Title: Functional metabolic annotations in the soil virosphere are rare but enriched for carbohydrate-active enzymes (CAZymes) and chitin decomposition functions

Dominik Merges*
Independent researcher

*Correspondence: mergesd01@gmail.com

Running title: Soil viral CAZyme and chitinase annotations

KEYWORDS

soil viromes; auxiliary metabolic genes; CAZymes; chitinase; metagenomics; functional potential; carbon cycling

ABSTRACT

Soil viruses are increasingly recognized as components of microbial communities that may alter how communities may function, yet the frequency and functional distribution of virus-encoded metabolic genes in soils remain poorly understood. Here, the Global Soil Virus Atlas gene catalog, comprising 1,432,147 viral genes from 1,223 soil samples across 13 ecosystem types, was analyzed to quantify the distribution of virus-encoded functional annotations and to estimate their representation in matched total metagenomic inventories. Functional annotations were assessed across KEGG Orthology, Pfam, and CAZy and grouped into carbon cycling, nitrogen cycling, and antibiotic resistance-associated categories. Only 1,903 viral genes (0.13%) had functional annotations. Carbon cycling dominated the annotated repertoire (1,840 genes; 96.5%), whereas nitrogen cycling (33 genes; 1.7%) and antibiotic resistance-associated functions (30 genes; 1.6%) were rare. Within carbon-cycling annotations, chitinase-associated genes were the most frequent named category (628 genes), followed by hemicellulase-associated functions. To estimate the representation of viral genes within broader metagenomic functional inventories, viral and total metagenomic annotations were compared across six exactly matched JGI studies. Viral contributions were usually low, with four of six studies showing less than 1% viral representation in targeted functions, but reached 9.86% for chitinase in one study. Together, these results show that virus-encoded metabolic genes are globally sparse in soils but are non-randomly concentrated in carbon-active CAZyme-linked functions, indicating that analyses restricted to microbial genes may underestimate predicted functional potential for selected degradation traits.

IMPORTANCE

Soil metagenomic studies usually interpret functional potential from microbial genes alone, even though soil viruses can also encode metabolic functions. In a global soil viral gene catalog, functionally annotated viral genes were rare overall, but the detectable signal was strongly concentrated in carbon-active CAZyme-linked functions, especially chitinase-associated annotations. Because viral diversity and activity can decouple from microbial responses to environmental gradients, virus-encoded functional genes may disproportionately affect gene-centric estimates of selected functions under stress or seasonal constraint (Zheng et al., 2022; Merges et al., 2023). A targeted comparison of viral and total metagenomic functional annotations further showed that viral contributions were usually small but could become non-trivial for selected carbon-degradation traits. These results identify viral CAZyme-associated functions as the clearest current case in which ignoring viruses may bias gene-centric estimates of predicted functional potential in soils. More broadly, the study provides a quantitative baseline for evaluating when virus-encoded genes are likely to matter for environmental metagenomic interpretation.

INTRODUCTION

Soil viruses are abundant members of terrestrial microbial communities and can influence ecosystem-relevant processes through host mortality, host metabolic reprogramming, and lateral gene flow. Global resources such as the Global Soil Virus Atlas now make it possible to examine viral genomic potential across broad spatial and ecosystem gradients (Graham et al., 2024). One implication of this expanding soil virosphere is that metagenomic estimates of community functional potential may be incomplete when they are derived from microbial genes alone.

Microbial and viral communities can respond to environmental change in distinct ways, creating conditions where virus-encoded functions may not track host-encoded functional inventories. In organochlorine pesticide-contaminated soils, bacterial diversity declines whereas viral diversity and the abundance of virus-encoded auxiliary metabolic genes (AMGs) linked to pesticide degradation increase

with contamination severity (Zheng et al., 2022). Metatranscriptomic analyses across an elevational gradient similarly indicate that bacterial metabolic activity declines with elevation while bacteriophage activity persists (Merges et al., 2023). Cohesive viral responses to soil moisture and habitat context have also been reported, suggesting that viral community dynamics can be structured by environmental drivers that do not map one-to-one onto microbial functional profiles (Santos-Medellín et al., 2022; Durham et al., 2022). Together, these observations motivate an inference risk: functional inventories derived from microbial genes alone may underestimate total gene-centric potential for selected traits and may overattribute those inventories to microbes when viral signals respond differently to environmental drivers.

This issue intersects with biodiversity-ecosystem functioning research, where links between diversity and process rates are increasingly recognized as context dependent in microbial systems. In contrast to plant communities, microbial communities often contain many taxa with overlapping core traits, and realized process rates can be buffered by regulation, resource supply, and species interactions. Recent synthesis has therefore argued that functional redundancy should be treated cautiously and only in relation to explicit traits, spatial scales, and environmental context, rather than as a generic explanation for weak diversity-function relationships (Louca et al., 2018; Eisenhauer et al., 2023a; Eisenhauer et al., 2023b).

Auxiliary metabolic genes provide one route by which viruses may alter gene-centric interpretations of microbial functional potential. AMGs are generally interpreted as host-derived genes that may modulate host metabolism during infection in ways that increase host fitness. Marine and permafrost studies have identified AMGs involved in carbon and nutrient transformations, and soil-focused studies now show that bacterial and viral genomic potential can differ across environmental contexts and management regimes (Roux et al., 2016; Emerson et al., 2018; Touceda-Suarez et al., 2025). Recent carbon-focused work in compost-amended soils has further highlighted phage-associated carbon metabolism genes as a current area of interest in soil viromics (Yuan et al., 2025). At the same time, AMG interpretation remains sensitive to annotation quality and curation, and gene presence alone does not establish expression, enzyme activity, or process-level effects (Pratama et al., 2021; Martin et al., 2025).

This study used the Global Soil Virus Atlas gene catalog to address three questions: First, how frequently do functionally annotated virus-encoded metabolic genes occur across global soils? Second, are these annotations concentrated in particular functional domains? Third, how large can the virus-encoded component become relative to total metagenomic functional annotations for selected functions in exactly matched datasets? These questions are addressed at the level of predicted functional potential and do not infer gene expression, enzyme activity, or ecosystem flux from gene presence alone.

RESULTS

Functionally annotated viral metabolic genes are rare across the global soil gene catalog.

Across 1,432,147 viral genes from 1,223 soils, only 1,903 genes (0.13%) had functional metabolic annotations supported by at least one of KEGG, Pfam, or CAZy. Among annotated genes, carbon-cycle annotations dominated (1,840 genes; 96.5%; Figure 1b, Table 1), whereas nitrogen-cycle annotations (33 genes; 1.7%, Fig. 1c, Table 1) and antibiotic resistance-associated annotations (30 genes; 1.6%; Fig.1d) were sparse. Most annotations were supported by a single database (66.0%), with 23.7% supported by two databases and 10.3% supported by all three (Table 1).

Carbon-active CAZyme-linked functions dominate the annotated repertoire.

Within carbon-cycle annotations, chitinase-associated annotations were the most frequent single activity (628 genes; 33.0% of all annotated genes; Fig, 1b). Other carbon-related annotations included

hemicellulase-associated genes (236; 12.4%), cellulase/hemicellulase-associated genes (93; 4.9%), pectinase-associated genes (72; 3.8%), and cellulase-associated genes (22; 1.2%), with additional heterogeneous carbon-related annotations (789; 41.5%; Fig, 1b). When restricted to CAZy-annotated genes (1,079), chitinase-associated genes accounted for 58.2% (628/1,079) and hemicellulase-associated genes for 21.9% (236/1,079). Together, these results indicate that the detectable viral metabolic signal in soils is strongly concentrated in carbon-active functions and, within that subset, in chitinase-associated annotations.

Carbon-active viral annotations occur across major soil ecosystem categories.

Carbon-related viral annotations were detected across all major ecosystem categories represented in the dataset, with the largest gene counts in unclassified environments (777 carbon genes), agricultural soils (426), and forest soils (309; Figure 1a). Nitrogen-related viral annotations were most frequent in agricultural soils (15 genes), followed by grasslands (6 genes; Fig. 1a). Antibiotic resistance-associated viral annotations were rare across all ecosystems, with forests showing the highest diversity of antibiotic mechanisms (7 genes; Fig. 1a). The ecosystem distribution therefore reinforces the broader pattern that functionally annotated viral genes are dominated by carbon-active categories rather than by nitrogen cycling or resistance-associated functions.

Viral genes represent a bounded but context-dependent non-trivial fraction of targeted metagenomic functional inventories.

To estimate the representation of virus-encoded functions within broader metagenomic inventories, viral-contig-specific annotations were compared with total metagenomic annotations across six matched JGI studies (Fig. 2a&b; Table 2; Supplementary table 1), targeting chitinase, nifH, glutamine metabolism, and two antibiotic resistance-associated functions. Viral contributions to KEGG-annotated targeted functions ranged from 0.10% to 9.86% (mean 2.60%), with four of six studies showing less than 1% contribution (Fig. 2a). The largest value occurred for chitinase in one sample (9.86%; 7 viral genes out of 71 total KEGG-annotated chitinase genes; Fig. 2a&b). Across studies, viral gene counts remained low across a wide range of total gene counts, indicating that viral functional gene counts did not scale proportionally with total metagenomic functional gene counts in this small matched set. These comparisons therefore provide a bounded estimate of when viral genes may measurably alter gene-centric inventories for selected functions.

DISCUSSION

The main contribution of this study is to place a quantitative bound on virus-encoded metabolic potential in global soils. Across the full gene catalog, functionally annotated viral genes were exceptionally rare, which argues against a broad interpretation in which viral genes dominate soil metagenomic functional inventories. At the same time, the annotated signal was strongly structured rather than random, with carbon-active functions accounting for nearly all detectable metabolic annotations and chitinase-associated genes representing the largest named category. In that sense, rarity and ecological clustering are both central features of the current dataset.

Placed in the context of microbial biodiversity-ecosystem functioning research, these results support a trait-specific rather than a general interpretation of viral effects. In microbial systems, links between community composition and process rates are often weaker and more context dependent than in macroscopic systems, and recent synthesis has cautioned against invoking functional redundancy as a generic explanation detached from explicit traits and environmental conditions (Louca et al., 2018; Eisenhauer et al., 2023a; Eisenhauer et al., 2023b). These results fit that perspective. They do not suggest that viruses broadly reshape microbial functional potential across all metabolic domains.

Rather, they indicate that any bias introduced by ignoring viruses is likely to be function specific and strongest for selected degradation traits.

A key implication of recent soil viromics is that viral diversity and activity can decouple from microbial responses, implying that virus-encoded functional potential may become relatively more prominent under environmental stress or seasonal constraint. Organochlorine-contaminated soils show decreases in bacterial diversity but increases in viral diversity and AMG abundance linked to pesticide degradation (Zheng et al., 2022). Across an elevational gradient, bacterial transcriptional activity declines while bacteriophage activity remains detectable (Merges et al., 2023). Soil viral communities can also show cohesive responses to moisture and habitat context (Santos-Medellín et al., 2022; Durham et al., 2022), and virus-host interactions extend through soil depth profiles (Muscatt et al., 2023). In this context, the concentration of annotated viral functions in CAZyme-linked carbon categories suggests a tractable pathway by which viruses could alter gene-centric interpretations in a function-specific manner, even when functionally annotated viral metabolic genes are rare in aggregate.

The concentration of annotated viral functions in carbon-active CAZyme-linked categories is consistent with current understanding of soil viral ecology. Previous work has argued that soil viruses are plausible participants in carbon processing and has identified viral glycoside hydrolases in soil and permafrost systems (Trubl et al., 2018; Emerson et al., 2018). Recent soil metagenomic studies have also shown that bacterial and viral genomic potential can differ across environmental contexts, including traits related to carbon and nitrogen transformations (Touceda-Suarez et al., 2025). In parallel, recent work in compost-amended soils has highlighted carbon-focused phage AMGs as potential functional reservoir (Yuan et al., 2025). The present analysis complements those site-specific studies by providing a global catalog-based estimate and by showing that the strongest detectable signal in soils is concentrated in CAZyme-linked carbon functions. At the same time, the scope of inference remains limited to predicted functional potential, because annotation alone does not establish transcription, translation, enzyme activity, or ecosystem flux effects (Pratama et al., 2021; Martin et al., 2025).

The six-study matched metagenome comparison further narrows interpretation. Viral contributions to targeted functions were usually small, but the elevated contribution observed for chitinase in one matched sample indicates that virus-encoded genes can become a measurable component of selected functional inventories. This supports a function-specific view of soil viral AMGs: viruses are unlikely to broadly reshape estimates of total functional potential across all metabolic categories, but they may alter predicted inventories for selected degradation traits, particularly CAZyme-mediated carbon turnover. The most informative implication of the present study is therefore not that soil viral AMGs generally control ecosystem functioning, but that viral CAZyme-associated functions represent the clearest current target for testing when virus-encoded traits modify estimates of microbial functional potential in soils.

MATERIALS AND METHODS

Global analysis of virus-encoded functional annotations

The Global Soil Virus Atlas published by Graham et al. 2024 was analyzed, focusing on the gene catalog associated with soil viral contigs. The underlying atlas compiles viral sequences from a global collection of soil metagenomes and reports 1,432,147 viral genes associated with the curated soil virosphere dataset. In the present study, the viral gene annotation tables derived from that resource were used to characterize the functional landscape of soil viral communities across 1,223 soil samples representing 13 ecosystem types.

Functional annotation and category assignment

Functional annotation coverage was assessed across three annotation systems: KEGG Orthology, Pfam protein families, and CAZy enzyme classifications. To distinguish broad annotation support from more weakly supported assignments, each viral gene was classified by annotation confidence according to the number of databases supporting the same functional interpretation: none, 0 databases; low, 1 database; medium, 2 databases; and high, 3 databases. Carbon-cycling functions were grouped by activity using CAZy family assignments together with associated KEGG and Pfam annotations. Categories included chitinase, cellulase, hemicellulase, pectinase, starch degradation, oxidative enzymes, and mixed cellulase/hemicellulase assignments where family-level interpretation spanned both substrate classes. Nitrogen-cycling functions were classified into mechanistic categories including nitrogen fixation, denitrification, nitrogen assimilation, and nitrogen transport using targeted keyword matching of KEGG and Pfam descriptions. Antibiotic resistance-associated functions were grouped into mechanistic classes including beta-lactam resistance, aminoglycoside resistance, fluoroquinolone resistance, glycopeptide resistance, multidrug efflux, and penicillin-binding proteins.

Comparative analysis of viral and total metagenomic functional annotations

To estimate how strongly virus-encoded functions could affect gene-centric functional inventories, a targeted comparison was performed between viral annotations and total metagenomic annotations in matched soil datasets retrieved from the JGI Integrated Microbial Genomes and Microbiomes platform on 10 January 2026. Graham et al. provide the study identifiers linking the viral atlas entries to their source metagenomes, which enabled exact matching between viral datasets and the corresponding total metagenomes. Six matched studies (JGI study IDs 3300025711, 3300025925, 3300038415, 3300038418, 3300005471, and 3300036781; Table S1) were selected for a bounded proof-of-concept analysis. The six-study comparison was designed as a targeted pilot rather than a comprehensive survey of all matched metagenomes, and included two studies each representing carbon-active CAZyme-linked functions (agricultural rhizosphere), nitrogen-related functions (peat soils), and antibiotic resistance-associated functions (agricultural rhizosphere and forest soils).

Targeted functions and statistics

For each matched study, total metagenomic KEGG and Pfam annotations were extracted from the IMG annotation framework and compared with the corresponding viral-contig-specific annotations. The targeted annotation pairs were chitinase activity, nitrogen fixation, glutamine metabolism, and two antibiotic resistance-associated functions, using the KO and Pfam identifiers reported in the dataset. Viral contribution was calculated for each functional category as the proportion of viral gene counts relative to the total number of annotated genes detected in the corresponding metagenome for that category. All data integration and analysis were carried out in Python using pandas-based workflows to merge annotation tables from heterogeneous file formats, including tab-delimited output, HMMER-derived annotation files, and gene feature tables. Figures were generated in Python using matplotlib and seaborn.

REFERENCES

Cardinale, B.J., Duffy, J.E., Gonzalez, A., et al. (2012). Biodiversity loss and its impact on humanity. Nature 486, 59-67. doi:10.1038/nature11148.

Delgado-Baquerizo, M., Maestre, F.T., Reich, P.B., et al. (2016). Microbial diversity drives multifunctionality in terrestrial ecosystems. Nature Communications 7, 10541. doi:10.1038/ncomms10541.

Durham, D.M., Sieradzki, E.T., ter Horst, A.M., et al. (2022). Substantial differences in soil viral community composition within and among four Northern California habitats. ISME Communications 2, 100. doi:10.1038/s43705-022-00171-y.

Eisenhauer, N., Hines, J., Maestre, F.T., Rillig, M.C. (2023a). Reconsidering functional redundancy in biodiversity research. npj Biodiversity 2, 9. doi:10.1038/s44185-023-00015-5.

Eisenhauer, N., Hines, J., Rillig, M.C. (2023b). Reply: Functional similarity is more appropriate than functional redundancy. npj Biodiversity 2, 24. doi:10.1038/s44185-023-00030-6.

Emerson, J.B., Roux, S., Brum, J.R., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. Nature Microbiology 3, 870-880. doi:10.1038/s41564-018-0190-y.

Graham, E.B., Camargo, A.P., Wu, R., et al. (2024). A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. Nature Microbiology 9, 1873-1883. doi:10.1038/s41564-024-01686-x.

Hector, A., Bagchi, R. (2007). Biodiversity and ecosystem multifunctionality. Nature 448, 188-190. doi:10.1038/nature05947.

Hooper, D.U., Chapin, F.S. III, Ewel, J.J., et al. (2005). Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. Ecological Monographs 75, 3-35. doi:10.1890/04-0922.

Louca, S., Polz, M.F., Mazel, F., et al. (2018). Function and functional redundancy in microbial systems. Nature Ecology and Evolution 2, 936-943. doi:10.1038/s41559-018-0519-1.

Martin, C., Emerson, J.B., Roux, S., Anantharaman, K. (2025). A call for caution in the biological interpretation of viral auxiliary metabolic genes. Nature Microbiology 10, 2122-2129. doi:10.1038/s41564-025-02095-4.

Merges, D., Schmidt, A., Schmitt, I., et al. (2023). Metatranscriptomics reveals contrasting effects of elevation on the activity of bacteria and bacterial viruses in soil. Molecular Ecology 32, 6552-6563. doi:10.1111/mec.16756.

Muscatt, G., Cook, R., Millard, A., et al. (2023). Viral metagenomics reveals diverse virus-host interactions throughout the soil depth profile. mBio 14, e02246-23. doi:10.1128/mbio.02246-23.

Pratama, A.A., Bolduc, B., Zayed, A.A., et al. (2021). Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. PeerJ 9, e11447. doi:10.7717/peerj.11447.

Roux, S., Brum, J.R., Dutilh, B.E., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537, 689-693. doi:10.1038/nature19366.

Santos-Medellín, C., Estera-Molina, K., Yuan, M., et al. (2022). Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. Proc Natl Acad Sci U S A 119, e2209132119. doi:10.1073/pnas.2209132119.

Touceda-Suarez, M., Ponsero, A.J., Barberan, A. (2025). Differences in the genomic potential of soil bacterial and viral communities between urban greenspaces and natural arid soils. Applied and Environmental Microbiology 91, e02124-24. doi:10.1128/aem.02124-24.

Trubl, G., Jang, H.B., Roux, S., et al. (2018). Soil viruses are underexplored players in ecosystem carbon processing. mSystems 3, e00076-18. doi:10.1128/mSystems.00076-18.

Yuan, S., Wu, Y., Balcazar, J.L., Wang, D., Zhu, D., Ye, M., Sun, M., Hu, F. (2025). Expanding the potential soil carbon sink: unraveling carbon sequestration accessory genes in vermicompost phages. Applied and Environmental Microbiology 91, e00296-25. doi:10.1128/aem.00296-25.

Zheng, X., Jahn, M.T., Sun, M., et al. (2022). Organochlorine contamination enriches virus-encoded metabolism and pesticide degradation associated auxiliary genes in soil microbiomes. The ISME Journal 16, 1397-1408. doi:10.1038/s41396-022-01188-w.

280 **TABLES**

281 Table 1. Global soil viral functional landscape overview.

Metric	Value	Notes
Total viral genes analysed	1,432,147	GSV atlas viral gene catalogue
Total soil samples	1,223	samples represented in the catalogue
Ecosystem types represented	13	ecosystem categories
Functionally annotated viral genes	1,903	0.13% of all viral genes
Carbon cycling genes	1,840	96.5% of annotated genes
Nitrogen cycling genes	33	1.7% of annotated genes
Antibiotic resistance-associated genes	30	1.6% of annotated genes
High-confidence annotations	197	3 databases; 10.3% of annotated genes
Medium-confidence annotations	451	2 databases; 23.7% of annotated genes
Low-confidence annotations	1,255	1 database; 66.0% of annotated genes
KEGG annotated genes	397	0.028% of all viral genes
Pfam annotated genes	1,272	0.089% of all viral genes
CAZy annotated genes	1,079	0.075% of all viral genes

282

283

284 Table 2. Function-specific representation of viral genes in matched total metagenomic annotations
285 across six JGI studies.

Function	JGI Study ID	Ecosystem	Target IDs (KO/Pfam)	Total KEGG	Viral KEGG	Viral %	Study #
Chitinase	3300025925	Agricultural	K03791/PF00182	95	2	2.11	S3
Chitinase	3300025711	Agricultural	K03791/PF00182	71	7	9.86	S4
Glutamine metabolism	3300038415	Peat	K01915/PF00120	366	1	0.27	S5
Nitrogen fixation (nifH)	3300038418	Peat	K02588/PF00142	34	1	2.94	S6
Beta-lactam-related	3300005471	Agricultural	K00662/PF02522	322	1	0.31	S1
Efflux pumps	3300036781	Forest	K03587/PF03717	973	1	0.10	S2

286 *Viral % is calculated as (viral KEGG count / total KEGG count) × 100 for the targeted KO in each study.*

287

FIGURES

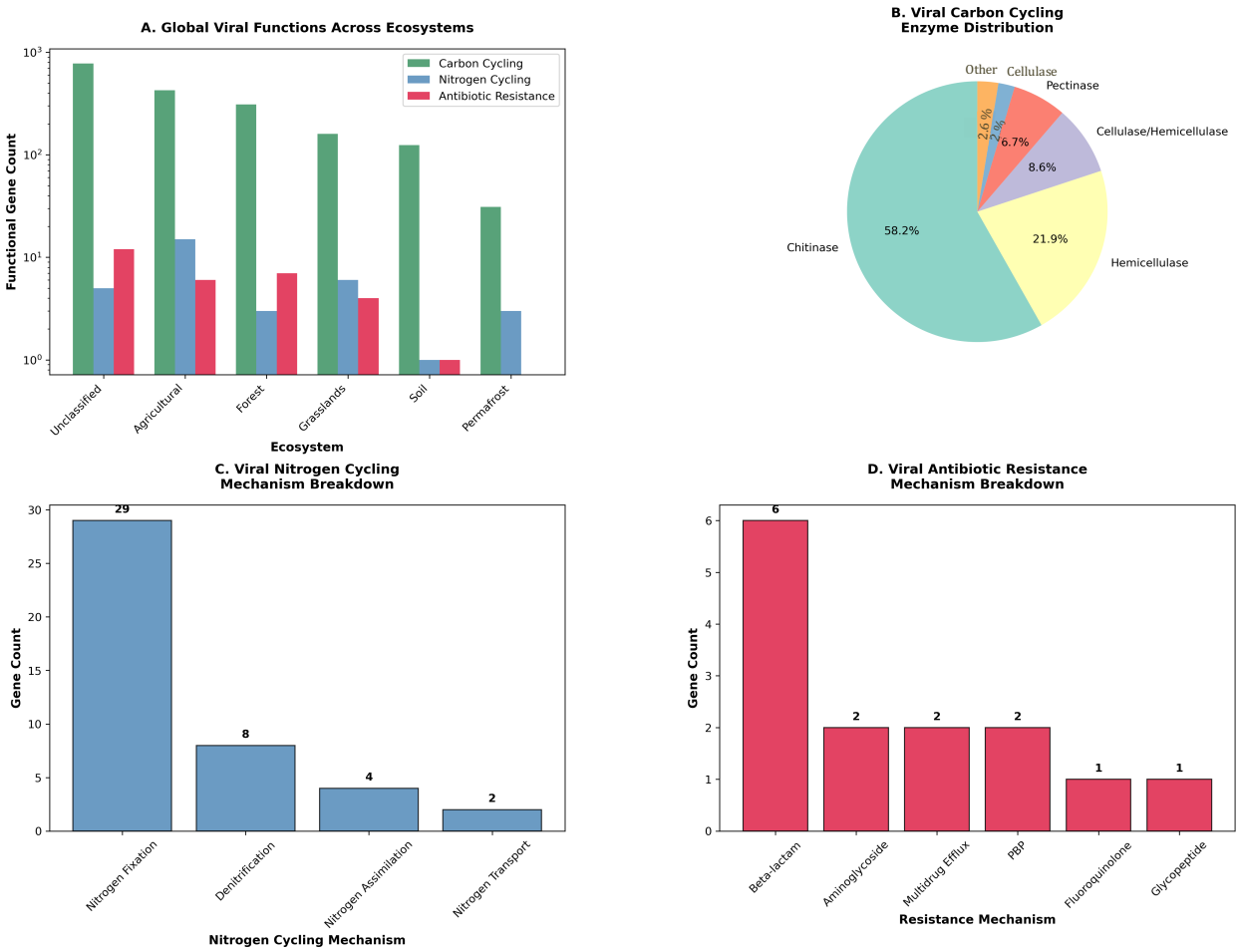

Figure 1. Global functional categories across ecosystems and within-category breakdowns. Panel A: counts of carbon-, nitrogen-, and antibiotic resistance-associated viral annotations across selected ecosystem categories (log scale). Panel B: distribution of CAZy-annotated carbon-cycle activities, highlighting chitinase-associated and hemicellulase-associated annotations. Panel C: nitrogen-cycle functional subcategories. Panel D: antibiotic resistance-associated mechanisms.

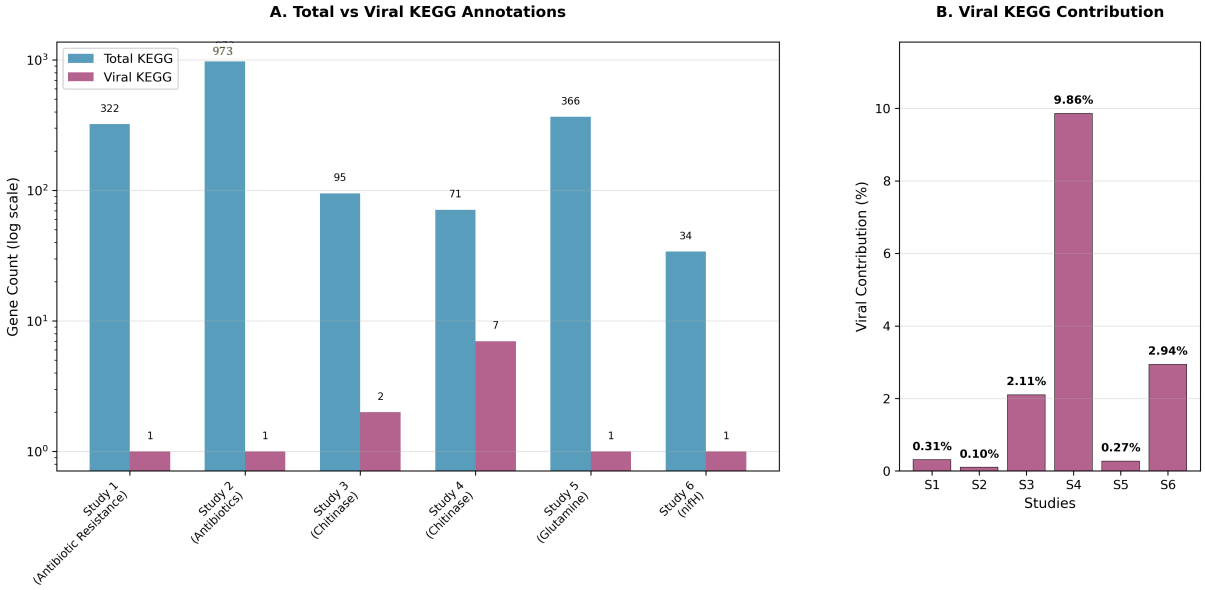

Figure 2. Viral contribution to targeted metagenomic functional annotations in six matched studies. Panels A and B: viral and total KEGG-annotated gene counts for targeted categories and viral contribution percentages.

336 **SUPPLEMENTAL TABLES**

337 Table S1. Metadata for the six matched JGI studies used in the viral-versus-total comparison.

JGI Study ID	Ecosystem	Sample name
3300025711	Agricultural	Switchgrass rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA - KBS S5-4 metaG (SPAdes) (version 2)
3300025925	Agricultural	Corn rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA - C1
3300036781	Forest	Soil fungal communities from native Pine forests in California, United States - CAS_2
3300005471	Agricultural	Corn, switchgrass and miscanthus rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA - KBS K1-50-2 metaG
3300038415	Peat	Peat soil microbial communities from Marcell Experimental Forest, MN, USA - P7_D10
3300038418	Peat	Peat soil microbial communities from Marcell Experimental Forest, MN, USA - P4_D8

338