Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization

Luis H. Orellana[a], Joanne C. Chee-Sanford[b], Robert A. Sanford[c], Frank E. Löffler[d,e], and Konstantinos T. Konstantinidis[a,#]

Georgia Institute of Technology, Atlanta, Georgia, USA[a]; US Department of Agriculture—Agricultural Research Service, Urbana, Illinois, USA[b]; University of Illinois, Urbana, Illinois, USA[c]; University of Tennessee, Knoxville, Tennessee, USA[d]; Bioscience Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA[e].

Supporting information

Experimental Procedures

Functional annotation of short-reads using SEED in soil and fresh water metagenomes

SEED functional categories examined in detail for pathways of secondary metabolism included the terms “Iron acquisition and metabolism”, “Membrane transport”, “Metabolism of aromatic compounds”, “Motility and chemotaxis”, “Nitrogen metabolism”, “Phosphorus metabolism”, “Potassium metabolism”, “Secondary metabolism”, and “Sulfur metabolism”. Categories having above 0.01% relative abundance, on average, for top and deep soil layers in both sites were used for the determination of coefficient of variation between and within samples. The same annotation strategy was used for Lake Lanier metagenomes over the course of 1 year (1101B, 1104A, 1107A, and 1108A) and 2 years (1007B, 1008A, 1009A, 1010A, 1101B, 1104A, 1107A, and 1108A) (1).

Identification and analyses of 16S rRNA gene sequences

Short-read sequences encoding 16S rRNA gene fragments were extracted from each metagenome by using SortMeRNA (2) and their taxonomy was assigned using RDP classifier (cutoff 50)(3). In addition to the taxonomic annotation, operational taxonomic units (OTUs) were determined using a closed-reference OTU picking strategy as implemented in QIIME (4) using the same recovered 16S rRNA gene fragments. Sequences were clustered into OTUs at 97% similarity using UCLUST (Edgar, 2010) and using references from SILVA database v111 (Quast *et al.*, 2013).

Identification of glycoside hydrolase genes

Glycoside hydrolase (GH) protein sequences in unassembled metagenomes were detected by querying the short-reads against the dbCAN database (5) using BLASTx (default settings and minimum 60% identity and 70% query coverage for a match). MAGs harboring GH proteins were detected using BLASTp (default settings and minimum 60% identity and 70% query

coverage for a match) against the previous database. In both cases, results were summarized based on the family classification from the CAZy database (6) and categories proposed previously (7).

Phylogenetic trees and placement of short-reads

Protein reference and assembled sequences were aligned using ClustalΩ (8) with default parameters. Resulting alignments were used to build phylogenetic trees in RAxML v8.0.19 (9). Identified short-reads encoding the protein of interest were extracted from soil metagenomes using ROCker (BLASTx) and their protein-coding sequences were predicted using FragGeneScan (10). The latter sequences were added to the corresponding protein alignment using MAFFT ("addfragments") (11) and were placed in the corresponding phylogenetic tree using RAxML EPA (-f v option) (12).

Visualization and clade classification of reads placements

The visualization of the generated jplace files (13) was performed using the "JPlace.to_iToL.rb" script from the enveomics collection (14) and subsequently visualized on iTol (15). Quantification of the number of reads assigned to a specific clade (e.g., to distinguish between *nxrA* or *narG* reads) was done using the "JPlace.distances.rb" script, also available in the enveomics collection.

To quantify *nirK* gene fragments assigned to specific clades we used the clades previously proposed (16). The same process as described above for *nxrA/narG* was repeated except that all reads detected by ROCker models (I+II, III and *Thaumarchaeotea*) were used for classification. Clade IV (e.g., *Actinobacteria*) was intentionally omitted from this analysis due to the limited number of available genomes harboring *nirK*, which limited the development of a robust ROCker model.

Results

Given the different amounts of organic matter (OM) observed between the two sites and soil layers, we hypothesized that there would be site-specific microbial communities involved in

the cycling and degradation of carbonaceous material. Specifically, we sought to find a link between the soil type and the dynamics of genes encoding enzymes (e.g., glycoside hydrolases) directly involved in the hydrolysis of glycosidic bonds in plant-derived carbon biomass. Even though genes encoding glycoside hydrolases (GH) showed a slight increase (8%) at the end of the year in the top soil depth of Havana, stable GH gene abundances were observed throughout the year within the same soil depth at each site (Fig. S9). For instance, GH genes encoding amylolytic enzymes showed high and stable abundance in both soils (up to 0.16% and 0.19% of total GH genes in Havana and Urbana, respectively), regardless of the differing soil texture and quantity of soil organic matter. Both sites showed significantly higher relative abundance of GH genes on the top compared to the deeper soil layers (two tailed *t*-test, $p < 0.001$) (Fig. S9), and Urbana showed, on average, 20.4% higher relative abundance of GH genes compared to Havana.

Taxonomic compositions of agricultural soils

For Havana, *Proteobacteria* (~40%), *Acidobacteria* (~18%), and *Actinobacteria* (~17%) represented the most abundant phyla in both the 0-5 and 20-30cm depths. *Bacteroidetes*, *Actinobacteria*, and *Firmicutes* were distinctive in the top soil metagenomes (*P*-value adjusted < 0.0001), whereas *Nitrospirae*, *Thaumarcheota*, and *Euryarchaeota* were characteristic of the deeper soil layer (*P*-value adjusted ≤ 0.001), in agreement with functional annotation results (Fig. S3b). At the order level, *Sphingomonadales*, *Sphingobacteriales*, *Actinomycetales*, and *Solirubrobacterales* were distinctive in the top layer, and *Nitrosopumilales*, *Neisseriales*, *Nitrospirales*, *Bacillales*, and *Rhodospirillales* were more abundant in the deeper metagenomes (*P*-value adjusted ≤ 0.0001). For Urbana, *Proteobacteria* (32%), *Actinobacteria* (22%) and *Acidobacteria* (~19%) represented the most abundant phyla in both depths (Fig. S3b). *Bacteroidetes* and *Gemmatimonadetes* were more abundant in the top layer, whereas *Verrucomicrobia*, *Chloroflexi* and *Thaumarchaeota* were distinctive of the deep layer (*P*-value adjusted < 0.001). At the order level, *Flavobacteriales*, *Sphingomonadales*, *Caulobacterales*,

Xanthomonadales, *Solirubrobacterales*, and *Burkholderiales* where characteristic of the top layer, whereas *Anaerolineales*, *Nitrospirales*, and *Nitrososphaerales* were distinctive of the lower layer (*P*-value adjusted < 0.05). Comparison of alpha diversity (Chao-Shen entropy index), based on the taxonomy at the phyla and order levels of the recovered 16S rRNA gene fragments, showed significant differences between the two soil layers in Urbana. For Havana, significant differences in alpha diversity were only detected at the phylum level between top and deep soils (Fig. S1b).

Using a closed reference OTU picking strategy, over 61% of the recovered 16S rRNA gene sequences in each site were clustered into an average of 3,482 and 2,170 OTUs (97% similarity) per sample in Havana and Urbana, respectively (defined at 97% 16S rRNA gene sequence identity). OTU projections per sample (Chao1 index) showed that Havana harbored more OTUs than Urbana soils (two-tailed *t*-test, $P < 0.01$). In addition, the latter estimates revealed that the detected OTUs in Havana ranged from 46% to 73% of the estimated total number of OTUs depending on the sample considered, whereas these values ranged from 49% to 82% in Urbana. Both sites shared 19.9% of the total detected OTUs (n=12,125) whereas 42.6% and 37.5% OTUs were specific to Havana and Urbana, respectively. A comparison of top vs. deep OTUs showed that in Havana, statistically overrepresented OTUs (Log 2-fold >=2 and p-adjusted <0.01) in the top layer belonged to *Actinobacteria* (25.3%), *Alphaproteobacteria* (22.6%), and *Chloracidobacteria* (16.6%) whereas enriched OTUs in the deep layer belonged to *Gemmatimonadetes* (16%), *Nitrospirae* (10.2%), and *Thaumarchaeota* (10.2%). Similarly, overrepresented OTUs in the top layer of Urbana samples belonged to *Alphaproteobacteria* (46.5%), *Thermoleophilia* (14%) and *Actinobacteria* (11.6%) whereas enriched OTUs in the deep layer were *Actinobacteria* (25.3%), *Alphaproteobacteria* (22.6%), and *Chloracidobacteria* (16.6%).

Denitrification genes

Hallmark denitrification genes showed stable abundances throughout the year but differences between soil layers and sites. For instance, in Havana, nitrate reductase (*narG*), nitrite

reductases (*nirK* and *nirS*), and nitrous oxide reductase (*nosZ*) showed significantly higher abundance in the deep compared to the top soil layer (Fig. S5). Even though both nitrite reductases were more abundant in the deeper soil layer, *nirK* was, on average, 9.5 and 6.1 times more abundant than *nirS* in the top and deep soil layers, respectively. On the other hand, opposite abundance patterns for denitrification genes were observed for Urbana. For instance, *narG*, *nirK*, *nirS*, and *norB* were statistically significantly more abundant in the surface soil layer compared to the deep soil layer (Fig. S5), probably as a result of the contrasting edaphic factors between the sites. In addition, in both sites, *nrfA* showed the opposite abundance patterns compared to denitrification genes. Consistent with our previous reports from composite soil samples from the same agricultural soils (17), clade II, or atypical *nosZ*, gene fragments showed higher abundance throughout the year in both sites. In Havana, clade II *nosZ* gene fragments were, on average, ~7 times more abundant than their clade I counterparts in both soil layers across the year. Interestingly, similar trends were observed in Urbana where atypical *nosZ* gene fragments were on average 9.7 and 15.9 times more abundant than their typical counterparts in the top and deep soil layers throughout the year, respectively.

Recovered populations from metagenomes

The assembly and binning resulted in 69 population MAGs in total from both sites, having over 50% completion and less than 20% of contamination based on the presence of 104 and 26 single-copy bacterial and archaeal genes, respectively. These genes might not always be present in all microbial lineages, therefore, gene content and completeness values were likely underestimated. The use of relatively low stringency criteria was due the low fraction of assembled metagenomic reads. Even at this level of stringency, only 69 MAGs, representing ~30% of the total MAGs obtained, were selected. The remaining MAGs were even more incomplete or contaminated despite efforts to refine binning by performing a second round of assembly (see Experimental Procedures for details). Genome sizes ranged from 1.1 to 6.7 Mbp, and G+C% content varied from 35 to 72% (Table S6). Inferred taxonomy revealed that most

MAGs represented members of *Proteobacteria*, *Acidobacteria*, and *Actinobacteria* in both soils whereas *Verrucomicrobia* and *Gemmatimonadetes* were characteristic of Urbana and Havana, respectively. As expected, genomic comparisons based on average amino acid (AAI) values (18) revealed that most of the obtained MAGs likely represented novel organisms when compared to the NCBI prokaryotic genome database (Table S6). For Havana, only 4.3% of the MAGs had AAI values greater than ~65% (i.e., shared genus) (19) compared to their close relatives. A similar trend was observed for Urbana MAGs where none of the MAGs likely corresponded to known genera. However, closely related MAGs, most likely representing member of the same genus (i.e., sharing AAI >65%), were detected in both sites. For instance, in Havana, *Nitrospira* MAGs HD017 and HD021 shared 81.69% AAI (SD: 15.43%, from 2288 proteins); *Gemmatimonadetes* MAGs HD002 and HD027 shared 77.47% AAI (SD: 15.80%, from 2429 proteins). In Urbana, *Verrucomicrobia* MAGs UD002 and UD007 shared 82.65% AAI (SD: 16.82%, from 1713 proteins). Several MAGs were specific to each site but shared relatively high AAI values such as the *Thaumarchaeota* MAGs HD032 and UD001 which shared 76.79% AAI (SD: 14.46%, from 1560 proteins).

Diversity of MAGs involved in carbon cycling

Differences in the number of genes encoding key polysaccharide degradation enzymes (i.e., glycoside hydrolase enzymes) were observed between the MAGs. For instance, MAGs from Urbana encode significantly more glycoside hydrolases (GH) compared to Havana MAGs (unpaired *t*-test, *P*-value < 0.05, see also Table S7). In addition, MAGs from Urbana showed almost double the number of cellulase genes encoding oligosaccharide-degrading enzymes and amylolitic enzymes compared to MAGs from Havana. Genes encoding beta-glucosidase enzymes GH3 (n=93) and the amylolytic enzymes GH13 (n=320) and GH15 (n=78), were among the most commonly detected glycoside hydrolases in recovered MAGs. These results were consistent with the results obtained from recovered short-reads and, in general, with a

higher soil organic matter content in the Urbana (silty loam) soil vs. its Havana (sandy) counterpart. The MAGs UD035 (*Actinobacteria*), UD029 (*Firmicutes*), and UT009 (*Acidobacteria*) from Urbana had the highest number of GH genes (n=67, 41 and 49 GH genes) and mostly corresponded to oligosaccharide-degrading and amylolytic enzymes. In Havana, MAGs HD112 and HD089 (*Acidobacteria*) and MAG HD116 (*Bacteroidetes*) had the highest number of HG genes also corresponding to cellulases, oligosaccharide-degrading and amylolytic enzymes.

Discussion

Unexpected genetic diversity in agricultural soils

The majority of the MAGs were predicted to belong to novel species, if not genera, reflecting the low representation of soil-dwelling microorganisms in current genomic databases. For instance, highly abundant archaeal and bacterial nitrifier (discussed above) and *Verrucomicrobia* populations obtained from Urbana (e.g., MAG UD002) only shared ~46% AAI to the closest reference genome. Microbial communities belonging to the this group are underrepresented in genomic databases and have been predicted to inhabit soils with high organic matter content such as those found in Urbana (20). It is important to note that while the MAGs were searched against the NCBI prokaryotic genome database (as implemented in MiGA) for close relatives, more recently sequenced genomes, which are not yet part of NCBI, would have been missed. For instance, MAG UD053 shared 61% AAI with recently described and novel phylum *Candidatus* Rokubacteri*a* (21). Further, abundant populations detected based on 16S rRNA gene fragments recovered in the metagenomes were not well represented in the recovered MAGs, such as *Gemmatimonadetes* in Urbana. Apparently, the latter genomes were not well binned, presumably due to high intra-population sequence diversity. Altogether, the MAGs reported here represent mostly novel and deep-branching taxa and offer a genomic reference for future studies targeting abundant natural microbial communities found in agricultural soils.

Recent findings have revealed that PCR-based surveys targeting N-cycle genes have

overlooked a vast amount of natural diversity related to nitrification (22-25) and denitrification genes such as *nirK* (16)*, nosZ* (17), and *nrfA* (26). These findings suggest that the previously unaccounted gene diversity might play an important role in key biogeochemical cycles. Our results show that the use of metagenomic approaches in combination with reliable detection tools (e.g., ROCker) can circumvent these limitations in samples of high sequence complexity. For instance, abundant *nirK* genes found in the soil samples were assigned to *Thaumarchaeota*, which has been inadvertently excluded in previous PCR-based gene surveys. Interestingly, the changes in relative abundance for *Thaumarchaeota nirK* gene fragments are consistent with recent findings that have proposed an alternative role for this archaeal NirK activity as part of the ammonia oxidation to nitrite mechanism in *Thaumarchaeota* (27).

Genes and microbial populations involved in biomass degradation

We explored the impact of the microbial communities in the breakdown and recycling of plant biomass in soils, by surveying genes associated with biomass and polysaccharide degradation. The two agricultural sites share a similar history of cropping where biomass derived from either corn or soybean represents a constant input of C at the end of the growing season and this was reflected by stable abundances in all GH categories studied. Even though a higher influence of plants (e.g., root-exudates) during crop-growing periods was expected (e.g., June and September), our core collecting regime was directed to sample in between plant rows, and thus, likely missed microorganisms in close proximity to roots. Overall, Urbana (silt-loam soil) showed a higher relative abundance of GH genes at both gene and genomic population levels compared to Havana, likely explained by the intrinsic characteristics of the soils. For instance, the differences in sorption and binding capacities particular to each soil type resulted in a higher OM availability in Urbana compared to Havana, which likely accounted for the differences in GH genes between the two sites. Further, previous reports have recognized that genes encoding GHs belonging to the family GH13 are among the most widespread and abundant amylolytic enzymes

208 found in microbial genomes (28) and soils (29), consistent with the findings based on the recovered MAGs reported here. Other abundant GHs in the recovered MAGs belonged to the
210 glucoamylase GH15 family, which in combination with debranching enzymes from GH13 have been proposed as part of the main enzymes for degradation of polysaccharides in bacteria (28).
212 Therefore, in addition to playing a role in the cycling of N in soils, MAGs encoding GH might also participate in maintaining and recycling labile carbon in the explored agricultural soils.

References

1. **Oh S**, **Caro-Quintero A**, **Tsementzi D**, **DeLeon-Rodriguez N**, **Luo C**, **Poretsky R**, **Konstantinidis KT**. 2011. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl Environ Microbiol **77**:6000–6011.

2. **Kopylova E**, **Noé L**, **Touzet H**. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28**:3211–3217.

3. **Wang Q**, **Garrity GM**, **Tiedje JM**, **Cole JR**. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol **73**:5261–5267.

4. **Caporaso JG**, **Kuczynski J**, **Stombaugh J**, **Bittinger K**, **Bushman FD**, **Costello EK**, **Fierer N**, **Peña AG**, **Goodrich JK**, **Gordon JI**, **Huttley GA**, **Kelley ST**, **Knights D**, **Koenig JE**, **Ley RE**, **Lozupone CA**, **McDonald D**, **Muegge BD**, **Pirrung M**, **Reeder J**, **Sevinsky JR**, **Turnbaugh PJ**, **Walters WA**, **Widmann J**, **Yatsunenko T**, **Zaneveld J**, **Knight R**. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods **7**:335–336.

5. **Yin Y**, **Mao X**, **Yang J**, **Chen X**, **Mao F**, **Xu Y**. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res **40**:W445–51.

6. **Lombard V**, **Golaconda Ramulu H**, **Drula E**, **Coutinho PM**, **Henrissat B**. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res **42**:D490–5.

7. **Allgaier M**, **Reddy A**, **Park JI**, **Ivanova N**, **D'haeseleer P**, **Lowry S**, **Sapra R**, **Hazen TC**, **Simmons BA**, **VanderGheynst JS**, **Hugenholtz P**. 2010. Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. PLoS ONE **5**:e8812.

8. **Sievers F**, **Wilm A**, **Dineen D**, **Gibson TJ**, **Karplus K**, **Li W**, **Lopez R**, **McWilliam H**, **Remmert M**, **Söding J**, **Thompson JD**, **Higgins DG**. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol **7**:539–539.

9. **Stamatakis A**. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690.

10. **Rho M**, **Tang H**, **Ye Y**. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res **38**:e191.

11. **Katoh K**, **Standley DM**. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30**:772–780.

12. **Berger SA**, **Krompass D**, **Stamatakis A**. 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. Syst Biol

60:291–302.

13. **Matsen FA**, **Hoffman NG**, **Gallagher A**, **Stamatakis A**. 2012. A format for phylogenetic placements. PLoS ONE **7**:e31009.

14. **Rodriguez-R LM**, **Konstantinidis KT**. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Prepr **4**:e1900v1.

15. **Letunic I**, **Bork P**. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res **39**:gkr201–W478.

16. **Wei W**, **Isobe K**, **Nishizawa T**, **Zhu L**, **Shiratori Y**, **Ohte N**, **Koba K**, **Otsuka S**, **Senoo K**. 2015. Higher diversity and abundance of denitrifying microorganisms in environments than considered previously. ISME J **9**:1–12.

17. **Orellana LH**, **Rodriguez-R LM**, **Higgins S**, **Chee-Sanford JC**, **Sanford RA**, **Ritalahti KM**, **Löffler FE**, **Konstantinidis KT**. 2014. Detecting nitrous oxide reductase (NosZ) genes in soil metagenomes: method development and implications for the nitrogen cycle. mBio **5**:e01193–14.

18. **Konstantinidis KT**. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA **102**:2567–2572.

19. **Konstantinidis KT**, **Tiedje JM**. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Curr Opin Microbiol **10**:504–509.

20. **Brewer TE**, **Handley KM**, **Carini P**, **Gilbert JA**, **Fierer N**. 2016. Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. Nat Microbiol **2**:16198.

21. **Hug LA**, **Thomas BC**, **Sharon I**, **Brown CT**, **Sharma R**, **Hettich RL**, **Wilkins MJ**, **Williams KH**, **Singh A**, **Banfield JF**. 2015. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environ Microbiol **18**:159–173.

22. **Daims H**, **Lebedeva EV**, **Pjevac P**, **Han P**, **Herbold C**, **Albertsen M**, **Jehmlich N**, **Palatinszky M**, **Vierheilig J**, **Bulaev A**, **Kirkegaard RH**, **Bergen von M**, **Rattei T**, **Bendinger B**, **Nielsen PH**, **Wagner M**. 2015. Complete nitrification by Nitrospira bacteria. Nature **528**:504–509.

23. **van Kessel MAHJ**, **Speth DR**, **Albertsen M**, **Nielsen PH**, **Op den Camp HJM**, **Kartal B**, **Jetten MSM**, **Lücker S**. 2015. Complete nitrification by a single microorganism. Nature **528**:555–559.

24. **Palomo A**, **Fowler SJ**, **Gülay A**, **Rasmussen S**, **Sicheritz-Ponten T**, **Smets BF**. 2016. Metagenomic analysis of rapid gravity sand filter microbial communities suggests novel physiology of Nitrospira spp. ISME J **10**:2569–2581.

25. **Pinto AJ**, **Marcus DN**, **Ijaz UZ**, **Bautista-de Lose Santos QM**, **Dick GJ**, **Raskin L**. 2016. Metagenomic evidence for the presence of comammox Nitrospira-like bacteria in a Drinking Water System. mSphere **1**:e00054–15.

26. **Nelson MB**, **Martiny AC**, **Martiny JBH**. 2016. Global biogeography of microbial nitrogen-cycling traits in soil. Proc Natl Acad Sci USA **113**:8033–8040.

27. **Kozlowski JA**, **Stieglmeier M**, **Schleper C**, **Klotz MG**, **Stein LY**. 2016. Pathways and key intermediates required for obligate aerobic ammonia-dependent chemolithotrophy in bacteria and Thaumarchaeota. ISME J **10**:1836–1845.

28. **Henrissat B**, **Deleury E**, **Coutinho PM**. 2002. Glycogen metabolism loss: a common marker of parasitic behaviour in bacteria? Trends Genet **18**:437–440.

29. **Howe A**, **Yang F**, **Williams RJ**, **Meyer F**, **Hofmockel KS**. 2016. Identification of the core set of carbon-associated genes in a bioenergy grassland soil. PLoS ONE **11**:e0166578.

Supplementary Figure Legends

Supplementary Figure 1. Alpha diversity values determined for metagenomic samples. **A.** Diversity of metagenomic reads as determined by Nonpareil. The Chao-Shen entropy values for **B.** the order level of taxonomy and **C.** functional annotations (SEED subsystems).

Supplementary Figure 2. A. Distributions of coefficients of variation for all SEED subsystems detected in soil metagenomes for all seasons. Panel **B** summarizes the distributions of coefficient of variation for all SEED subsystems (left) and the subset devoted to secondary metabolism (right) for the three cores obtained for the 20-30 cm soil samples during June in Havana and Urbana. **C.** Distributions of coefficients of variation for all SEED subsystems (left) and a subset consisting of secondary metabolism annotations (right panel) in Lake Lanier metagenomes.

Supplementary Figure 3. Functional clustering and taxonomy for sandy (Havana) and silt-loam (Urbana) soils. A. Non-metric multidimensional scaling analysis based on SEED subsystems annotation of short-reads of the metagenomic samples showed independent clustering by site and depth. The length of the arrow is proportional to the correlation between measured metadata and determined ordination values. **B** Summary of the taxonomic affiliation (figure key) of the recovered 16S rRNA gene fragments obtained from soil metagenomes.

Supplementary Figure 4. Differential abundance of SEED subsystems between top (0-5cm) and deep (20-30 cm) soil layers. Predicted-protein sequences from short-reads were annotated using UniProt and subsequently classified into functional categories using SEED subsystems. Significant differences in abundance of SEED subsystems between top and deep layers were identified using a negative binomial test as implemented in DESeq2. Selected SEED subsystems showing $\log_2$-fold change >=1 or <=-1 and adjusted *P*-values < 0.01 are shown.

Supplementary Figure 5. Abundance of N-cycle genes in sandy (Havana) and silt-loam (Urbana) soils. Heatmaps show calculated relative abundance for N-cycle genes as genome

equivalents for Havana (left panel) and Urbana (right panel). Manually-curated databases for each N gene were searched against soil metagenomes using BLASTx and outputs were filtered using ROCker models for each gene (see Methods for more details). Values for the 20-30 cm layer in June represent the average of the three soil cores.

Supplementary Figure 6. Abundance and diversity for *hao* and *nxrA* in Havana. Phylogenetic reconstruction of Hao (**A**) and NxrA (**B**) protein sequences including assembled sequences from soil metagenomes. For reconstructed sequences, names in parentheses indicate corresponding metagenomic bins. The pie charts represent the placing of Havana metagenomic reads for archaeal and bacterial *amoA* genes using RAxML EPA. Pie chart radii represent the read abundance for each node (calculated as genome equivalents) and the colors of the slices represent the depth and month for the origin of the metagenomic reads.

Supplementary Figure 7. Abundance and diversity for archaeal and bacterial *amoA, hao,* and *nxrA* in Urbana. Phylogenetic reconstruction of archaeal (**A**) and bacterial (**B**) AmoA, Hao (**C**) and NxrA (**D**) protein sequences including assembled sequences from soil metagenomes. For reconstructed sequences, names inside parentheses indicate corresponding metagenomic bins. The pie charts represent the placement of Havana metagenomic reads for archaeal and bacterial *amoA* genes using RAxML EPA. Pie chart radii represent the read abundance for each node (calculated as genome equivalents) and the colors of the slices represent the depth and month for the origin of the metagenomic reads.

Supplementary Figure 8. Changes in abundance of metagenomic populations. Log_2 fold changes in abundance (y-axis) between months (x-axis) were calculated using individual bin abundances.

Supplementary Figure 9. Relative abundances of categories for glycoside hydrolases in both agricultural soils. Glycoside hydrolases (GH) gene fragments were detected in each metagenome and individual GH abundances were summarized in six functional categories.

Supplementary Tables

Supplementary Table 1. Soil metadata for Havana and Urbana samples.

Supplementary Table 2. Agricultural management for Havana and Urbana sites during 2012.

Supplementary Table 3. Metagenomic sequences and Nonpareil estimations for Havana and Urbana sites.

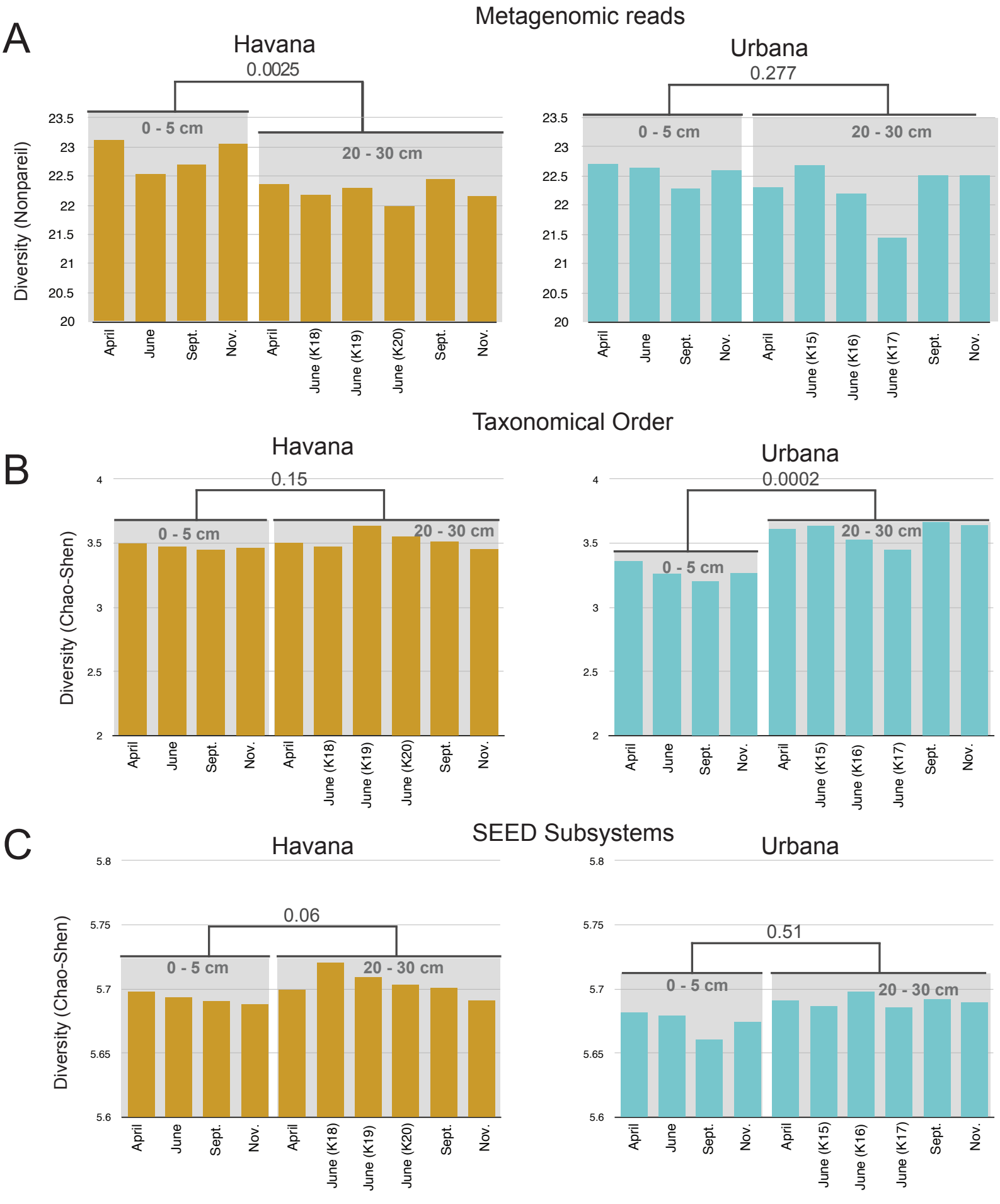
Supplementary Table 4. Summary for co-assemblies from Havana and Urbana.

Supplementary Table 5. Summary of ROCker models used for detecting N genes in metagenomic soil samples.

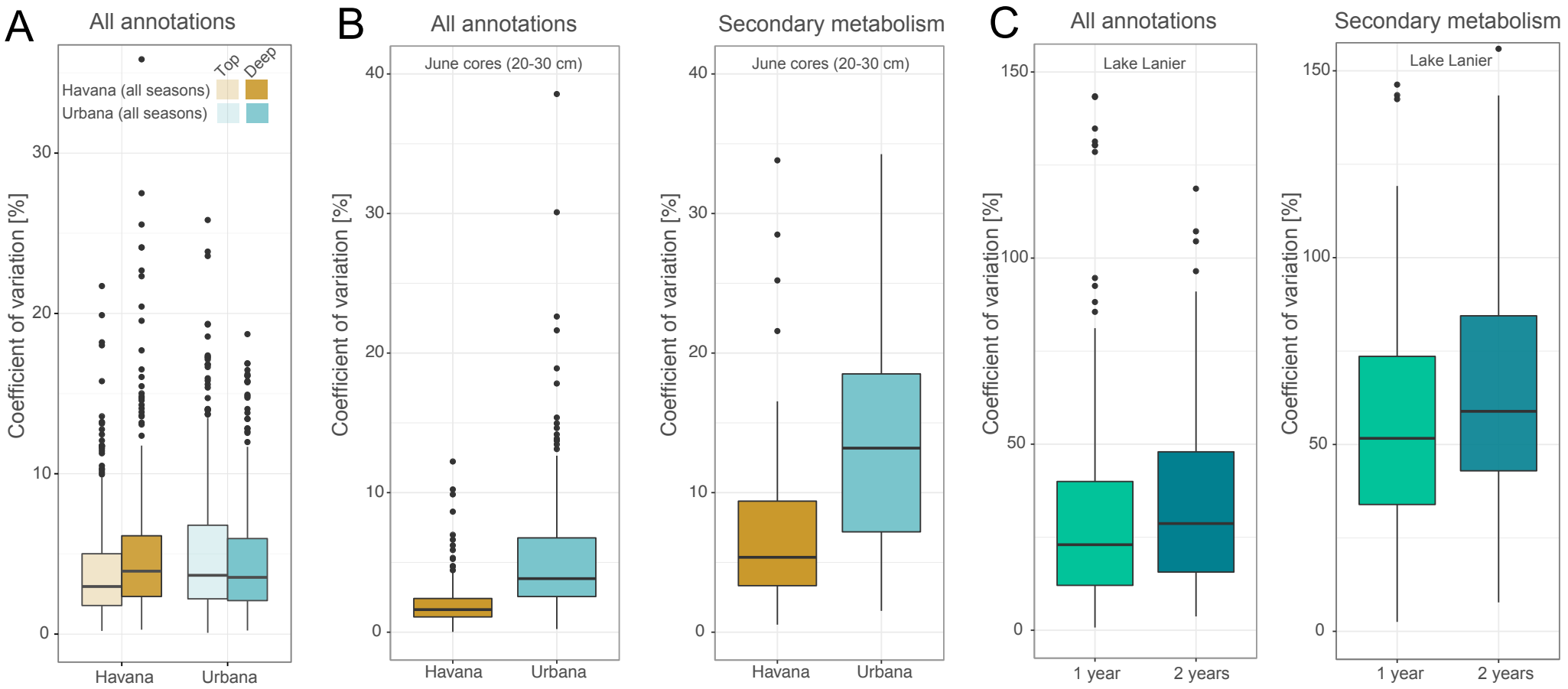
Supplementary Table 6. Summary for obtained bins from Havana and Urbana.

Supplementary Table 7. Summary of Glycoside hydrolase enzymes found in metagenomic bins

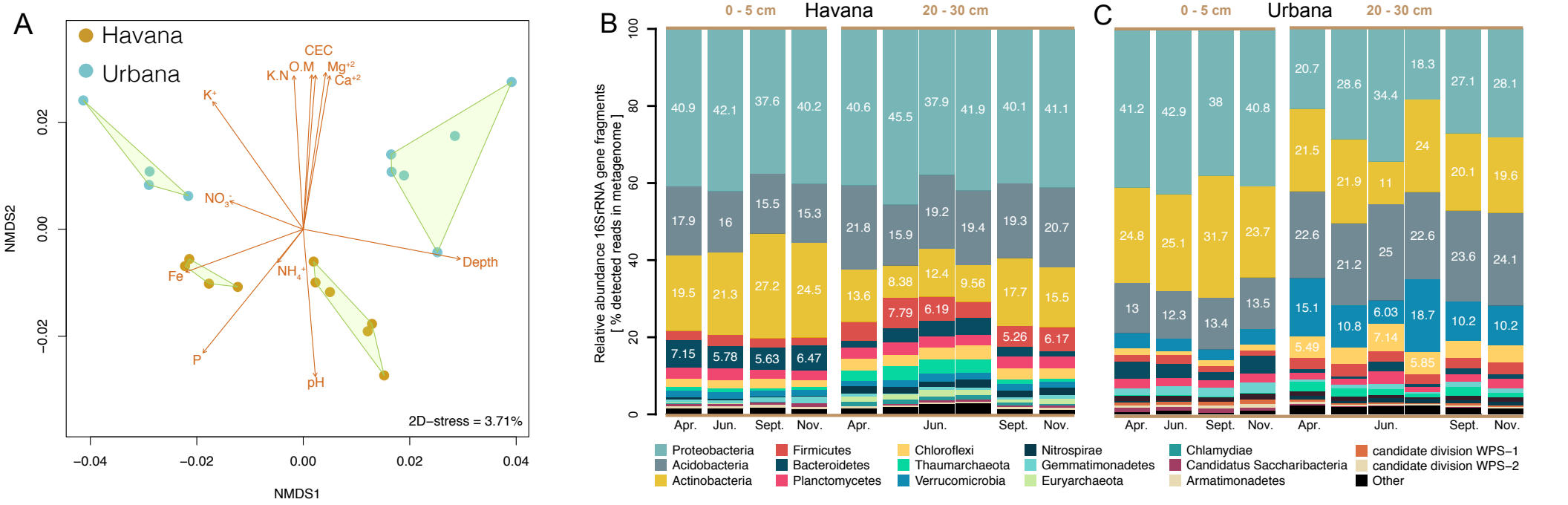
Supplementary Figure 1



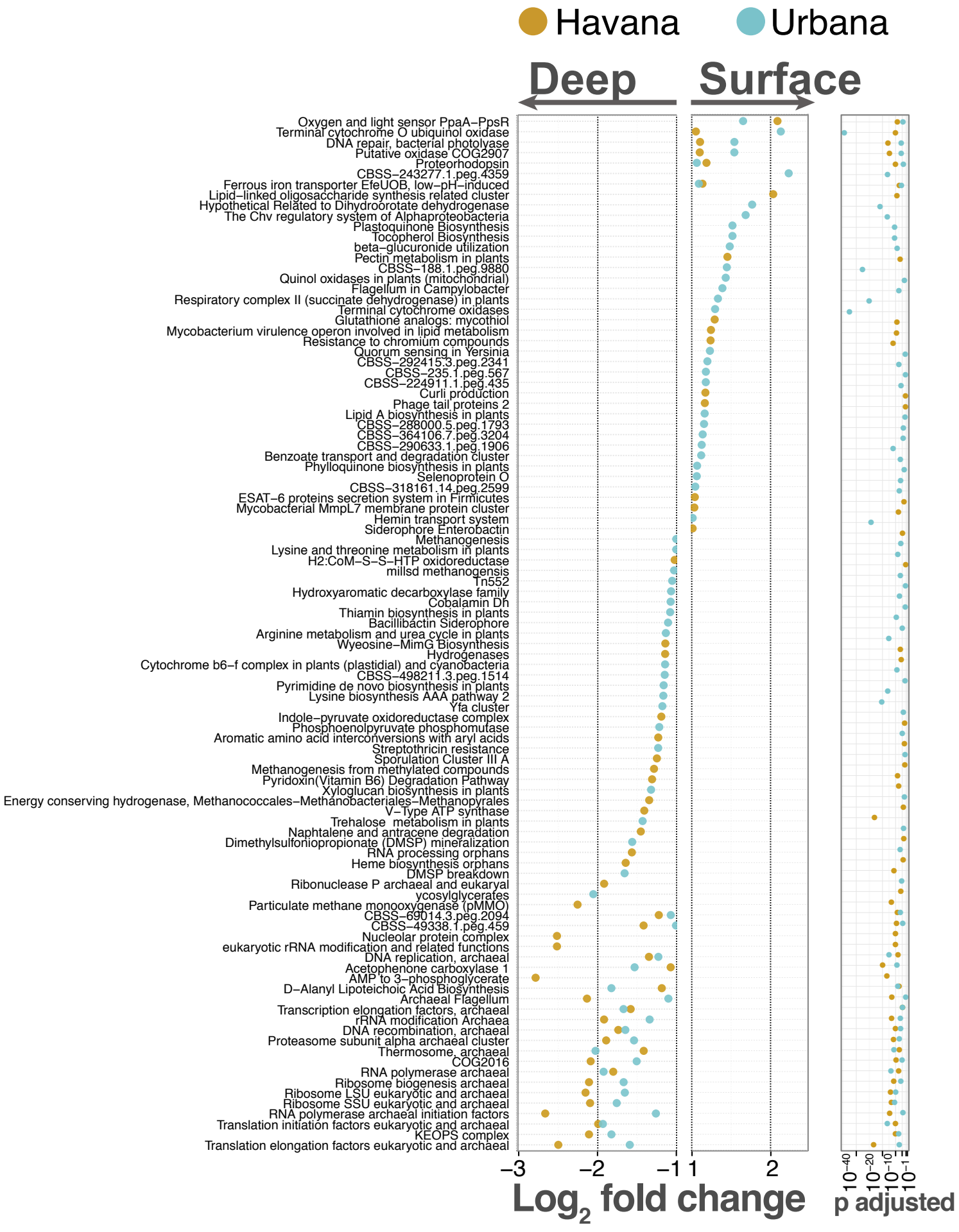
Supplementary Figure 2



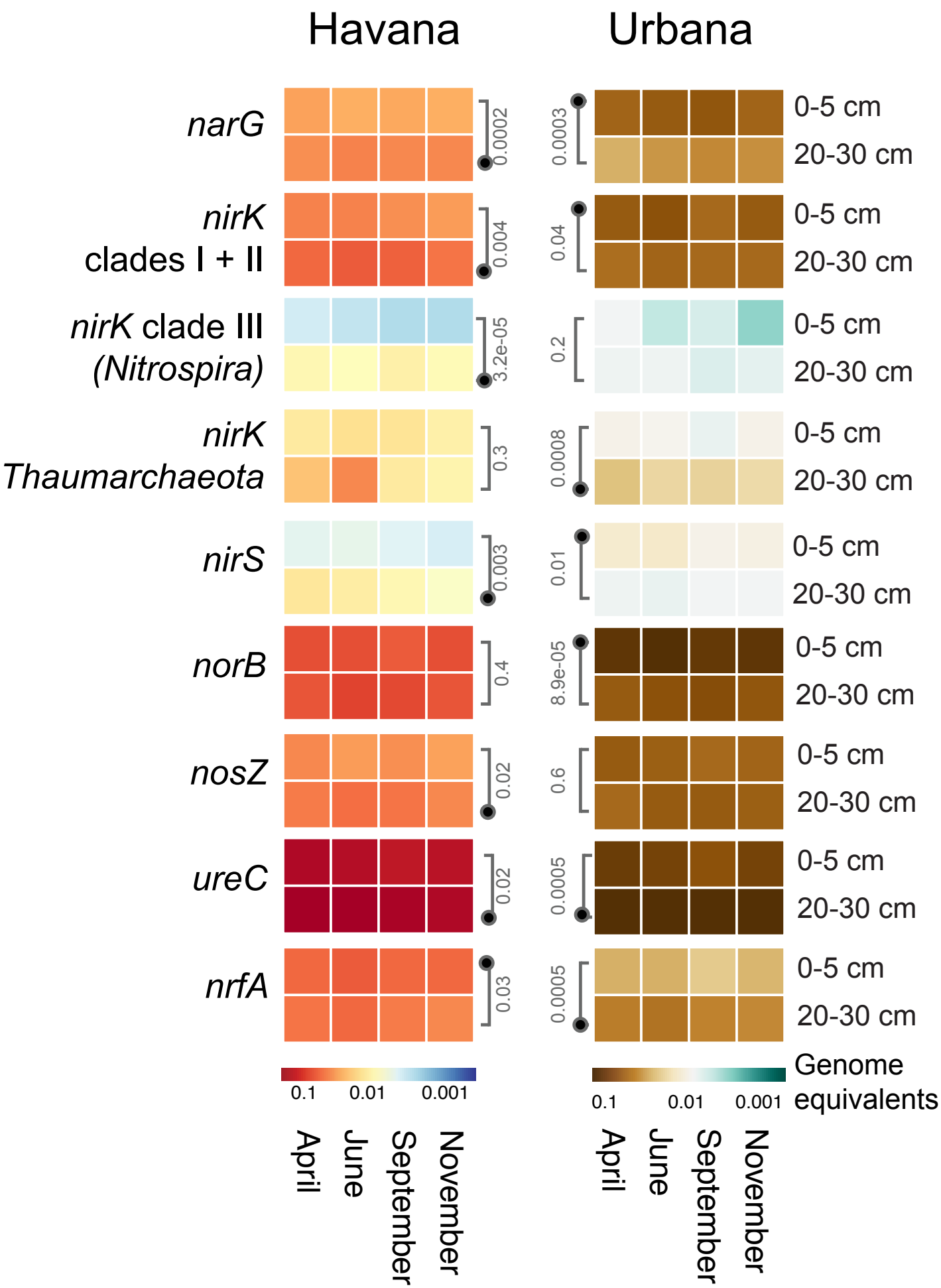
Supplementary Figure 3



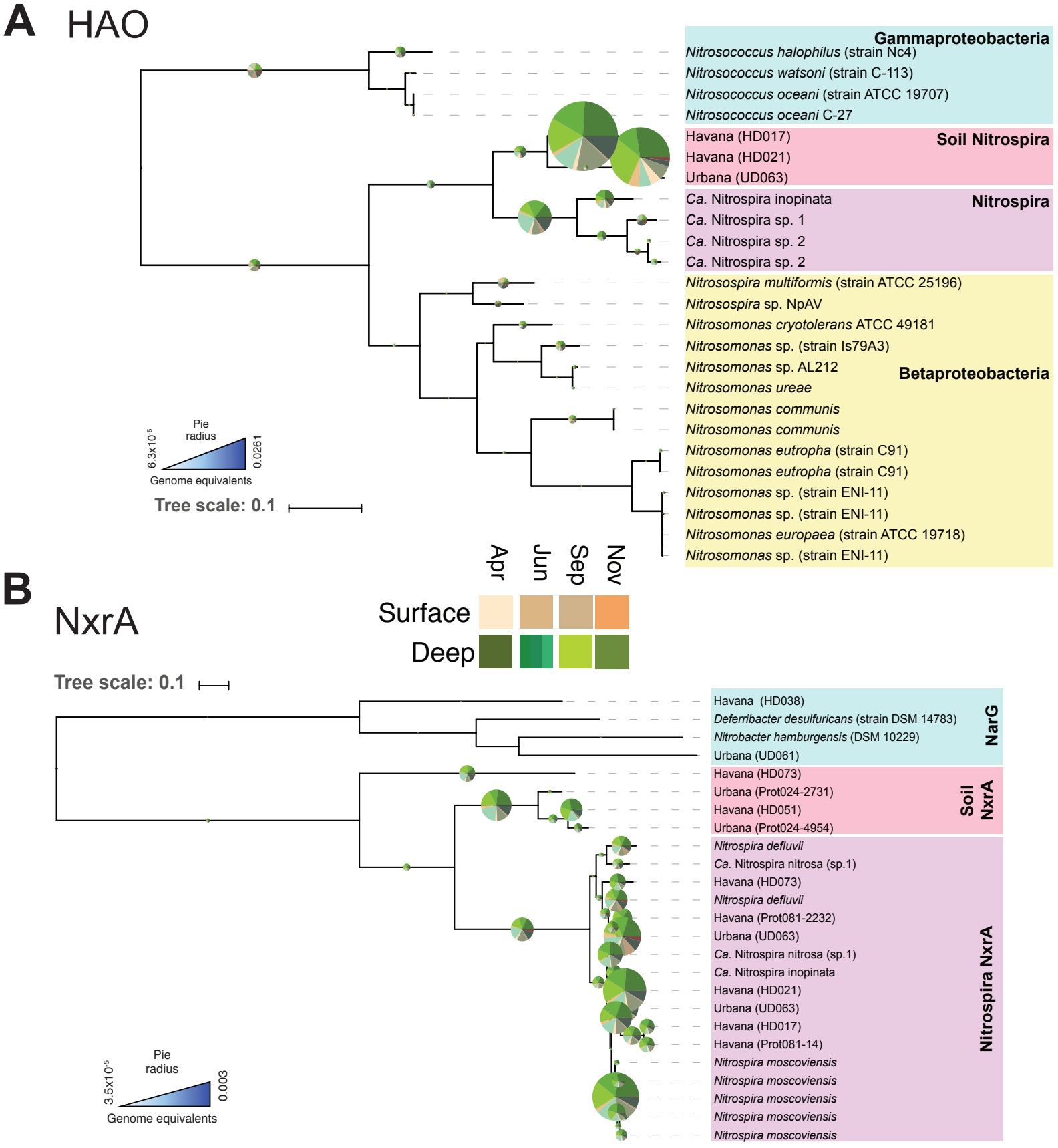
Supplementary Figure 4



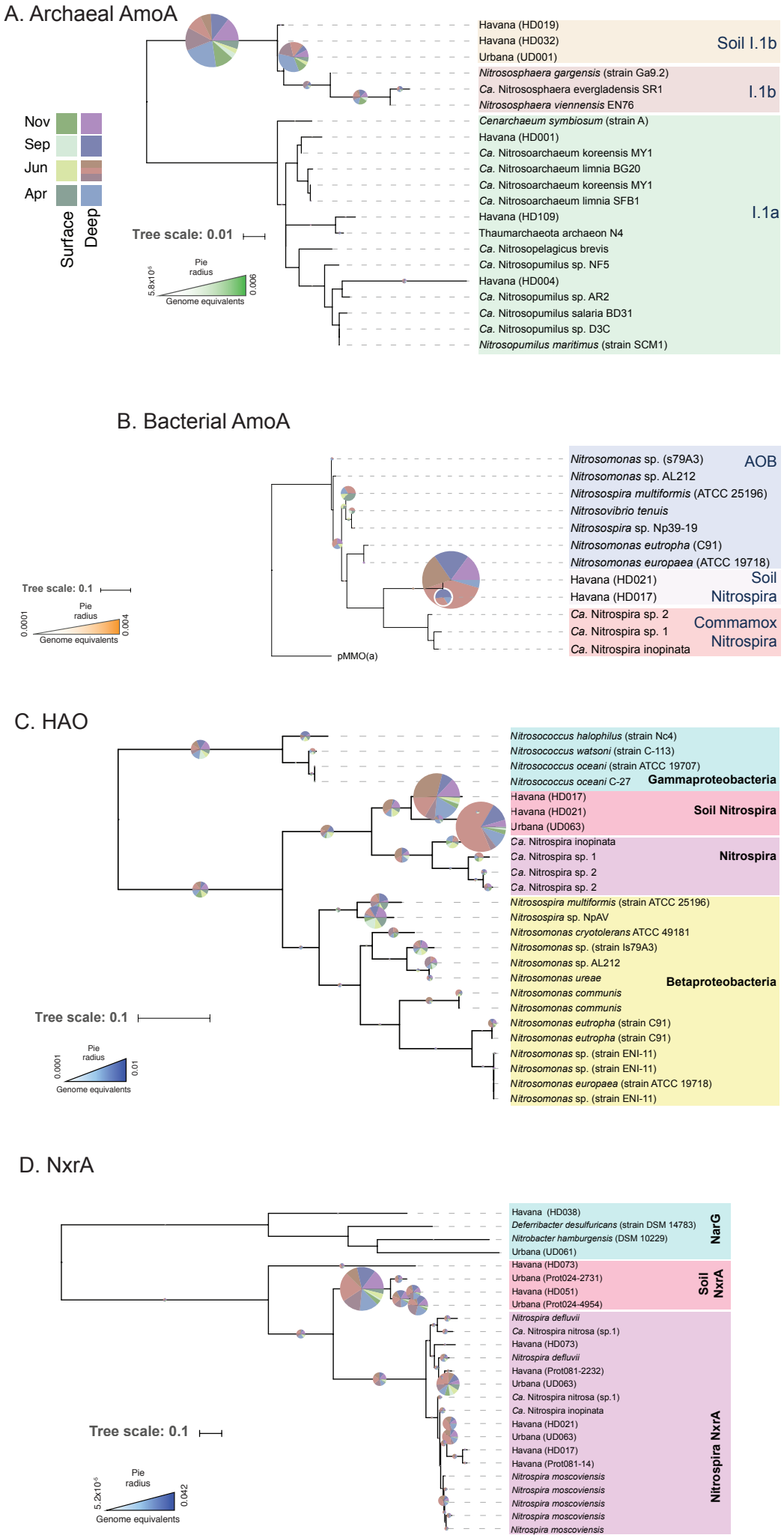
Supplementary Figure 5



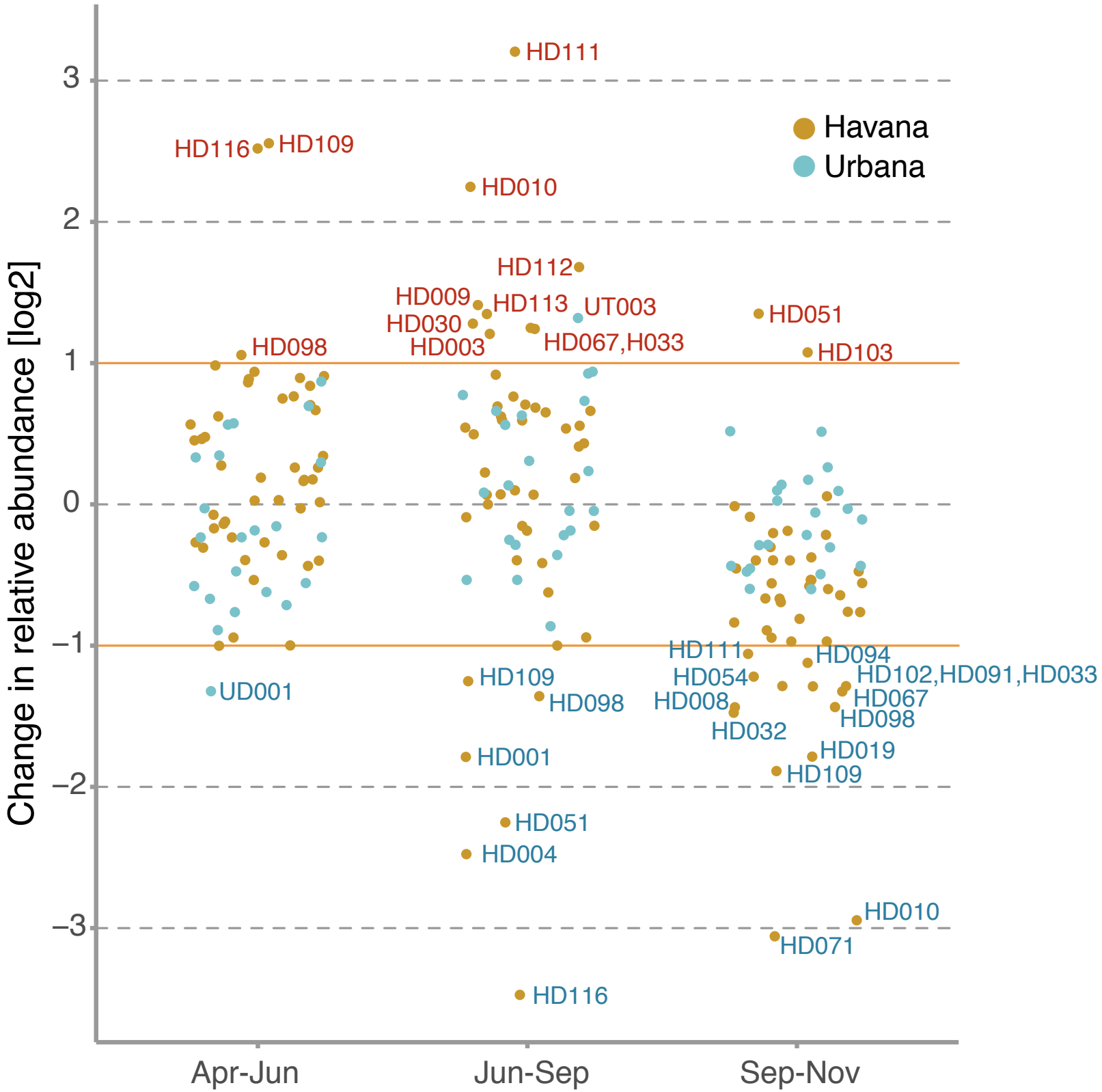
Supplementary Figure 6



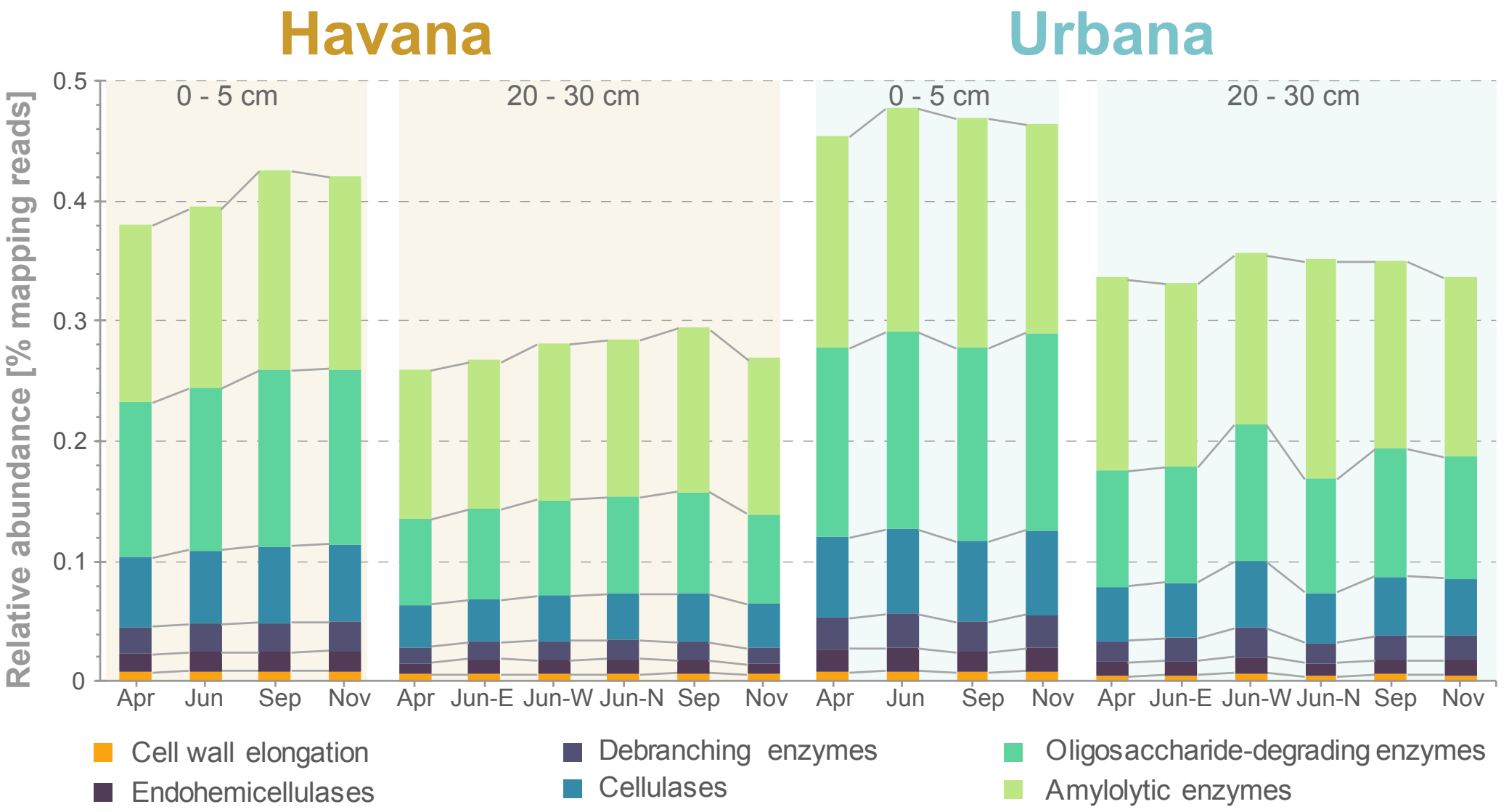
Supplementary Figure 7



Supplementary Figure 8



Supplementary Figure 9



Supplementary Table 1

Site	ID	Depth	Sampling date during 2012	Soil metadata												
				pH	Total organic matter	Available P	K	Mg	Ca	NO_3^--N	NH_4^+-N	Total Kjeldahl N	Extractable Fe	CEC	Temp	Moisture
		[cm]			[%]	[ppm-P]	[ppm]	[ppm]	[ppm]	[ppm]	[ppm]	[%]	[ppm]	[meq/100g]	[°C]	[%]
Havana (sandy soil)	K10	0 - 5	Apr 4	7.7	0.7	49	59	118	729	4	5	0.039	160	4.8	17.9	4.63
	K14		Jun 6	7.7	1.3	55	91	126	838	74	139	0.083	138	5.5	32.6	4.99
	K6		Sep 5	7.3	1.2	53	68	144	851	6	2	0.064	150	5.6	23.1	6.33
	K2		Nov 6	7.6	0.9	54	78	139	816	8	5	0.048	142	5.4	8.4	4.65
	K12	20 - 30	Apr 4	7.4	0.4	43	41	60	443	1	2	0.02	132	2.8	17.4	4.93
	K18 (E)		Jun 6	7.35	0.6	50	38	56	572	1	4	0.022	163	3.4	22.8	6.74
	K19 (M)		Jun 6	7.3	0.6	50	38	56	572	1	4	0.022	163	3.4	22.8	3.53
	K20 (W)		Jun 6	7.48	0.6	50	38	56	572	1	4	0.022	163	3.4	22.8	5.49
	K8		Sep 5	7	0.3	48	49	70	489	1	1	0.021	160	3.2	23.6	4.85
	K4		Nov 6	7.4	0.4	58	43	77	471	1	3	0.027	162	3.1	10	4.55
Urbana (silt-loam soil)	K9	0 - 5	Apr 2	5.9	3.7	46	179	355	1,800	12	4	0.158	172	18.6	22.7	19.31
	K13		Jun 4	5.3	3.5	45	202	369	1,998	26	4	0.16	161	19.6	20.4	18.65
	K5		Aug 29	5.6	4.2	54	234	425	2,412	29	3	0.167	194	21	21.3	19.33
	K1		Nov 8	6	3.7	46	187	373	2,051	6	4	0.152	182	17.4	9.1	21.82
	K11	20 - 30	Apr 2	6.2	4.1	21	122	558	3,135	4	5	0.15	138	24.2	18.2	21.6
	K15 (M)		Jun 4	5.92	3.8	18	59	456	2,550	7	4	0.14	113	19.1	19.7	20.33
	K16 (S)		Jun 4	6.92	3.8	18	59	456	2,550	7	4	0.14	113	19.1	19.7	19.71
	K17 (N)		Jun 4	6.2	3.8	18	59	456	2,550	7	4	0.14	113	19.1	19.7	22.69
	K7		Aug 29	6.1	4.1	25	102	498	2,913	4	3	0.155	138	22.6	21.8	18.78
	K3		Nov 8	6.2	3.7	20	79	449	2,669	7	4	0.127	126	20.9	7.8	20.93

Supplementary Table 2

Site	Sampling date during 2012	Crop information	Tillage & N-fertilizer input	Notes
Havana	Apr 4	Pre-tillage, pre-fertilizer, pre-planting at time of sampling (winter fallow)	Pre-Tillage, Pre-fertilizer	
	Jun 6	Corn planted May 12	Spring tillage, UAN28 applied late April (180 lb N/acre)	Herbicide applied June 15
	Sep 5	Full canopy corn; beginning senesce		
	Nov 6	Post-soybean harvest by time of sampling; harvested few days prior		
Urbana	Apr 2	Pre-planting at time of sampling (winter fallow)	Pre-Tillage	
	Jun 4	Pre-planting at time of sampling	Spring tillage No UAN28 application this crop year	Soybean planted Jun 6, 2012, glyphosate late June
	Aug 29	Full canopy soybean		Full growing season rain-fed only
	Nov 8	Post-harvest; Soybean harvested Nov 1	No Fall tillage yet	

Supplementary Table 3

Site	ID	Depth	Month	Sequences		Coverage
				Trimmed Reads*	Trimmed reads length (avg)	
Havana	K2	0 - 5 cm	November	27,808,182	123.8	0.117
	K4	20 - 30 cm	November	24,373,825	123.7	0.192
	K6	0 - 5 cm	September	32,787,009	124.3	0.116
	K8	20 - 30 cm	September	33,047,556	124.7	0.178
	K10	0 - 5 cm	April	38,337,187	124.5	0.105
	K12	20 - 30 cm	April	29,017,415	124.5	0.172
	K14	0 - 5 cm	June	53,543,681	126.6	0.294
	K18	20 - 30 cm (E)	June	30,610,876	129.2	0.226
	K19	20 - 30 cm (M)	June	31,784,017	129.1	0.203
	K20	20 - 30 cm (W)	June	49,463,716	126.8	0.427
Urbana	K1	0 - 5 cm	November	21,681,291	124.5	0.101
	K3	20 - 30 cm	November	26,427,577	123.9	0.155
	K5	0 - 5 cm	September	28,018,675	123.7	0.188
	K7	20 - 30 cm	September	26,864,164	124.3	0.159
	K9	0 - 5 cm	April	32,535,582	124.6	0.127
	K11	20 - 30 cm	April	30,652,664	124.3	0.215
	K13	0 - 5 cm	June	34,023,870	124.4	0.162
	K15	20 - 30 cm (M)	June	29,187,308	127.0	0.237
	K16	20 - 30 cm (S)	June	72,914,672	126.9	0.492
	K17	20 - 30 cm (N)	June	32,057,255	126.0	0.466

Supplementary Table 4

Samples	Depth	Million reads	IDBA co-assembly				Gene Prediction	
			Contigs	N50	Avg. length	Longest contig	Total bp	Genes
Havana top	0-5 cm	136,453,108	118,687	1,130	1,160.5	46,851	137,742,067	220,365
Havana deep	20-30 cm	179,133,698	419,023	1,779	1,568.9	388,680	657,447,015	938,759
Urbana top	0-5 cm	104,056,954	147,610	1,349	1,308.9	60,203	193,216,907	301,988
Urbana deep	20-30 cm	195,425,606	430,724	1,524	1,409.7	78,105	607,223,845	883,376

Supplementary Table 5

Target Protein	ROCker build (125 bp read length)				
	Positive references	Negative references	Sensitivity	Specificity	Accuracy
AmoA bacteria	7	14	92.60%	99.64%	98.10%
AmoA archaea	5	16	100%	100%	100%
Hao	22	9	98.72%	99.89%	99.14%
NarG/NxrA	311	0	99.68%	99.98%	99.96%
NirK (Clade I and II)	140	8	96.79%	99.99%	99.95%
NirK (Thaumarchaeota)	18	0	98.15%	100%	100.00%
NirK (Clade III)	10	0	96.38%	100%	100.00%
NirS	74	33	97.83%	100.00%	99.97%
NorB	309	0	98.99%	99.97%	99.94%
NosZ	166	0	98.55%	99.99%	99.96%
UreC	103	0	99.42%	99.99%	99.98%
NrfA	260	8	98.11%	99.97%	99.94%
RpoB	756	0	99.71%	99.34%	99.62%