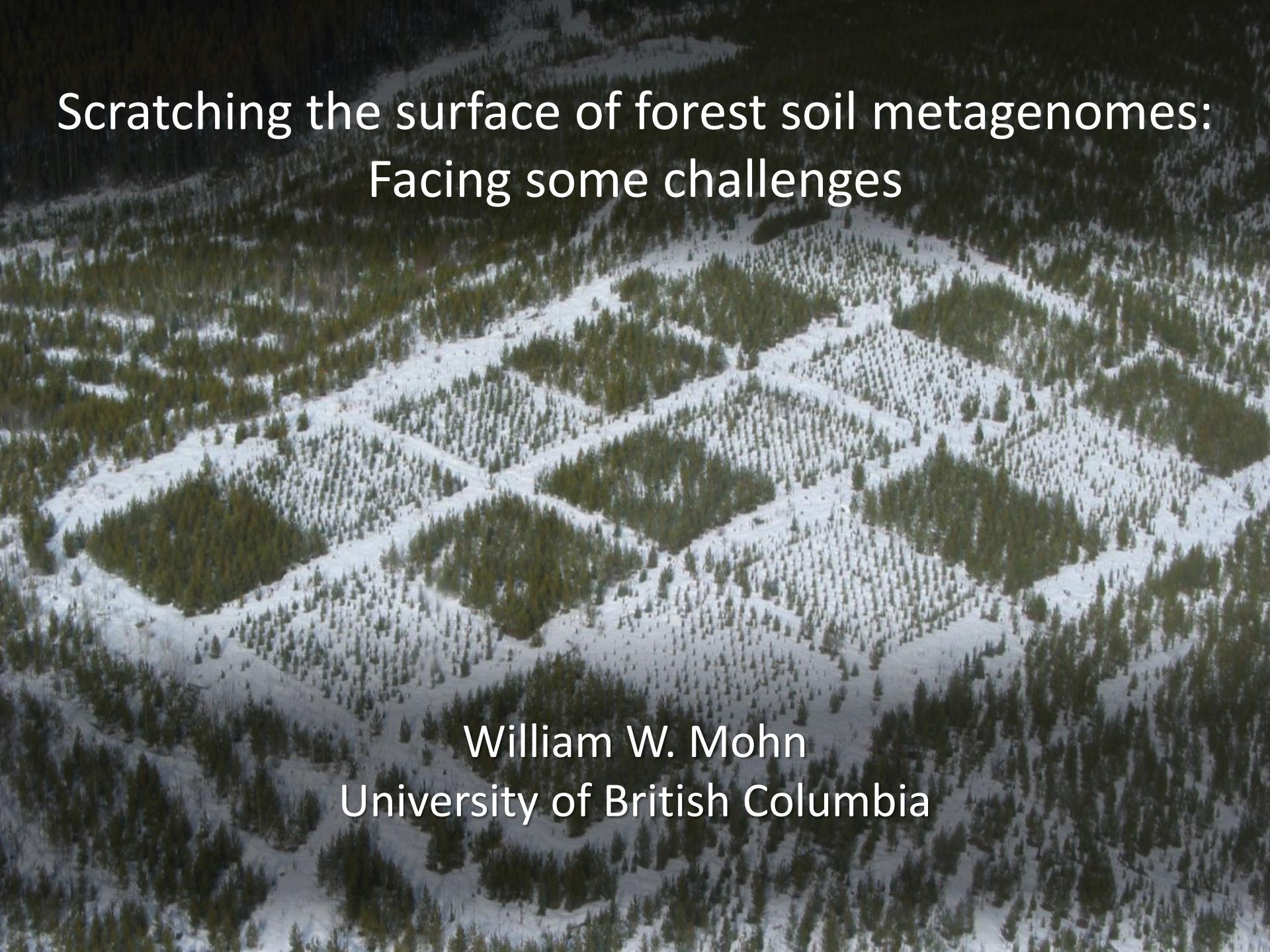


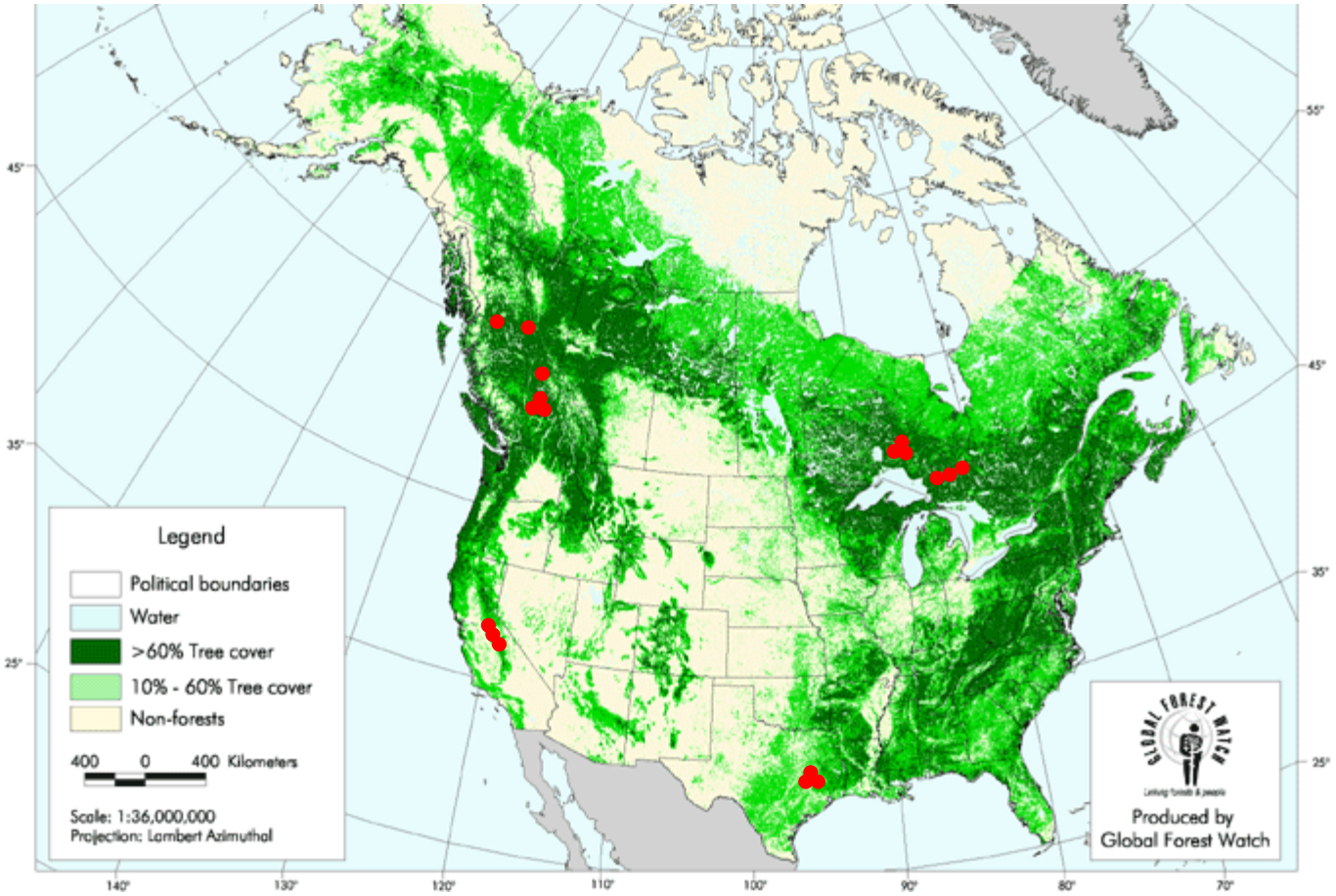
Scratching the surface of forest soil metagenomes: Facing some challenges

William W. Mohn
University of British Columbia

Long Term Soil Productivity (LTSP) Study

- Robust field experiment examining harvesting disturbances
- One focus is effects of organic matter removal
- Metagenomic approach; little precedent for forest soil

LTSP sites under investigation



Objectives of metagenomic analysis:

1. To compare metabolic potential of communities among diverse ecozones
2. To evaluate effects of organic matter removal among ecozones

Assembly of O'Connor Lake metagenome

From 7 lanes of Illumina HiSeq 75-b reads, 71 Gb

	Gbases sequenced	Mbases assembled	Percent reads assembled	Contigs >1kb
OM0C0-O	11.1	15.8	1.3	325
OM1C0-O	10.1	6.1	0.5	335
OM2C0-O	9.5	2.1	0.3	89
OM0C0-M	10.6	54.3	4.6	9818
OM1C0-M	10.5	64.7	2.6	1758
OM2C0-M	8.7	18.4	2.3	826
OM3C0-M	10.3	40.1	3.3	3231
All Organic layer samples	30.7	13.9	0.8	396
All Mineral layer samples	40.1	48.3	3.1	4222
Human gut *	4.5 - 7.3	19 - 237	14 - 56	N/A

* Quin et al 2010. Nature 464, 59-65

- Overall conclusion: inadequate assembly

Sequencing platform alternatives

Platform	Total sequence	Reads	Read length	No. genes (coverage)	Cost
Shotgun Illumina HiSeq	27,000 Mb	180 M	150 b	??? (10%)	$4,000
Shotgun 454 Titanium	750 Mb	1.5 M	500 b	~1 M (50%)	$10,000
Multiplexed fosmids (HiSeq)	16 Mb	400 fosmids	40 kb	16 K (100%)	$8,000

Outstanding question:
Can short, un-assembled HiSeq reads provide a meaningful gene inventory?

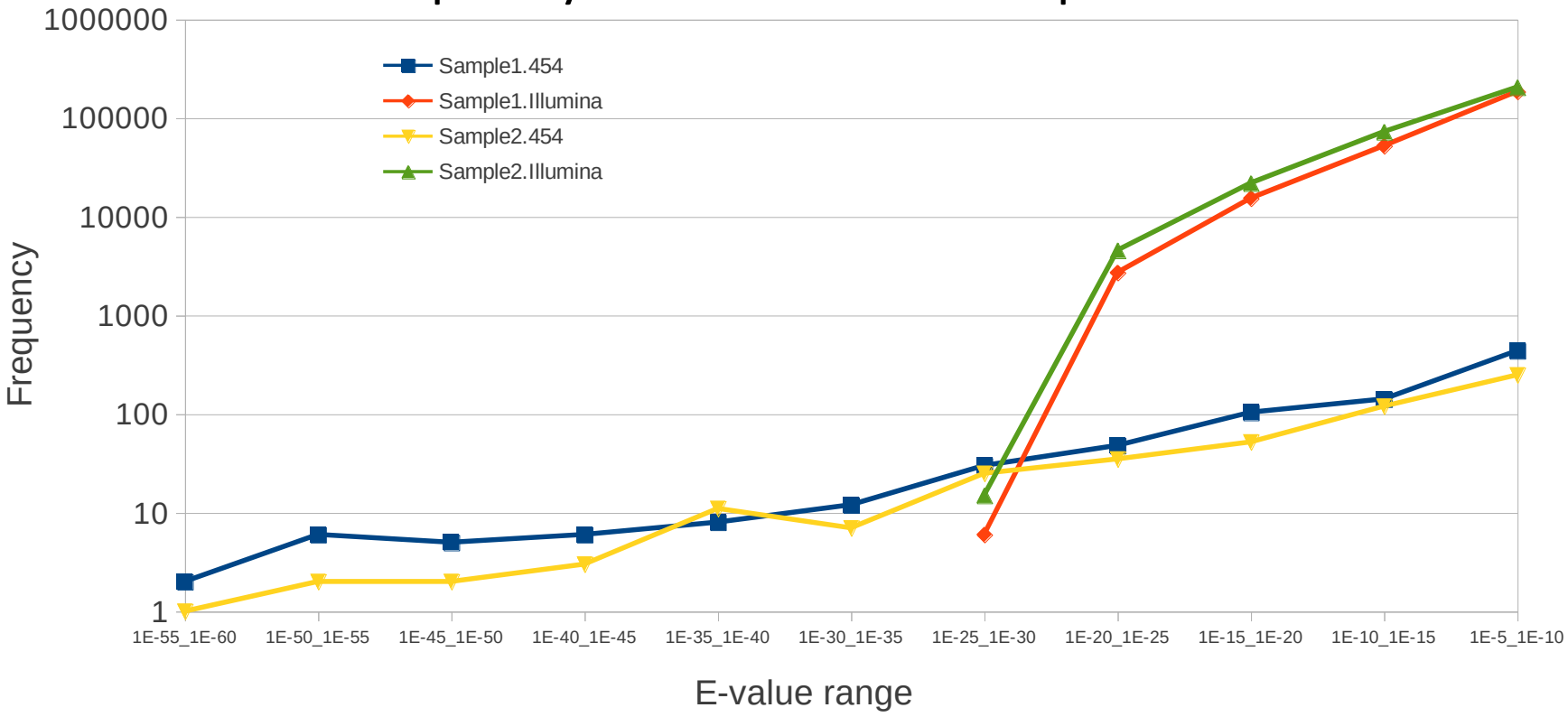
454 vs Illumina, empirical comparison of genes identified

- Un-assembled reads Blasted against Fungal Oxidative Lignin enzymes (FOLy) database (2516 proteins)
- E value threshold, 10^{-5}

	Reads (Million)		Hits (evalue <1E-5)		Hits /million reads	
	454	Illumina	454	Illumina	454	Illumina
Sample1	0.578	419	801	256749	1385	613
Sample2	0.426	373	507	305383	1190	818

- Illumina reads efficiently find genes
- Outstanding question: How accurate is gene identification with Illumina reads?

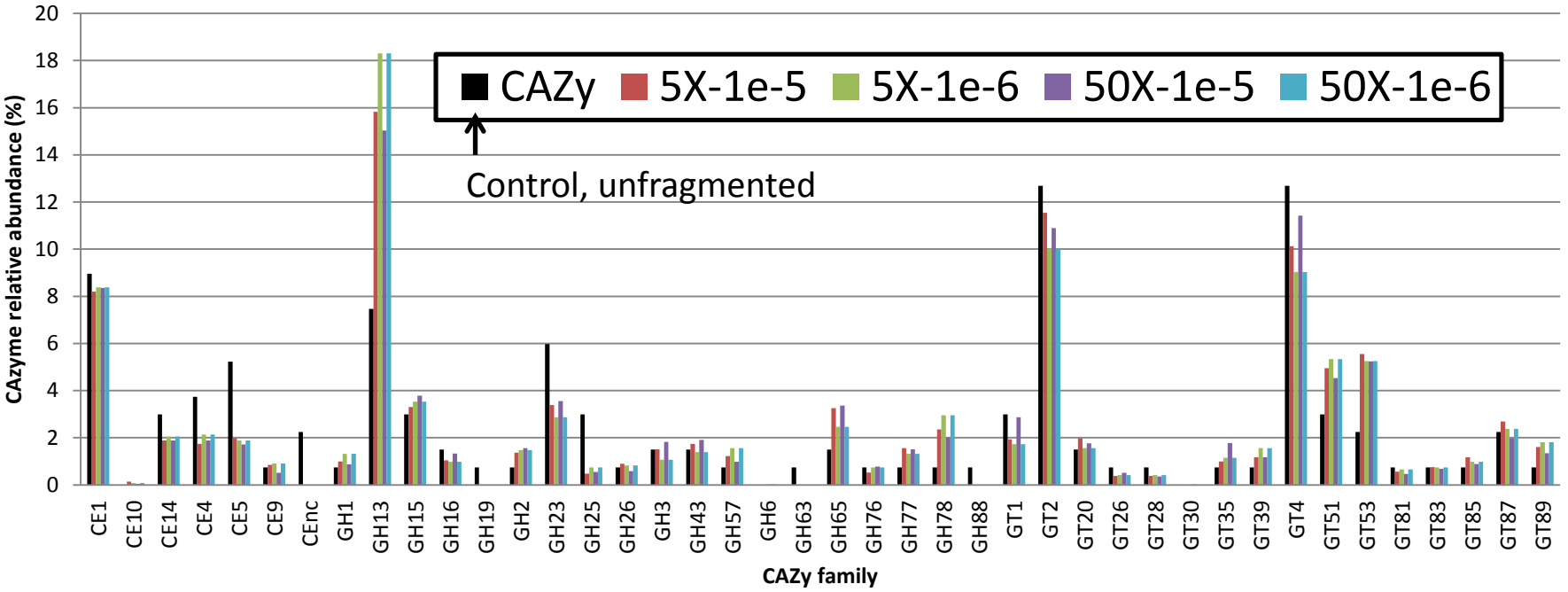
454 vs Illumina, empirical comparison of genes identified: Frequency distribution of blastp E values



- Illumina identifies far more genes meeting minimal criterion
- 454 identifies small number of additional genes with very high confidence
- We are developing hidden Markov models for gene families of interest, which will identify genes with greater accuracy than blast

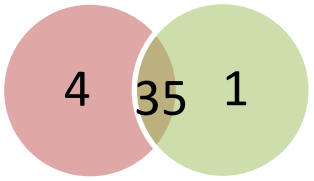
In silico validation

Rhodococcus genome randomly fragmented into 75-b "reads" and Blasted against the Carbohydrate Active Enzyme (CAZy) database

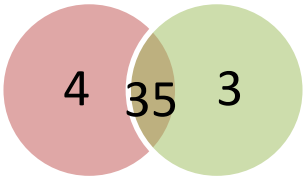


Families detected: Predicted (Red) vs Detected (Green)

2X, 1e-5



2X, 1e-6

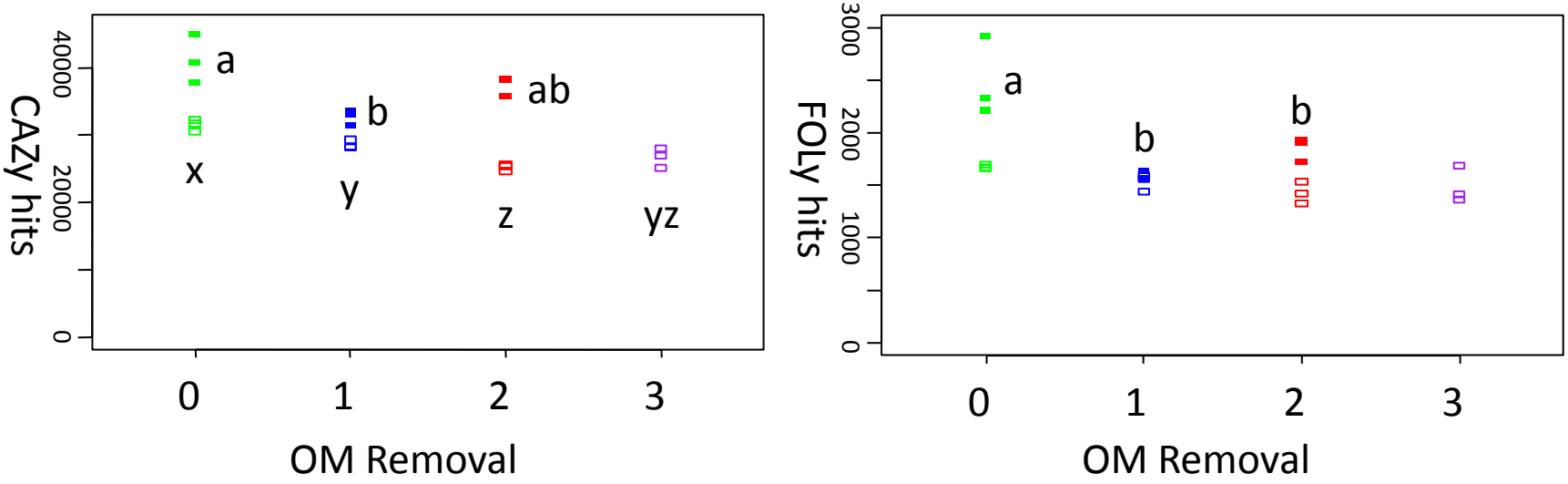


20X, 1e-5

20X, 1e-6

Effect of OM removal on relative abundance of lignocellulose decomposition enzymes

Normalized numbers of CAZy and FOLy hits in un-assembled metagenomes

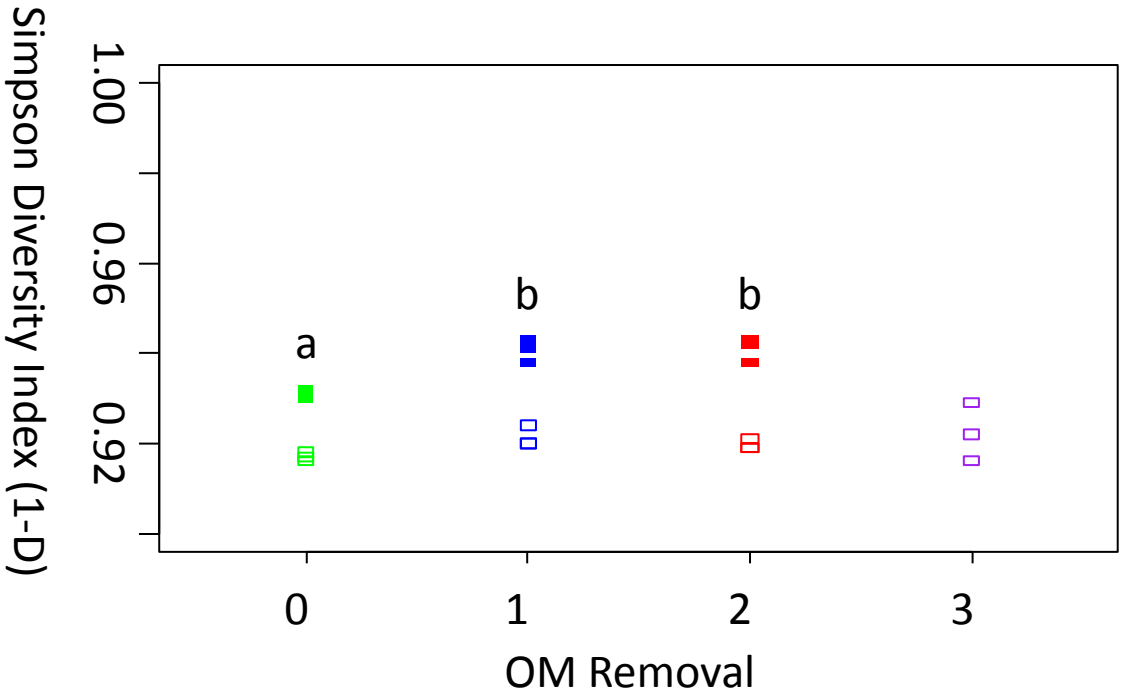


Solid symbols, organic horizon; open symbols, mineral horizon; letters indicate significant differences ($p < 0.01$) according to ANOVA; n = 3 metagenomes

- Harvesting reduces abundance of CAZy genes in both horizons
- Harvesting reduces number of FOLy genes in organic horizon

Effect of OM removal on diversity of lignocellulose decomposition enzymes

CAZy family diversity in un-assembled metagenomes



Solid symbols, organic horizon; open symbols, mineral horizon; letters indicate significant differences ($p < 0.05$); n = 3 metagenomes

- Harvesting increases diversity in organic horizon

Analyzing un-assembled short reads

- Genes can be efficiently found among reads as short as 75 b
- Meaningful gene identification is possible
- Functional potential can be compared among communities
- Large Illumina datasets advantageous
- Illumina read lengths of 150 b now available, and 250 b is coming

Difficulty identifying catabolic genes 1

Best blastp hits of putative ortholog of *ditA1,* encoding a key enzyme involved in resin acid degradation

Hit	Organism	Identity (%)	Similarity (%)	Alignment length	Evalue
gi|346421700 CmtAb	*Pseudomonas* sp. 19-rlim	49.4	63.9	393	4.44E-130
gi|386285754 p-cumate dioxygenase large subunit (CmtAb)	Gamma proteobacterium BDW918	47.4	66.1	392	8.51E-130
gi|353193549 Benzoate 1,2-dioxygenase	*Mycobacterium rhodesiae* JS60	48.6	65.2	399	2.46E-129
gi|373856464 Benzoate 1,2-dioxygenase	*Bacillus* sp. 1NLA3E	46.5	64.9	396	1.91E-128
gi|148548110 Ring hydroxylating dioxygenase subunit alpha	*Pseudomonas putida* F1	49.9	62.7	397	2.79E-127
gi|383822102 2-chlorobenzoate 1,2-dioxygenase	*Mycobacterium phlei* RIVM601174	49.0	62.5	400	2.92E-127
gi|386289217 p-cumate dioxygenase large subunit (CmtAb)	Gamma proteobacterium BDW918	48.2	63.6	390	4.72E-127
gi|91782278 p-cumate dioxygenase large subunit (CmtAb)	*Burkholderia xenovorans* LB400	48.9	62.3	401	2.10E-125
gi|148554676 ring hydroxylating dioxygenase subunit alpha	*Sphingomonas wittichii* RW1	49.9	63.5	397	1.08E-123
gi|331698063 2-chlorobenzoate 1,2-dioxygenase	*Pseudonocardia dioxanivorans* CB1190	46.9	61.2	397	6.63E-121

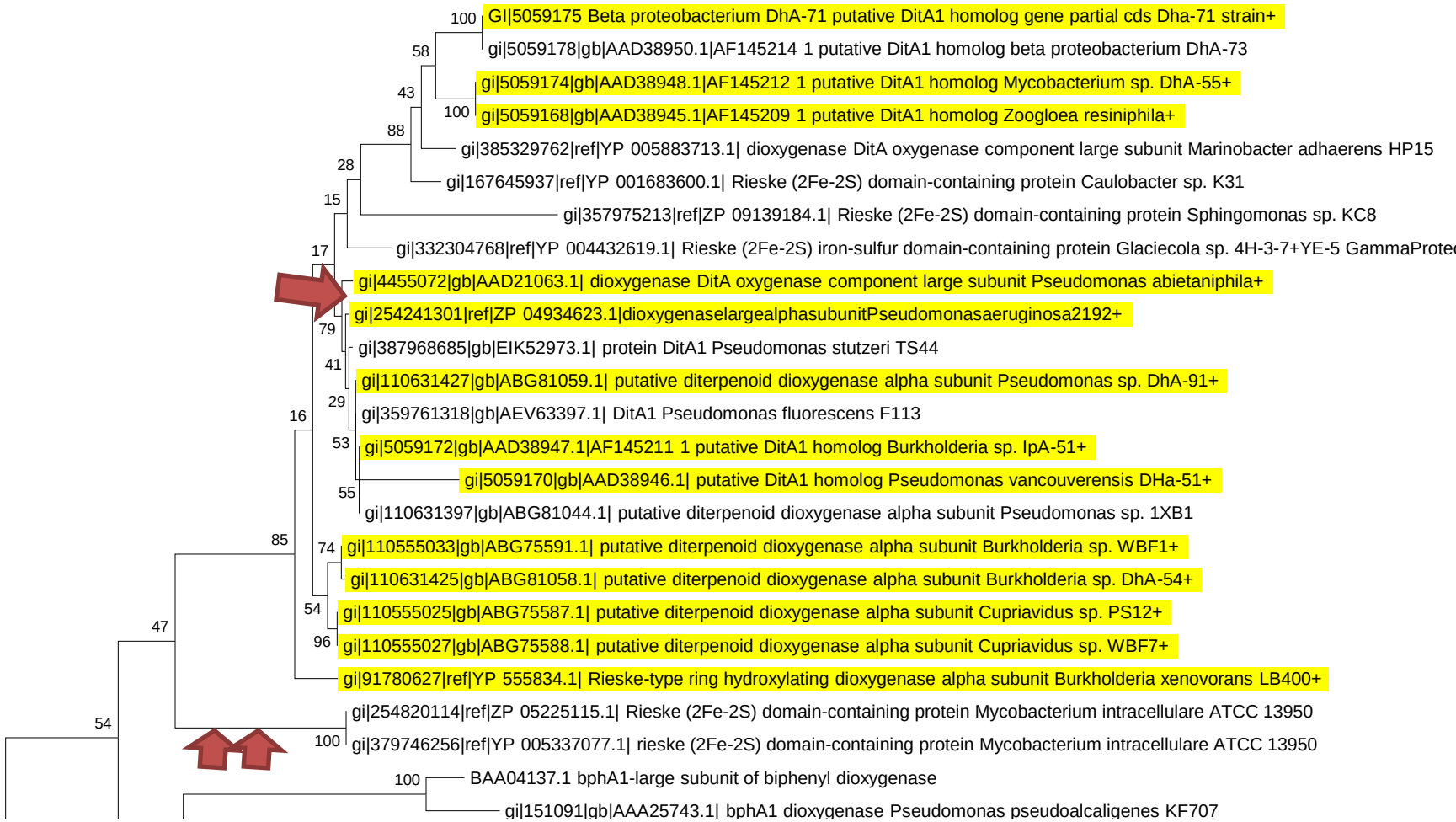
- Annotation of hits likely incorrect
- Leads to misidentification

Difficulty identifying catabolic genes 2

Best blastp hits of another putative ortholog of *ditA1,* encoding a key enzyme involved in resin acid degradation

Hit	Organism	Identity (%)	Similarity (%)	Alignment length	Evalue
gi|254820114 Rieske (2Fe-2S) domain-containing protein	Mycobacterium intracellulare MOTT-02	96.1	98.4	128	1.69E-83
gi|126436767 Rieske (2Fe-2S) domain-containing protein	Mycobacterium sp. JLS	93.0	97.7	128	1.09E-79
gi|167645937 Rieske (2Fe-2S) domain-containing protein	Caulobacter sp. K31	45.5	65.7	134	5.31E-28
gi|332304768 Rieske (2Fe-2S) iron-sulfur domain-containing protein	Glaciecola sp. 4H-3-7+YE-5	46.2	61.5	130	3.04E-27
gi|5059172 putative DitA1 homolog	Burkholderia sp. IpA-51	44.5	64.8	128	7.02E-27
gi|387968685 protein DitA1	Pseudomonas stutzeri TS44	45.4	63.8	130	2.13E-26
gi|378951304 ditA1 gene product	Pseudomonas fluorescens F113	44.5	64.8	128	2.51E-26
gi|91780627 Rieske-type ring hydroxylating dioxygenase alpha subunit	Burkholderia xenovorans LB400	44.9	63.8	127	3.67E-26
gi|5059178 putative DitA1 homolog	Beta proteobacterium DhA-73	44.0	64.9	134	8.92E-26
gi|5059176 putative DitA1 homolog	Beta proteobacterium DhA-71	44.0	64.9	134	9.87E-26

- Annotation of closest hits is vague
- Less similar hits reflect correct function
- Leads to uncertain identification



Phylogeny can assist identification

Highlighted genes have experimentally verified function

Phylogenetic association of 3 putative *ditA1* orthologs from a metagenome

One falls within verified cluster – confident identification

Two fall within unverified cluster – uncertain identification

Identifying catabolic genes

- Accurate gene identification is vital for interpretation of metagenomic data
- Annotation is not consistent in databases
- Automated annotation creates and massively propagates errors
- Phylogeny can assist identification, but requires a robust set of diverse, validated reference sequences
- Curated databases (eg, CAZy, FOLy) important resources, but lacking for most types of catabolic genes/proteins