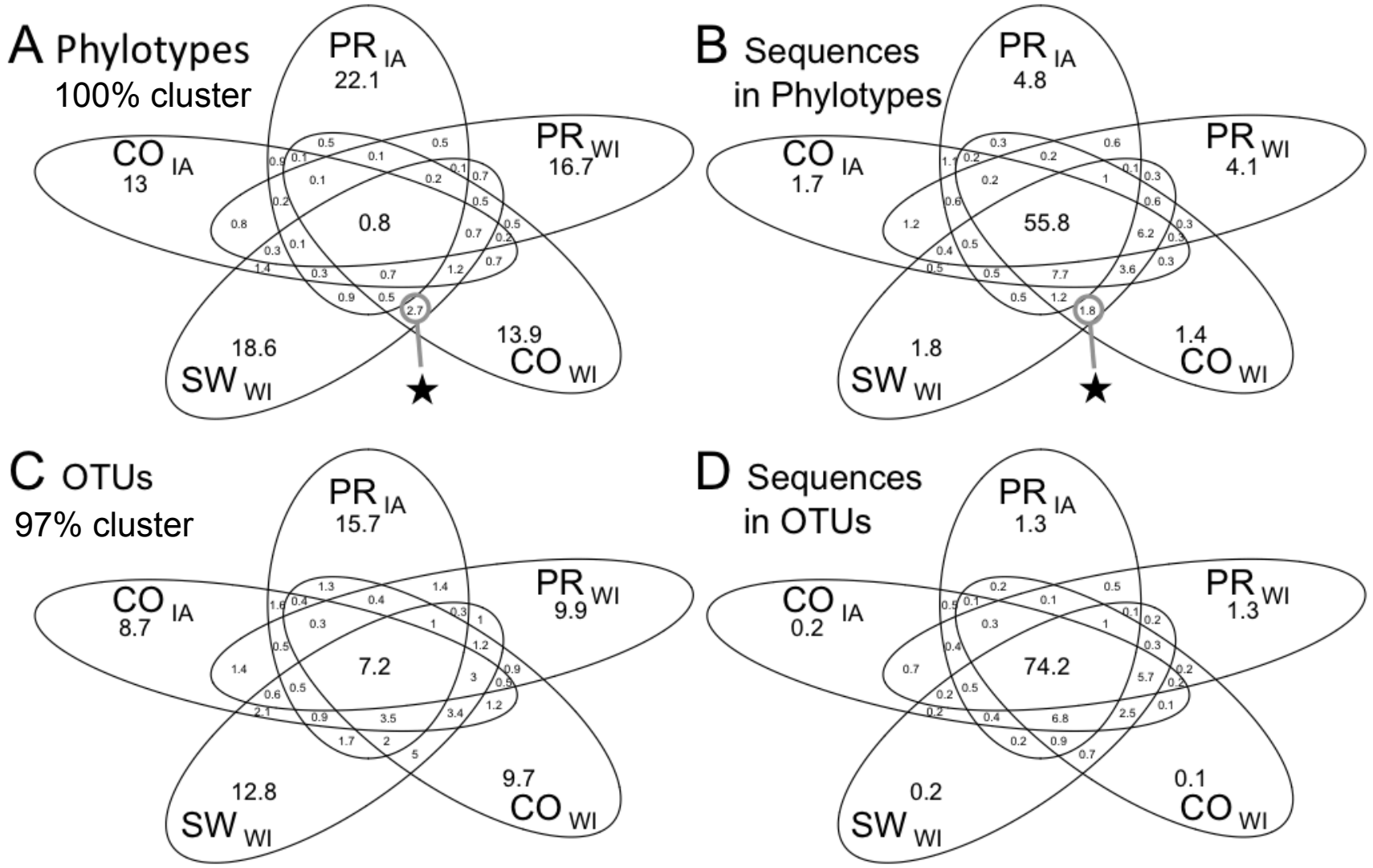
Depth estimates, based on linear extrapolation from k-mer mark/recapture analysis:

Iowa prairie (136 GB):	est 1.26 x
Iowa corn (62 GB):	est 0.86 x
Wisconsin corn (190 GB):	est 2.17 x

- Need 1-2 TB of sequence to see majority of current critters @ ~5-10x coverage. Compare with rumen at avg 56x coverage, ~300 GB.
- Need 2-5 TB of sequence to get good read content on top 80%

Q: What is the size of the soil sample? An aggregate vs an area composite

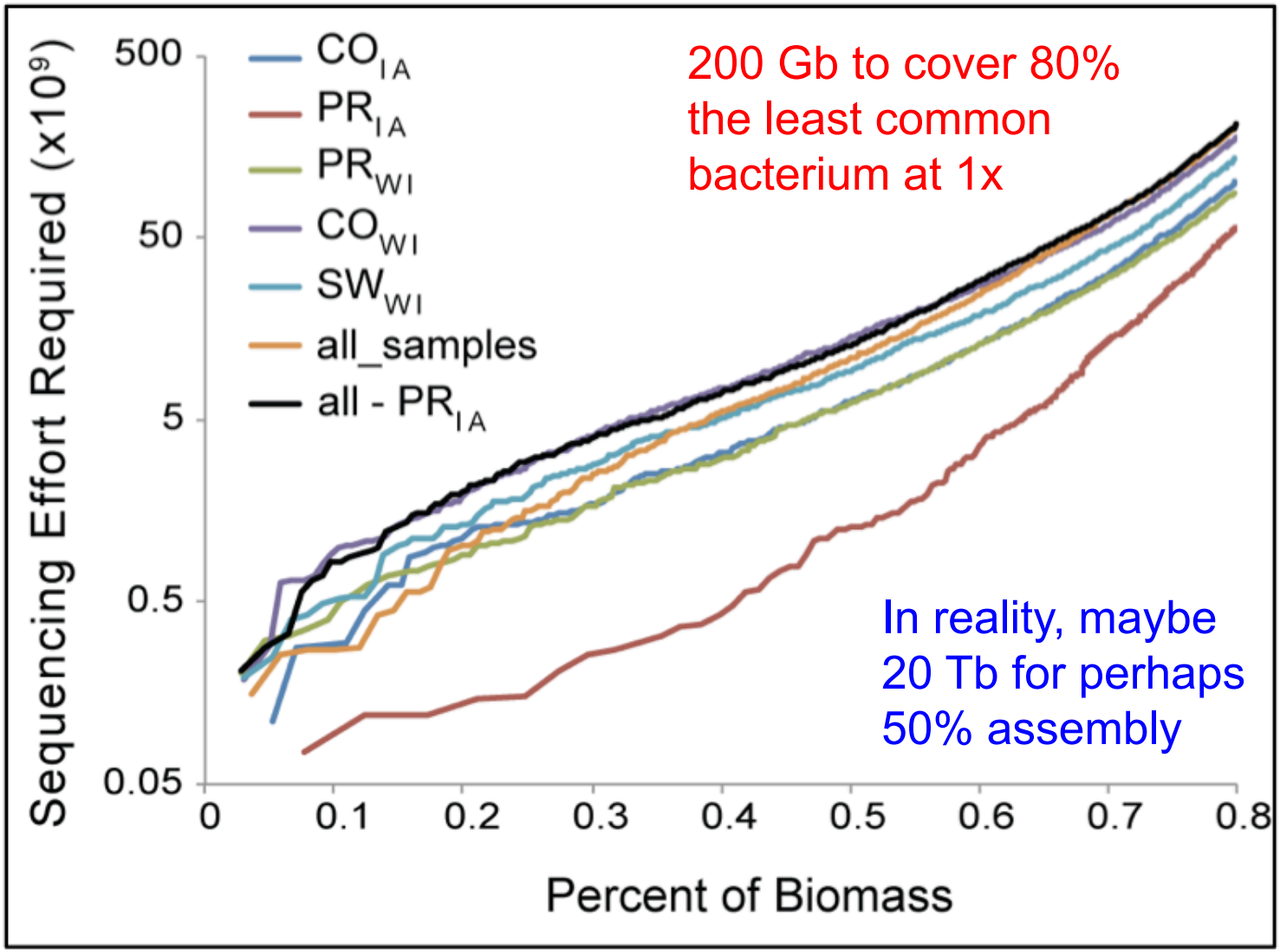
Commonality of individual aggregrate communities over 500 km distance of US midwest prairie soil



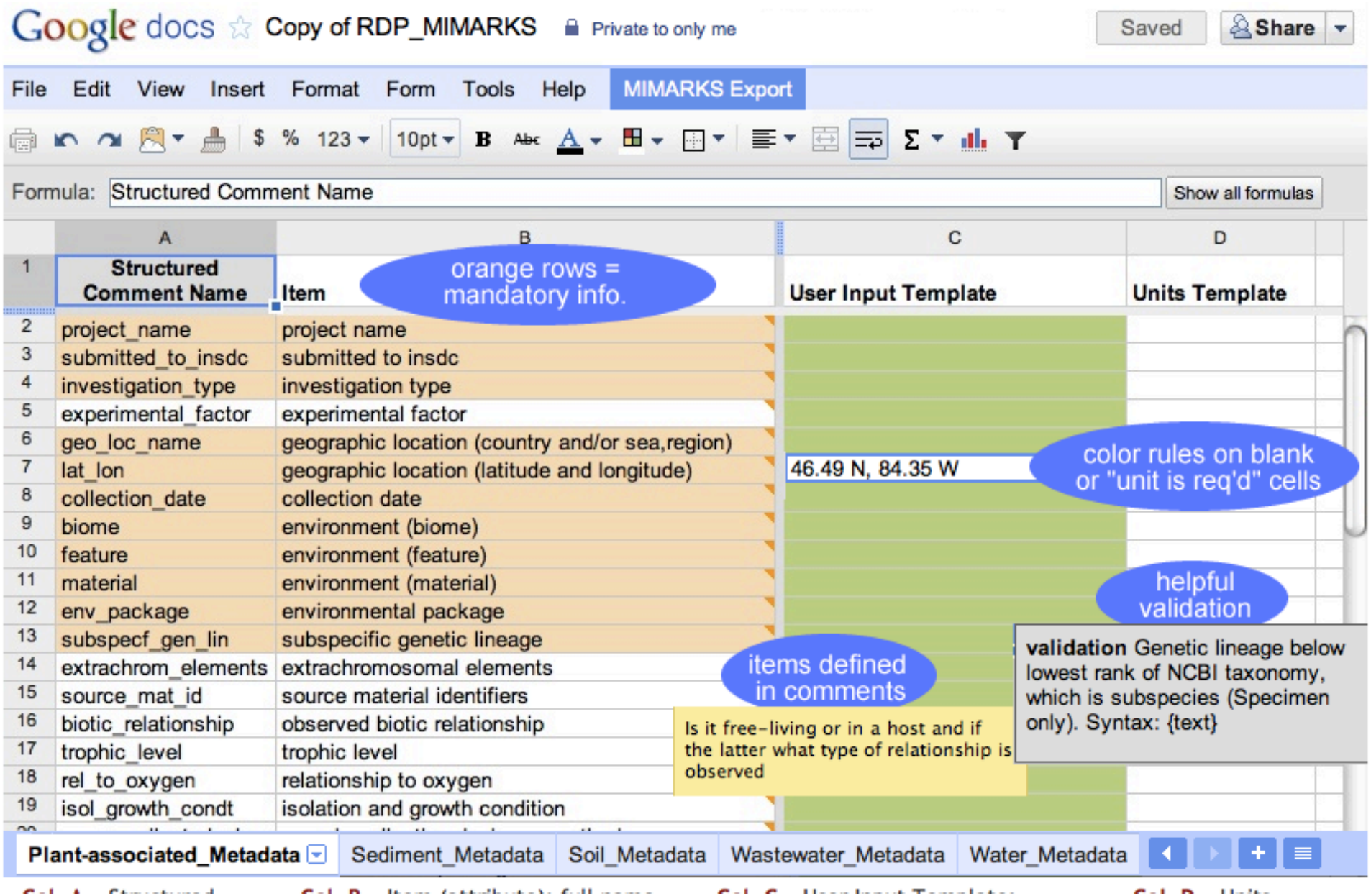
Summary

- ~1-1.5 mil 16S rRNA sequences/ sample
- ~20,000 OTUs (97% cluster distance)/ 0.25 g soil
- But, ~ 40% are singletons
- 3,900 OTUs are common for all samples, contain 74% of sequences
- 4,700 phylotypes are common and contain 56% of the sequences.
- 646,585 total phylotypes and 54,000 OUT's in 5 soil samples
- There are a huge number of phylotypes and OTUs/ 0.25 g so soil, but they are very rare and mostly unique to one sample and make up a very small portion of the total sequences.

Estimated sequencing effort for 1 x coverage, not considering pangenomes



GSC Metadata standards (MIMARKS): Also SRA Prepkit [At RDP}



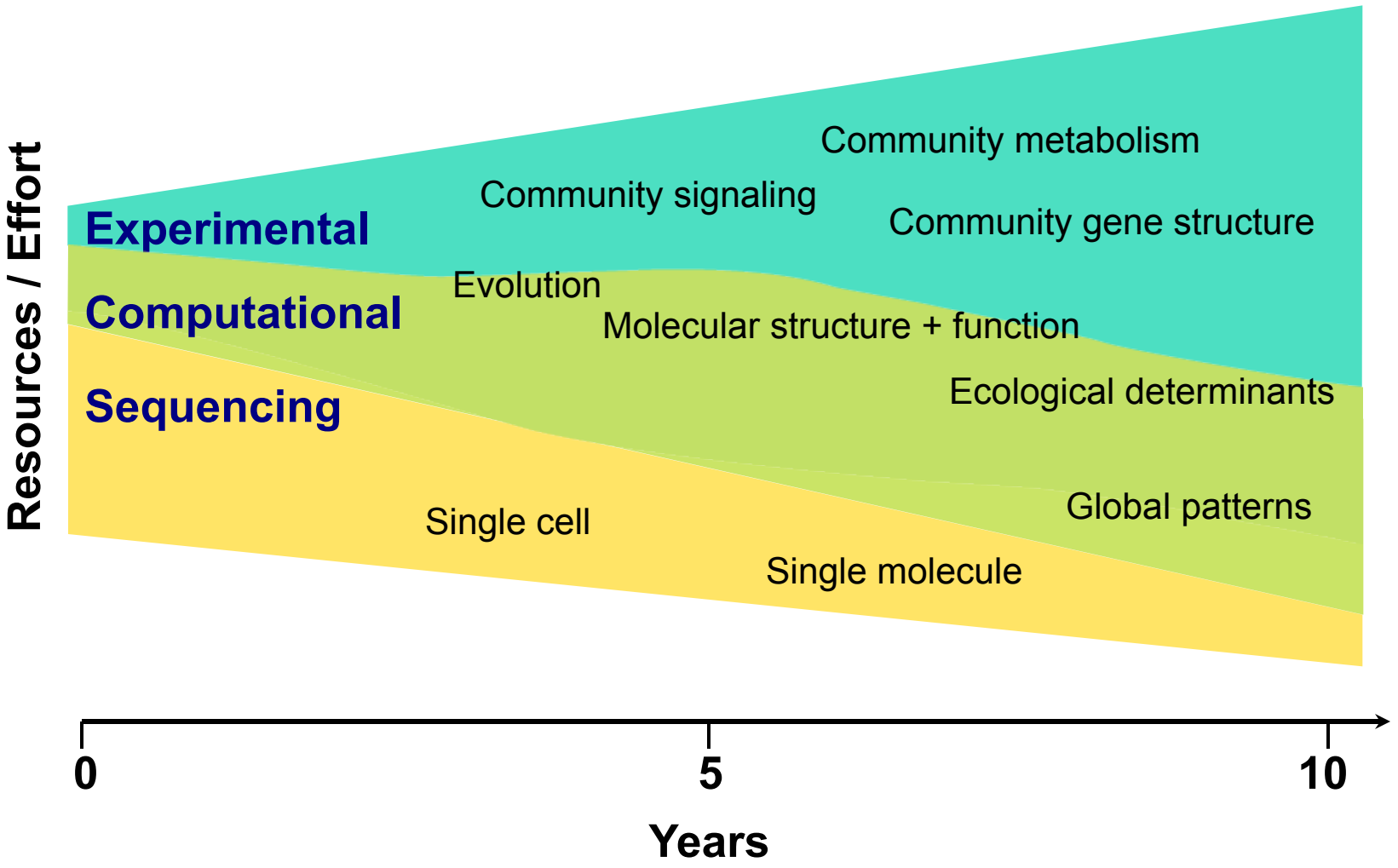
	A	B	C	D
1	**Structured Comment Name**	**Item**	**User Input Template**	**Units Template**
2	project_name	project name		
3	submitted_to_insdc	submitted to insdc		
4	investigation_type	investigation type		
5	experimental_factor	experimental factor		
6	geo_loc_name	geographic location (country and/or sea,region)		
7	lat_lon	geographic location (latitude and longitude)	46.49 N, 84.35 W	
8	collection_date	collection date		
9	biome	environment (biome)		
10	feature	environment (feature)		
11	material	environment (material)		
12	env_package	environmental package		
13	subspecf_gen_lin	subspecific genetic lineage		
14	extrachrom_elements	extrachromosomal elements		
15	source_mat_id	source material identifiers		
16	biotic_relationship	observed biotic relationship		
17	trophic_level	trophic level		
18	rel_to_oxygen	relationship to oxygen		
19	isol_growth_condt	isolation and growth condition		

What is Needed?

- Greater resolution of community composition, faster clock gene(s)
- More computational tools especially for short reads
- More reference genomes AND improved annotation/reference data
- Ability to sample and analyze at soil community scale
- Delineating the active from the inactive
- Ability to link sequence to function
- Much longer reads

A question: how important is it to know in which microbe (species) a ecofunctional gene resides?

Possible Shifts in Emphasis as the Field of Metagenomics Develops



Acknowledgements

Michigan State Univ.

- Kostas Konstantinidis
- WooJun Sul
- Debbie Himes-Yoder
- Shoko Iwai
- Ederson Jesus
- Ryan Penton
- Adriana Lopez
- Jorge Rodrigues
- John Quensen

Burkholderia team:

- J. Davies, L. Eltis, B. Mohn, UBC
- Eshwar Mahenthiralingam, Cardiff
- John LiPuma, U of Michigan

Metagenomics team:

- C. Titus Brown
- Adina Howe
- Patrick Chain
- Erick Cardenas
- Jiarong Guo
- Aaron Garoutte
- Janet Jansson
- Susannah Tringe
- Tijana Glavina del Rio
- Rachel Mackelpring
- Gina Lamendella

Amazon Microbial Observatory:

- Vivian Pellizari,
- Sui Tsai
- Brigitte Feigl
- Jorge Rodrigues
- Klaus Nuesslein
- Brendan Bohannan

RDP team at MSU

- Jim Cole
- Benli Chai
- Ryan Farris
- Jordan Fish
- Donna McGarrell
- Qiong Wang