Phage defence system abundances vary across environments and increase with viral density

Sean Meaden[1,2,*], Edze Westra[3,†], Peter Fineran[4,5,6,7,†]

1. Biology Department, University of York, York YO10 5DD, UK
2. York Biomedical Research Institute (YBRI), University of York, York YO10 5NG, UK
3. Environment and Sustainability Institute, Biosciences, University of Exeter, Penryn TR10 9FE, UK
4. Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand
5. Genetics Otago, University of Otago, Dunedin, New Zealand
6. Bioprotection Aotearoa, University of Otago, Dunedin, New Zealand
7. Maurice Wilkins Centre for Molecular Biodiscovery, University of Otago, Dunedin, New Zealand

* corresponding author
† equal contribution

Abstract

The defence systems bacteria use to protect themselves from their viruses are mechanistically and genetically diverse. Yet the ecological conditions that predict when defences are selected for remain unclear, as substantial variation in defence prevalence has been reported. Experimental work in simple communities suggests ecological factors can determine when specific defence systems are most beneficial, but applying these findings to complex communities has been challenging. Here, we use a comprehensive and environmentally balanced collection of metagenomes to survey the defence landscape across complex microbial communities. We also assess the association between the viral community and the prevalence of defence systems. We identify strong environmental effects in predicting overall defence abundance, with animal-host-associated environments and hot environments harbouring more defences overall. We also find a positive correlation between the density and diversity of viruses in the community and the abundance of defence systems. This study provides insights into the ecological factors that influence the composition and distribution of bacterial defence systems in complex microbial environments and outlines future directions for the study of defence system ecology.

Introduction

The immense diversity and complexity of the global virome has resulted in a multitude of ways that bacteria can defend themselves against viral infections. These defences span a range of mechanisms, including restriction modification (RM) and CRISPR systems, which recognise and cleave infecting viral genomes, while abortive infection systems often trigger degradation of essential molecules of the host cell, preventing further spread of viral particles [1]. In recent years, many more types of defence systems have been identified, expanding the known 'defensome' [2] and this diversity of defences and virus-encoded counter defences [3] suggest ongoing co-evolution and ecological significance. Many of these defence systems frequently co-occur within the same genome [4] with some combinations providing additional levels of defence through both additive and synergistic interactions [5–7]. In general, defence systems exhibit broad genetic and mechanistic diversity, encompassing those that degrade invading nucleic acids, such as RM and CRISPR, those that trigger cell death or dormancy, such as abortive infection (Abi) systems like toxIN, or type-III CRISPR and CBASS, and those that exhibit numerous other mechanisms (reviewed in [8]). These diverse mechanisms may also be favoured under specific ecological scenarios (reviewed in [9]), for example systems that protect neighbouring cells may be more beneficial when environmental spatial structure is high [10]. However, linking these defence systems to environmental factors in complex environments is challenging and the ecological drivers that shape their distributions in natural environments are less well understood. Finally, defence systems are frequently carried by mobile genetic elements (MGEs), allowing rapid mobilisation into new hosts and facilitating competition between MGEs for shared hosts (reviewed in [11]).

Both theory and experiments have revealed potential drivers of the evolutionary ecology of defence systems (reviewed in [9]), but we lack a synthesis of this knowledge and many open questions remain in complex microbial communities (reviewed in [12]). Abiotic factors generally determine microbial community composition (e.g. pH or salinity, [13]) and these vary substantially across environments, as shown by the Earth Microbiome Project (EMPO)[13]. In turn, biotic factors such as virus:microbe ratios, population sizes and interaction rates also vary

across environments and these biotic factors are likely to shape the composition of defence systems present [12]. For example, high relatedness between neighbouring bacteria has been shown to determine when abortive infection is a successful strategy [10], while the benefits of CRISPR systems outweigh receptor-based resistance when microbial biodiversity is high [14], whereas defence can be selected against in the presence of beneficial antibiotic resistance encoding plasmids [15]. As defence systems can be readily gained and lost, or inactivated, from bacterial genomes [15–22], their distributions may be optimised by ecological factors.

Genomic surveys of defence systems across bacterial and archaeal genomes have revealed some of these drivers of defence system composition. Multiple studies have found strong effects of genome size on the abundance of defences [23,24]. Furthermore, by linking genomes to predicted traits, additional drivers can be inferred, such as aerobicity [25] and temperature [26] for CRISPR systems, temperature for RM systems and fast growth rates for overall defence abundance [27]. Expanding this approach to include genomes assembled from metagenomic data has shown that genomes from the gut environment carry substantially more defence systems than genomes from soils and the oceans, respectively [2], and that plant-associated bacteria have fewer defence systems than non-associated close relatives [28].

An additional factor predicted to shape the abundance and type of defence systems present in an environment is the density of viruses. Higher density would likely lead to more frequent infections and, in turn, stronger selection for defence systems. We previously identified that the abundance of CRISPR defence systems is both positively correlated with the abundance of viruses, and varies widely across different microbial environments, with host associated environments carrying more CRISPR than free-living environments [29]. Here, we collected a diverse range of metagenome assemblies from a public sequence data repository (MGnify, European Nucleotide Archive, [30]) and mined these assemblies for both defence systems and viral sequences. We then used coverage information as a proxy for the relative abundance of defence systems and viruses. We describe the distribution of defence systems across environments, variation in the total amount of defence systems and link defence abundance to the abundance, or density, of viruses in each sample.

Methods

Data curation

We collected a wide-ranging and standardised collection of metagenomes that had been processed using consistent methods. The MGnify database (European Nucleotide Archive, ENA) was accessed via the API using the MGnifyR R package (https://github.com/EBI-Metagenomics/MGnifyR) on 14/06/2024. A search was conducted for all MGnify samples labelled as 'assembly' under the 'experiment-type' field and filtered to retain those processed using the MGnify pipeline version 5.0. Metadata from the resulting 34,799 assemblies was collected with the getMetadata function from MGnifyR. Samples were then selected from categories in the 'biome_string' field that had > 99 samples per group with a requirement that only Illumina sequence data was included. The locations of the assemblies within the ENA database were located with searchFile function of MGnifyR and the resulting URLs downloaded via a curl command on the University of York's HPC server. The associated fastq data files for each assembly were downloaded using enaBrowserTools [31] and subsampled to 1 million reads per sample using seqtk [32]. Samples with fewer than 1 million reads were excluded from downstream analyses.

Metagenome community composition and diversity

From the subsampled 1 million reads from each sample, 500,000 were taxonomically profiled using Kraken2 with the Kraken2 standard database [33]. Taxonomic groups that made up less than 0.01% of the relative abundance were removed. Diversity at the Genus level was calculated and PCoA clustering performed using the 'vegan' R package [34].

Defence system search

Contigs were first annotated for coding sequences using prodigal (default settings, translation table 11) [35]. Comparison on a subset of samples found that when 'meta' mode was used the results were almost identical. Assemblies shorter than 100kbp were excluded. Contigs were searched for defence systems using PADLOC ([36], version 2.0) with the PADLOC database (version 2.0). Defences labelled as 'DMS_other', 'dXTPase', 'PDC', 'HEC' and 'VSPR' were discarded as these are

either non-defence, not experimentally verified or unpublished predicted defence systems at the time of analysis, although some HEC systems have been subsequently verified [37]. We also used DefenseFinder to identify defense systems and compared these results to our PADLOC results. We found that the number of defences identified by each tool correlated very strongly and that these correlations were consistent across environments (Fig. S1), although PADLOC recovered slightly more systems. Benchmarking between tools and for specific systems has been previously described for complete genomes [38] and we solely used the PADLOC outputs for downstream analysis. Assemblies were searched for CRISPR arrays with metaCRT using default parameters [39].

Viral sequence search

Assemblies were searched for viral sequences using the geNomad pipeline [40]. The tool uses a dual strategy approach of both marker-based and alignment free classification. Sequences (contigs in this study) are either used for gene prediction and annotation against a reference catalogue of markers or grouped based on sequence similarity using a neural network model. These results are then aggregated for each sequence to provide a classification as either plasmid, viral or chromosomal. Full details of the methods and construction of the marker protein profiles are available in Camargo et al. [39]. During post-processing, sequences shorter than 1 kb and those lacking any ‘hallmark’ (as classified by geNomad) viral genes were discarded. Sequences were classified as chromosomal, plasmid or phage based on the maximum score ascribed by geNomad. The genomic locations of defences were identified using the geNomad predictions for chromosome, phage, plasmid or prophage again using the maximum score.

Abundance estimation

The associated sequencing reads were mapped to the assemblies with bwa ([41], version 0.7.17) and processed with samtools ([42], version 1.9). The resulting alignment files (BAM format) were used to calculate coverage and read recruitment values for all contigs in each assembly using coverM with the ‘contig’ command (version 0.7.0) and the ‘metabat’ and ‘count’ methods respectively. The contigs identified from the defence system and viral sequence searches were then extracted from the coverM results tables and used for downstream analysis. These per-contig

abundance files were then merged with the PADLOC results with 1 count value recorded per defence system-type per contig. In cases where one contig carried multiple defence systems, the count value was recorded for each defence type. Notably, of all the contigs identified as carrying a defence system, >95% carried a single defence system type, likely due to the highly fragmented nature of the assemblies (Fig. S2). The total defence abundance per sample was calculated by the sum of reads mapping to contigs carrying each defence system. We opted here to focus on measuring the abundance of defence systems and therefore count multiple unique defence systems on a single contig independently, effectively 'double counting' contig read recruitment if it carried multiple unique defences. In practice, this represented <5% of all contigs due to the fragmented nature of the assemblies (4% carrying 2 systems, 0.4% carrying 3 systems and 0.5% carrying 4 or more systems). Count data was also obtained using the same mapping-based approach for those contigs predicted to be of viral origin: abundance tables were extracted from the coverM results tables based on the geNomad predictions and collated into a master file containing the results from all samples.

Sequencing effort and eukaryotic contamination controls

We assessed the effect of sequencing depth by first collecting the associated fastq data and counting the number of reads. We also assessed eukaryotic (human) DNA contamination, predicted from the Kraken2 analysis, and found higher levels of eukaryotic DNA associated with human-associated samples (Fig. S3). When analysis was restricted to samples with < 50% of classified reads being of eukaryotic origin, and with the inclusion of assembly N50 and original sequencing depth, the results were qualitatively the same. There remained a strong environmental effect on defence abundance and a consistent correlation between viral abundance and defence abundance.

Taxonomic identification of plant-derived defence systems

We observed a high number of argonaute and Tiamat systems in the plant-associated samples. To assess the origin of these defence systems we extracted the corresponding contigs, on which these defences were located, and assigned a taxonomic classification using the MMSeqs2 taxonomy pipeline against the NCBI

non-redundant (nr) nucleotide database (downloaded 25/08/2024). Contig classifications based on the last common ancestor (LCA) were visualised in R.

Statistical analysis

The effect of environment on total defence abundance was assessed using a GLM with a 'quasipoisson' error structure. A measure of assembly fragmentation (N50 value) was included in the model, as preliminary analysis found a significant relationship between N50 value and defence abundance. Significance was assessed with ANOVA by comparison against a null model with the environment term removed.

The effect of environment of defence system composition was assessed by first converting the results to a sample by defence system abundance matrix. Permutational ANOVA was then applied using the 'adonis2' function from the 'vegan' R package. This function converts the abundance matrix to Bray-Curtis distances of dissimilarity between samples and applies a permutational ANOVA. 999 permutations were used. Ordinations were conducted using the 'NMDS' and 'pco' functions in 'vegan' based on Bray-Curtis dissimilarity values. Taxonomic dissimilarity scores were extracted from the Kraken2 analysis and compared to the defence composition dissimilarity scores using Spearman's correlation.

Correlations between defence abundance and viral abundance were assessed using a linear model including N50 as a covariate and a gaussian error structure. Defence abundance and viral abundance counts were log10 transformed to improve model fit and model residuals were assessed visually. Significance was assessed by ANOVA against a null model with viral abundance removed. We repeated the above statistical tests on a restricted dataset that had the following criteria: all samples had < 50% of reads classified as eukaryotic and both the assembly N50 value and the read count of the original samples included in the model. This aimed to account for assembly fragmentation and the sequencing depth, based on the assumption that all the data were used to generate each the assemblies.

Results

Defence abundance, diversity and composition varies across environmental categories

The discovery of novel defence systems continues at pace, but our knowledge of the ultimate drivers of their evolution and ecology is lacking. Here, we collated an ecologically diverse collection of 1075 metagenomes from 12 environments. We then surveyed the defence repertoire and abundances using a homologue-search based tool [36] to assess which environments carry the most defences and the composition of defence systems in those environments. We found that the total abundance of defence carrying contigs significantly varied among environments ($F_{11,1037} = 46.0$, $p < 0.0001$), suggesting that environmental conditions strongly affect when defence systems are optimal. Notably, animal-host associated environments (urethra, gut, vagina, oral and skin) were highest in overall defence abundance, with hot environments similarly high in defence abundance (Fig. 1).

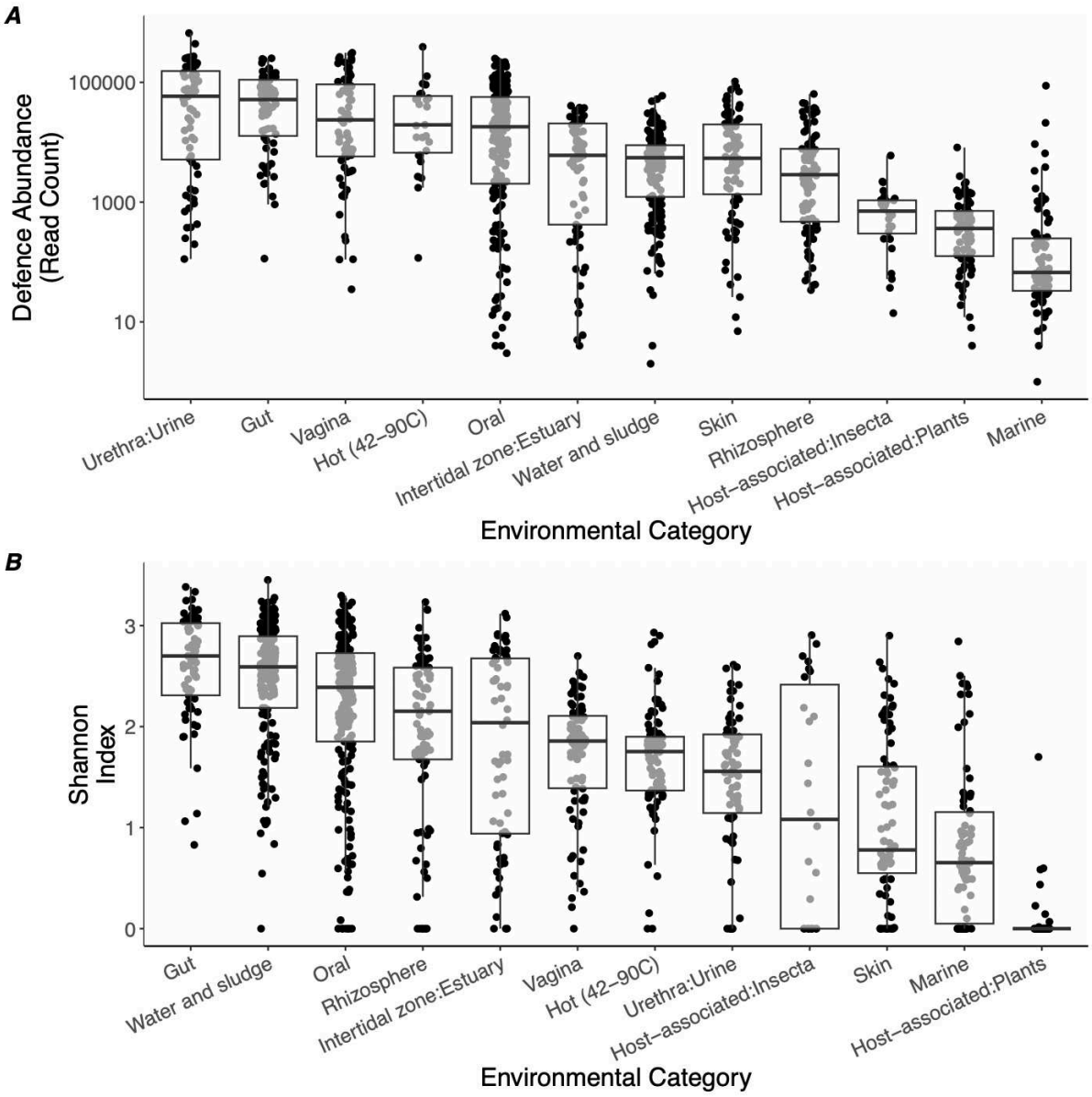


Fig. 1. Defence abundance and diversity varies across microbial environments. A) Points represent individual metagenomes grouped by biome metadata categories provided from the European Nucleotide Archive. Defence systems were identified from metagenomic contigs using PADLOC [36], a standardised subsample of 1 million reads were mapped to each assembly and the counts of contigs carrying defence systems were summed to count total defence per sample (>95% of contigs had a single defence system). Boxplots show median values and first and third quartiles. B) Boxplots showing defence diversity (Shannon Index) for each environment. Points represent individual metagenomic assemblies. Shannon's diversity index was calculated from a subsample of 1 million reads mapped to contigs carrying defence systems.

We then calculated the diversity of defences (Shannon's index) for each sample, which also varied significantly with environment ($F_{11,1064} = 75.2$, $p < 0.0001$, Fig. 1). Overall, defence diversity was correlated with defence abundance ($F_{1,987} = 719.9$, $p < 0.0001$); however, there were minor differences in the ranking of the environmental categories. Water and sludge, estuary and soil (rhizosphere) environments had similarly high defence diversity to the host-associated samples, despite lower defence abundances. We also observed substantial variation in the abundance and distribution of specific defences, with a notably sparse distribution, meaning that most defences were rare or absent in the majority of samples (Fig. S4). To assess the role of the environment in shaping the defence composition we applied a permutational ANOVA to a Bray-Curtis matrix of dissimilarity between samples using environment as a fixed effect. We found a significant effect of environment ($F_{11,1133} = 23.0$, $p < 0.001$) with an R^2 value of 0.18. Despite this significant effect, inspection of NMDS ordinations showed substantial overlap between groups (Fig. 2A), preventing accurate prediction of defence composition from environmental information alone. Importantly, taxonomic profiling of the same samples did show strong clustering by environment (Fig. 2B) and we observe a substantial mismatch between overall community diversity and defence diversity (Fig. S5 and Fig. 1B). To assess the contribution of taxonomic composition on defence composition we assessed the correlation between these pairwise distance scores for all samples. We found a significant but weak association between defence composition and taxonomic composition (Spearman's rho (n = 427,286): 0.223, $p < 0.0001$). The pan-immune hypothesis predicts that insufficient defences can be carried on a single genome for lasting resistance, but novel defences can be acquired or lost to a wider pool [43]. Our results are broadly consistent with this view in that differences in defence composition are unlikely to be driven solely by bacterial taxonomic effects, presumably due to frequent gain and loss via horizontal gene transfer.

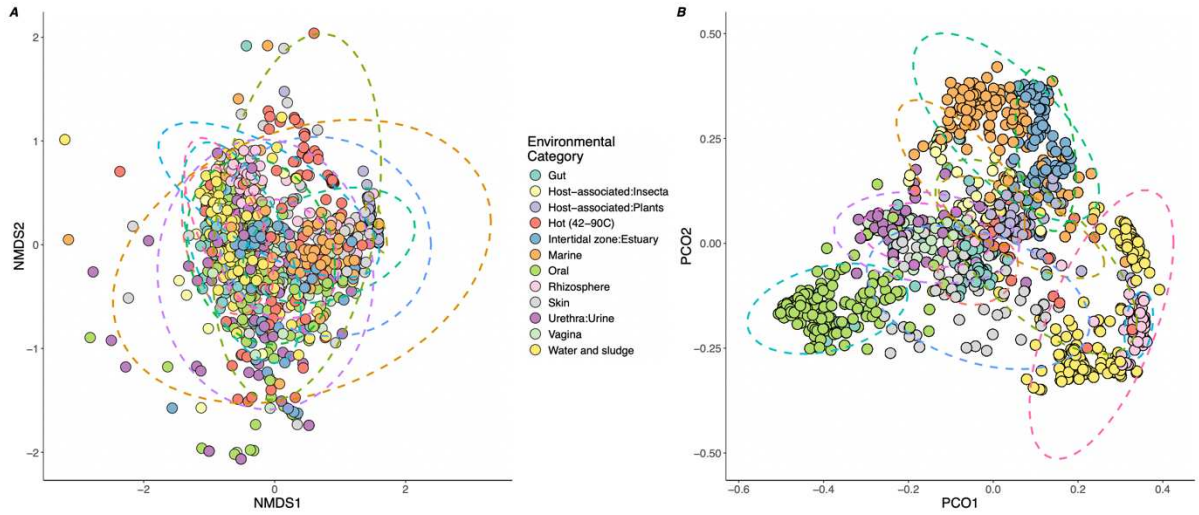


Fig. 2. Defence and taxonomic composition variation between environments Ordinations of metagenomic samples based on similarity in defence composition (A) or taxonomic similarity (B). Groups that cluster together share more defence systems or prokaryotic taxa respectively. Points represent individual metagenomic assemblies, ellipses represent 95% confidence intervals for a multivariate t-distribution and colours show the environment sampled. A) NMDS analysis was performed on a defence system abundance table constructed from a subsample of 1 million reads mapped to contigs carrying defence systems. B) A subsample of 500,000 reads were classified with Kraken2 and the frequency of each genus collected. Both ordinations use Bray-Curtis dissimilarity scores calculated from their respective abundance tables.

Defence abundance correlates with viral abundance across environments

In addition to abiotic factors, biotic factors (and in particular the viral community composition), are likely to influence the abundance of phage defence systems. To assess this, we first estimated the abundance of viruses in the sample based on coverage of viral contigs from a standardized subset of reads for each sample (1 million). These estimations therefore do not represent absolute viral abundances which likely vary substantially between environments, but measure viral abundance relative to microbial DNA sequences as the majority of reads are of bacterial origin. We found a significant correlation between overall defence abundance and viral abundance ($F_{1,997} = 115$, $p < 0.0001$, Fig. 3), suggesting that the density of viruses is a strong selective force for phage defence systems. As an additional test, we restricted the analysis to just those viruses annotated as Caudoviricites as a way of

excluding effects caused by non-phage environmental viruses. In this case we also observed a significant positive correlation with overall defence abundance and Caudoviricites abundance ($F_{1, 978}$ = 135.0, p < 0.0001, Fig. 3). Unsurprisingly, viral diversity and viral abundances were strongly correlated (Pearson correlation coefficient: 0.39, p < 0.0001). We refer herein to viral abundance but note that the accompanying viral diversity may also be contributing to the observed effects. We also identified the predicted genomic location of each defence system to assess the proportions of defences carried on MGEs. We found 54% of defences were located on chromosomal contigs, 32% on plasmid contigs, 13% on viral contigs and 0.3% on integrated prophages (Fig. S6). We note that many of the viral contigs likely represent fragments of prophages and the low number of integrated prophages results from the scarcity of fully intact prophage genomes complete with chromosomal flanks due to fragmented metagenomic assemblies.

CRISPR array abundance correlates with viral abundance across environments

Our previous work assessed the abundance of CRISPR defence systems across environments. We aimed to use a similar approach across the wider pool of samples used here, again focusing on CRISPR arrays rather than effector proteins. By focusing on CRISPR arrays we aimed to mitigate the difficulties of detecting multi-gene systems in fragmented, short contig, metagenomic data. As before, we found a positive correlation between CRISPR array abundance and viral abundance ($F_{1,911}$ = 4.33, p = 0.04) and a similar relationship when we restricted the analysis to Caudoviricites abundances ($F_{1,896}$ = 9.56, p < 0.01). We also assessed the origins of the arrays and found that similar numbers of arrays were predicted to be located on chromosomes (40.2%) and viral contigs (44.7%), which include prophages, with fewer predicted on plasmids (15.1%). When we repeated the correlations between CRISPR abundance and viral abundance for these subsets of the data, we found significant positive correlations between CRISPR abundance and viral abundance for the chromosomal ($F_{1,838}$ = 208.0, p < 0.0001) and plasmid ($F_{1,697}$ = 93.7, p < 0.0001) derived array abundances, but not virally derived arrays ($F_{1,814}$ = 0.94, p = 0.99). Taken together these results suggest that the selective forces that determine when CRISPR is beneficial may differ between bacteria and plasmids vs. phages, potentially due to stronger selection for streamlined genomes in viruses.

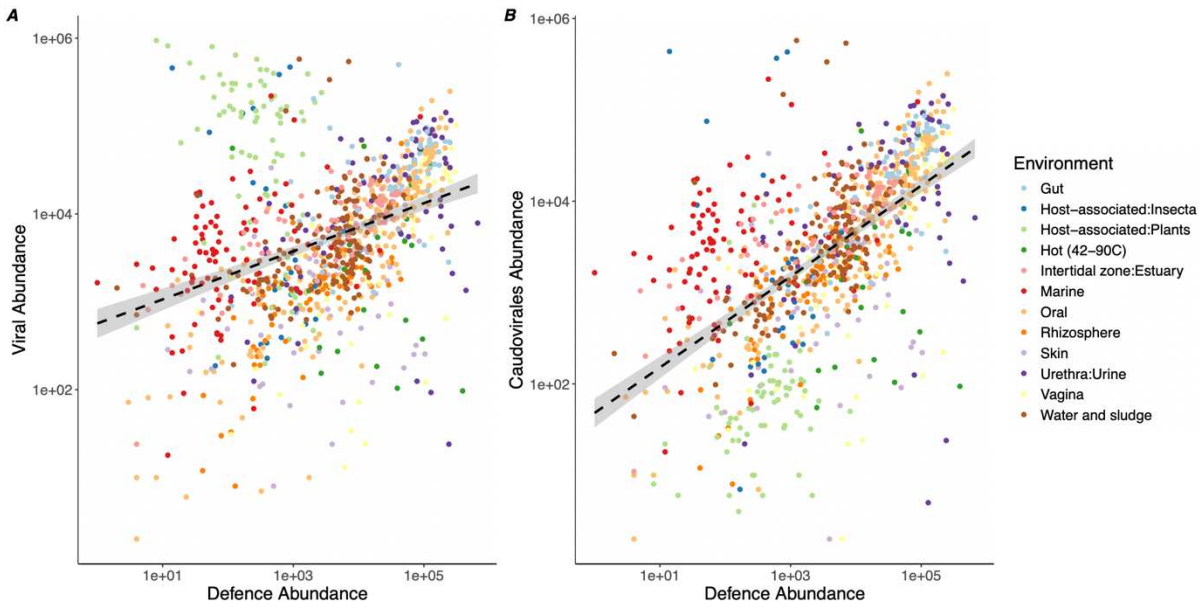


Figure 3. Defence abundance correlates with viral abundance.
Metagenomic assemblies were mined for phage defence systems and viral sequences. Coverage values were collected from a subsampled of 1 million reads per sample. Viral abundance represents the sum of all read counts for virally classified contigs; defence system abundance represents the sum of reads mapping to contigs carrying defence systems. The dashed lines represent linear models and the shaded area 95% confidence intervals. Panel A) shows the sum of all viral contig read counts while B) shows a subset of counts restricted to contigs annotated as the dsDNA tailed phage group Caudoviricetes.

Reduced defence system prevalence in plant-associated environments

While the abundance of defences generally correlated with viral abundance across environments, the plant metagenomes used in this study were obvious outliers, harbouring a highly reduced number of defence systems and elevated viral abundance (Fig. 3A). Plant microbiome samples typically suffer from high levels of plant DNA contamination due to sampling techniques [44]; however, recent work has also found phage defence systems to be underrepresented in plant environments [28]. The defence systems most abundant in the plant samples in our dataset were PD-T4, argonautes and Tiamat. Argonautes are well characterised plant immune effectors in RNA silencing immunity [45] and are therefore unsurprising to be

prevalent, if the metagenome is largely plant-host derived. To assess the taxonomic origins of argonaute and Tiamat systems we extracted the contigs containing these systems and classified them using the NCBI non-redundant nucleotide database. For the plant associated metagenomes, 86% of argonaute and 100% of Tiamat systems were located on plant genome sequences (Fig. S7). These observations of plant-derived sequences are supported by the presence of many viral sequences annotated as RNA viruses, despite the data being from metagenomic data, which can occur in the data as endogenous retroviruses integrated into the plant genome. Despite these technical aspects of plant microbiome sampling, further work must assess the biological reasons for an underrepresentation of phage defence systems in plant environments [28].

Discussion

Previous work has identified variation in both the frequency and types of phage defence systems found across different natural environments using metagenome assembled genomes (MAGs) [2]. Here, we search the full metagenome, at contig-level, for both defence systems and viral sequences to conduct a survey of the defence systems present in a broader range of environments using publicly available metagenomic data. We have previously shown that the abundance of a specific defence system, CRISPR-Cas, is strongly correlated with the relative abundance of viruses present in the environment [29]. We found a strong correlation with the total abundance of defence systems and viral abundance, consistent with the notion that the viral community is a strong selective force for the acquisition and retention of phage defence systems.

In agreement with other work, we found that gut samples harboured a greater abundance of defence systems than soil or marine environments respectively [2]. Notably, five out of the top six environments for viral abundance are human-host associated, with the exception being samples derived from environments 42°C or higher (such as hot-spring thermal environments). Hot-spring environments are suggested to be a hotspot for virus-defence systems due to higher costs associated with mutations, via reduced protein stability, in turn reducing viral diversity and the

potential for defence evasion [46]. By contrast, bacteria and archaea from mesophilic environments may be more able to tolerate a wider range of mutations, requiring more robust resistance mechanisms, such as mutation of surface receptors and subsequent phage resistance. Indeed, recent work suggests surface receptor variation can be a stronger predictor of successful phage infection than intracellular defence systems [47]. Our results are consistent with the notion of human-host environments providing a resource rich environment for microbes, in turn hosting a greater density of viruses and selection for defence systems.

Interestingly, marine environments appear to be particularly depleted in defence systems. Although the marine environment is predicted to have a high daily viral lysis, which is consistent with strong selection for defence, the virus-to-microbe ratio (VMR) is low [12]. Further, infection assays of culturable marine microbes found low predation pressures likely due to low encounter rates [48]. In addition, both the dominant marine clade, SAR11, and the marine cyanobacterium *Prochlorococcus*, has undergone genome reduction [49], presumably reducing the capacity for carrying diverse intracellular defence systems. The low overall density of defences we observed is consistent with relatively weak selection from phages and the typically low nutrient conditions may make the carriage of defence systems costly compared to selection for reduced genome sizes. In support of this conclusion, analysis of cyanobacteria and their phages found far more defence systems in freshwater genomes, which typically have higher nutrient availability, than those from marine environments [50]. In contrast to human host-associated samples, insect associated samples also carried far fewer defences. It has been observed previously that insect-associated bacterial genomes have few or no defences [23,51] and is likely either due to the general genome reduction processes that occur in intracellular insect symbionts [52] or reduced phage predation in the endosymbiotic environment (discussed in [53]).

Plant-associated samples also carried far fewer phage defence systems than human host-associated samples. The phyllosphere is typically low in carbon and nitrogen and relatively oligotrophic [54] again potentially increasing costs of phage defence system carriage. Recent work has found that plant-associated bacteria are depleted in defence systems relative to non-plant-associated relatives [28]. However, our

results may also be partially due to technical artefacts of sampling plant tissue and the typically high levels of host contamination. Specifically, in our dataset we found argonautes to be the most abundant defence, which are common in plant genomes, functioning as RNAi effectors [45]. The viral community was also consistent with this conclusion as although Caudoviricites was the most frequently identified viral group, we found many groups of Riboviria. These are RNA viruses capable of integrating into plant host genomes as endogenous retroviruses. Surprisingly, along with argonautes we also identified a high frequency of the Tiamat and PD-T4-6 defence systems. When we identified the origins of the Tiamat and argonaute systems, these were almost exclusively from plant sequences in the plant samples, versus a wide range of bacteria in the other environmental samples (Fig. S7). Further work is needed to assess the reasons why the plant-associated metagenomes were so depleted in defence systems [28] and the extent of defence conservation across domains of life [55].

By mining existing metagenomic assemblies, our results may be skewed towards viruses that are enclosed within a cell, either as prophages or undergoing active replication [56], although some viral particles will be present, as bulk metagenomes typically containing the most abundant viral genomes [57]. Yet experimental work has shown that the extracellular viromic fraction in an environment can change quickly, both temporally and spatially [57]. It is also challenging with metagenomic data to disentangle viral diversity and viral abundance as these factors strongly covary. We suggest experimental work in this area will yield valuable insights into the ultimate driver of defence system composition. Theory suggests that increased viral mutation rates limit the effectiveness of adaptive immunity [46] but further studies are needed. In addition, we focus entirely on DNA viruses, but RNA viruses are common [58,59] albeit less well studied. Assessing patterns of defence prevalence in light of RNA virus abundances, and integrating spatial and temporal information will be important future work. We also cannot rule out some biases in our search strategy, and currently available methods, as most defence systems, including those derived from a wide range of non-model bacteria and archaea, have been functionally validated in a limited number of model organisms [60,61]. It is possible that some incompatibility between defence-system origin host and the taxa chosen as model organism creates biases in defence system discovery. We suggest that future efforts

to identify viruses and defence systems from non-model organisms and environments will greatly expand our understanding of the role of the environment in shaping these interactions. We also note that our results are skewed towards smaller defence systems, with fewer core genes, due to the fragmented nature of typical metagenomic assemblies (Fig. S8). Longer, multi-gene defence systems are more likely to span multiple contigs in the assembly, and will therefore not be detected by defence identification tools, which rely on finding core genes and/or a minimum number of genes depending on the system. Finally, metagenomic assemblies are rarely exhaustive and likely represent the most abundant organisms in an environment; as such our defence system survey is representative of those that exist in the most abundant species and many others almost certainly exist in those environments at lower frequencies. Future efforts must focus on more contiguous assemblies derived from long-read sequencing that will be vital for more fine scale metagenomic analysis.

We found a consistent positive correlation between defence abundance and viral abundance; however, viruses and other MGEs are well-known to carry defence systems. Our analysis found up to 46% of defences predicted to be located on MGEs. Therefore, this high percentage of MGE associated defence systems may be driving the observed correlations. This leads to two possible interpretations: firstly, that defence accumulation on MGEs is a neutral process or ‘lottery’ effect, and will occur more frequently when MGE abundances are high, or secondly, that when MGE abundances are high there is greater competition between MGEs for susceptible hosts, requiring more defence systems to target competitors [11,62]. We suggest that experimental work in this area will yield valuable insights into natural microbial community dynamics of MGE competition and the interplay with defence systems.

Overall, we have identified wide variation in defence abundances across microbial environments and the density of viruses as a likely driver of selection for defence systems. Despite differences in defence abundance, there were minor differences in defence system composition across environments. This was surprising given the strong clustering of samples at the taxonomic level (Fig. 2) and suggests that while the environment predicts the overall abundance of defence, it does not strongly shape the defence composition. Clearly, further work that integrates community

ecology and metagenomic analysis is needed to assess whether the accumulation of specific defences is a stochastic process or determined by other unmeasured parameters. We also predict that further work may identify the ultimate selective forces acting on individual defence systems. As ever more defence systems are discovered we anticipate future studies will focus on the individual ecology of system types and classes of defence [8] and anti-defence [3], potentially identifying environmental hotspots that would allow targeted search strategies for defence discovery.

Data and Code Availability

All data used are publicly available from the MGnify database. A list of accession numbers is found in supplementary file 1. Code used for the analysis is available at github.com/s-meaden/mg_dfs.

Acknowledgements and funding statement

SM was supported by funding from the Biotechnology and Biological Sciences Research Council [BB/X009793/1]. This work was further supported by the Biotechnology and Biological Sciences Research Council (sLoLa grant BB/X003051/1), a UK Research and Innovation grant under the UK Government's Horizon Europe funding guarantee (EP/X030377/1), and the Philip Leverhulme Prize (PLP-2020-008) to E.R.W. PCF was supported by Bioprotection Aotearoa (Tertiary Education Commission, NZ) and by a James Cook Research Fellowship (RSNZ, Te Apārangi).

References

1. Hampton HG, Watson BNJ, Fineran PC. 2020 The arms race between bacteria and their phage foes. *Nature* **577**, 327–336. (doi:10.1038/s41586-019-1894-8)

2. Beavogui A, Lacroix A, Wiart N, Poulain J, Delmont TO, Paoli L, Wincker P, Oliveira PH. 2024 The defensome of complex bacterial communities. *Nature Communications* **15**, 2146.

3. Mayo-Muñoz D, Pinilla-Redondo R, Camara-Wilpert S, Birkholz N, Fineran PC. 2024 Inhibitors of bacterial immune systems: discovery, mechanisms and applications. *Nature Reviews Genetics* **25**, 237–254.

4. Tesson F, Bernheim A. 2023 Synergy and regulation of antiphage systems: toward the existence of a bacterial immune system? *Current Opinion in Microbiology* **71**, 102238. (doi:10.1016/j.mib.2022.102238)

5. Wu Y *et al.* 2024 Bacterial defense systems exhibit synergistic anti-phage activity. *Cell Host & Microbe* **32**, 557-572.e6. (doi:10.1016/j.chom.2024.01.015)

6. Maestri A *et al.* 2024 The bacterial defense system MADS interacts with CRISPR-Cas to limit phage infection and escape. *Cell Host & Microbe* **32**, 1412-1426.e11. (doi:10.1016/j.chom.2024.07.005)

7. Millman A, Bernheim A, Stokar-Avihail A, Fedorenko T, Voichek M, Leavitt A, Oppenheimer-Shaanan Y, Sorek R. 2020 Bacterial Retrons Function In Anti-Phage Defense. *Cell* **183**, 1551-1561.e12. (doi:10.1016/j.cell.2020.09.065)

8. Georjon H, Bernheim A. 2023 The highly diverse antiphage defence systems of bacteria. *Nature Reviews Microbiology* **21**, 686–700.

9. Van Houte S, Buckling A, Westra ER. 2016 Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiol Mol Biol Rev* **80**, 745–763. (doi:10.1128/MMBR.00011-16)

10. Berngruber TW, Lion S, Gandon S. 2013 Evolution of suicide as a defence strategy against pathogens in a spatially structured environment. *Ecology Letters* **16**, 446–453. (doi:10.1111/ele.12064)

11. Rocha EPC, Bikard D. 2022 Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biology* **20**, e3001514. (doi:10.1371/journal.pbio.3001514)

12. Chevallereau A, Pons BJ, van Houte S, Westra ER. 2022 Interactions between bacterial and phage communities in natural environments. *Nature Reviews Microbiology* **20**, 49–62.

13. Thompson LR *et al.* 2017 A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463.

14. Alseth EO, Pursey E, Luján AM, McLeod I, Rollie C, Westra ER. 2019 Bacterial biodiversity drives the evolution of CRISPR-based phage resistance. *Nature* **574**, 549–552. (doi:10.1038/s41586-019-1662-9)

15. Bikard D, Hatoum-Aslan A, Mucida D, Marraffini LA. 2012 CRISPR Interference Can Prevent Natural Transformation and Virulence Acquisition during In Vivo Bacterial Infection. *Cell Host & Microbe* **12**, 177–186. (doi:10.1016/j.chom.2012.06.003)

16. LeGault KN, Hays SG, Angermeyer A, McKitterick AC, Johura F, Sultana M, Ahmed T, Alam M, Seed KD. 2021 Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science* **373**, eabg2166. (doi:10.1126/science.abg2166)

17. Hussain FA *et al.* 2021 Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* **374**, 488–492. (doi:10.1126/science.abb1083)

18. Beamud B, Benz F, Bikard D. 2024 Going viral: The role of mobile genetic elements in bacterial immunity. *Cell Host & Microbe* **32**, 804–819.

19. Rousset F *et al.* 2022 Phages and their satellites encode hotspots of antiviral systems. *Cell Host & Microbe* **30**, 740-753.e5. (doi:10.1016/j.chom.2022.02.018)

20. Rollie C *et al.* 2020 Targeting of temperate phages drives loss of type I CRISPR–Cas systems. *Nature* **578**, 149–153. (doi:10.1038/s41586-020-1936-2)

21. Vercoe RB *et al.* 2013 Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. *PLOS Genetics* **9**, e1003454. (doi:10.1371/journal.pgen.1003454)

22. Watson BNJ, Staals RHJ, Fineran PC. 2018 CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. *mBio* **9**, e02406-17. (doi:10.1128/mBio.02406-17)

23. Tesson F, Hervé A, Mordret E, Touchon M, D'humières C, Cury J, Bernheim A. 2022 Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature communications* **13**, 2561.

24. Koonin EV, Makarova KS, Wolf YI. 2017 Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu. Rev. Microbiol.* **71**, 233–261. (doi:10.1146/annurev-micro-090816-093830)

25. Weissman JL, Laljani RM, Fagan WF, Johnson PL. 2019 Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. *The ISME journal* **13**, 2589–2602.

26. Wu R, Chai B, Cole JR, Gunturu SK, Guo X, Tian R, Gu J-D, Zhou J, Tiedje JM. 2020 Targeted assemblies of cas1 suggest CRISPR-Cas's response to soil warming. *ISME J* **14**, 1651–1662. (doi:10.1038/s41396-020-0635-1)

27. Liu Z-L, Liu J, Niu D-K. 2024 Bacterial defenses and their trade-off with growth are not ubiquitous but depend on ecological contexts. *bioRxiv* , 2024–03.

28. Bograd A, Oppenheimer-Shaanan Y, Levy A. 2024 Plasmids, Prophages and Defense Systems are Depleted from Plant Microbiota Genomes. *bioRxiv* , 2024–11.

29. Meaden S, Biswas A, Arkhipova K, Morales SE, Dutilh BE, Westra ER, Fineran PC. 2022 High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems. *Curr Biol* **32**, 220-227.e5. (doi:10.1016/j.cub.2021.10.038)

30. Richardson L *et al.* 2023 MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research* **51**, D753–D759.

31. Burgin J *et al.* 2023 The European nucleotide archive in 2022. *Nucleic acids research* **51**, D121–D125.

32. Shen W, Le S, Li Y, Hu F. 2016 SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one* **11**, e0163962.

33. Wood DE, Lu J, Langmead B. 2019 Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257. (doi:10.1186/s13059-019-1891-0)

34. Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M. 2007 The vegan package. *Community ecology package* **10**, 719.

35. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010 Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119. (doi:10.1186/1471-2105-11-119)

36. Payne LJ, Todeschini TC, Wu Y, Perry BJ, Ronson CW, Fineran PC, Nobrega FL, Jackson SA. 2021 Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic acids research* **49**, 10868–10878.

37. Payne LJ, Hughes TCD, Fineran PC, Jackson SA. 2024 New antiviral defences are genetically embedded within prokaryotic immune systems. , 2024.01.29.577857. (doi:10.1101/2024.01.29.577857)

38. Olijslager LH, Weijers D, Swarts DC. 2024 Distribution of specific prokaryotic immune systems correlates with host optimal growth temperature. *NAR Genomics and Bioinformatics* **6**, lqae105. (doi:10.1093/nargab/lqae105)

39. Rho M, Wu Y-W, Tang H, Doak TG, Ye Y. 2012 Diverse CRISPRs evolving in human microbiomes. *PLoS genetics* **8**, e1002441.

40. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PS, Nayfach S, Kyrpides NC. 2024 Identification of mobile genetic elements with geNomad. *Nature Biotechnology* **42**, 1303–1312.

41. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760.

42. Li H *et al.* 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

43. Bernheim A, Sorek R. 2020 The pan-immune system of bacteria: antiviral defence as a community resource. *Nat Rev Microbiol* **18**, 113–119. (doi:10.1038/s41579-019-0278-2)

44. Huang W, Gilbert S, Poulev A, Acosta K, Lebeis S, Long C, Lam E. 2020 Host-specific and tissue-dependent orchestration of microbiome community structure in traditional rice paddy ecosystems. *Plant Soil* **452**, 379–395. (doi:10.1007/s11104-020-04568-3)

45. Fang X, Qi Y. 2016 RNAi in plants: an argonaute-centered view. *The Plant Cell* **28**, 272–285.

46. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV. 2012 Viral Diversity Threshold for Adaptive Immunity in Prokaryotes. *mBio* **3**, 10.1128/mbio.00456-12. (doi:10.1128/mbio.00456-12)

47. Gaborieau B *et al.* 2024 Prediction of strain level phage–host interactions across the Escherichia genus using only genomic information. *Nature Microbiology* **9**, 2847–2861.

48. Kauffman KM, Chang WK, Brown JM, Hussain FA, Yang J, Polz MF, Kelly L. 2022 Resolving the structure of phage–bacteria interactions in the context of natural diversity. *Nature communications* **13**, 372.

49. Giovannoni SJ *et al.* 2005 Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science* **309**, 1242–1245. (doi:10.1126/science.1114057)

50. Lin W, Li D, Pan L, Li M, Tong Y. 2024 Cyanobacteria-cyanophage interactions between freshwater and marine ecosystems based on large-scale cyanophage genomic analysis. *Science of The Total Environment* **950**, 175201. (doi:10.1016/j.scitotenv.2024.175201)

51. Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, Fineran PC, Jackson SA. 2022 PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic acids research* **50**, W541–W550.

52. Moran NA. 1996 Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences* **93**, 2873–2878.

53. Siozios S *et al.* 2024 Genome dynamics across the evolutionary transition to endosymbiosis. *Current Biology* **0**. (doi:10.1016/j.cub.2024.10.044)

54. Bringel F, Couée I. 2015 Pivotal roles of phyllosphere microorganisms at the interface between plant functioning and atmospheric trace gas dynamics. *Frontiers in microbiology* **6**, 486.

55. Ledvina HE, Whiteley AT. 2024 Conservation and similarity of bacterial and eukaryotic innate immunity. *Nature Reviews Microbiology* , 1–15.

56. Coutinho FH, Rosselli R, Rodríguez-Valera F. 2019 Trends of Microdiversity Reveal Depth-Dependent Evolutionary Strategies of Viruses in the Mediterranean. *mSystems* **4**, 10.1128/msystems.00554-19. (doi:10.1128/msystems.00554-19)

57. Santos-Medellin C, Zinke LA, Ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. 2021 Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *The ISME journal* **15**, 1956–1970.

58. Neri U *et al.* 2022 Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037.e18. (doi:10.1016/j.cell.2022.08.023)

59. Hillary LS, Adriaenssens EM, Jones DL, McDonald JE. 2022 RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME communications* **2**, 34.

60. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R. 2018 Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120. (doi:10.1126/science.aar4120)

61. Gao L *et al.* 2020 Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* **369**, 1077–1084. (doi:10.1126/science.aba0372)

62. Koonin EV, Makarova KS, Wolf YI, Krupovic M. 2020 Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nature Reviews Genetics* **21**, 119–131.

Supplemental Information

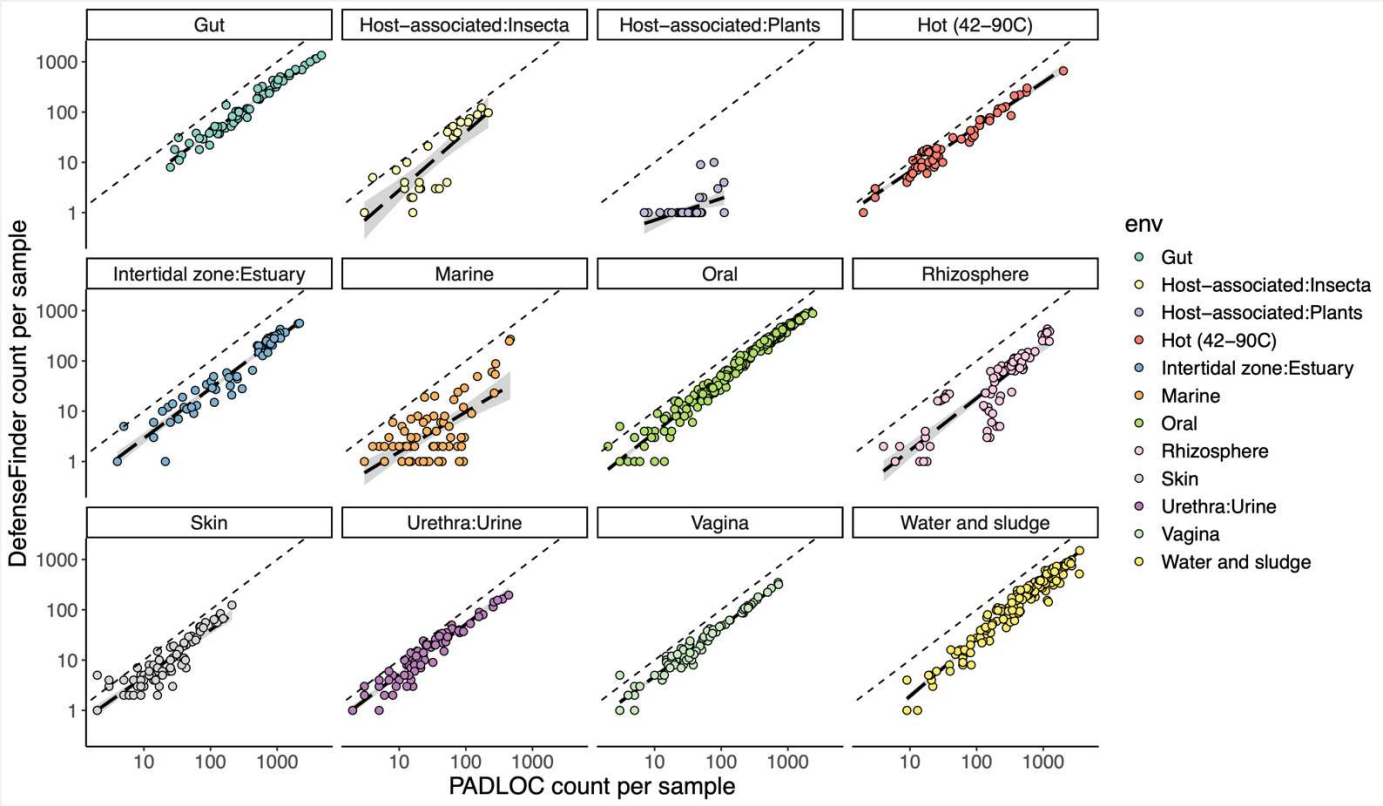


Fig. S1. Comparison of defence system identification tools

Correlations of the counts of contigs carrying defences identified with DefenseFinder (y-axis) or PADLOC (x-axis). Samples are grouped by environmental classification. Single dash lines represent a 1:1 correlation and double dash lines represent linear model fits with shaded areas showing 95% confidence intervals.

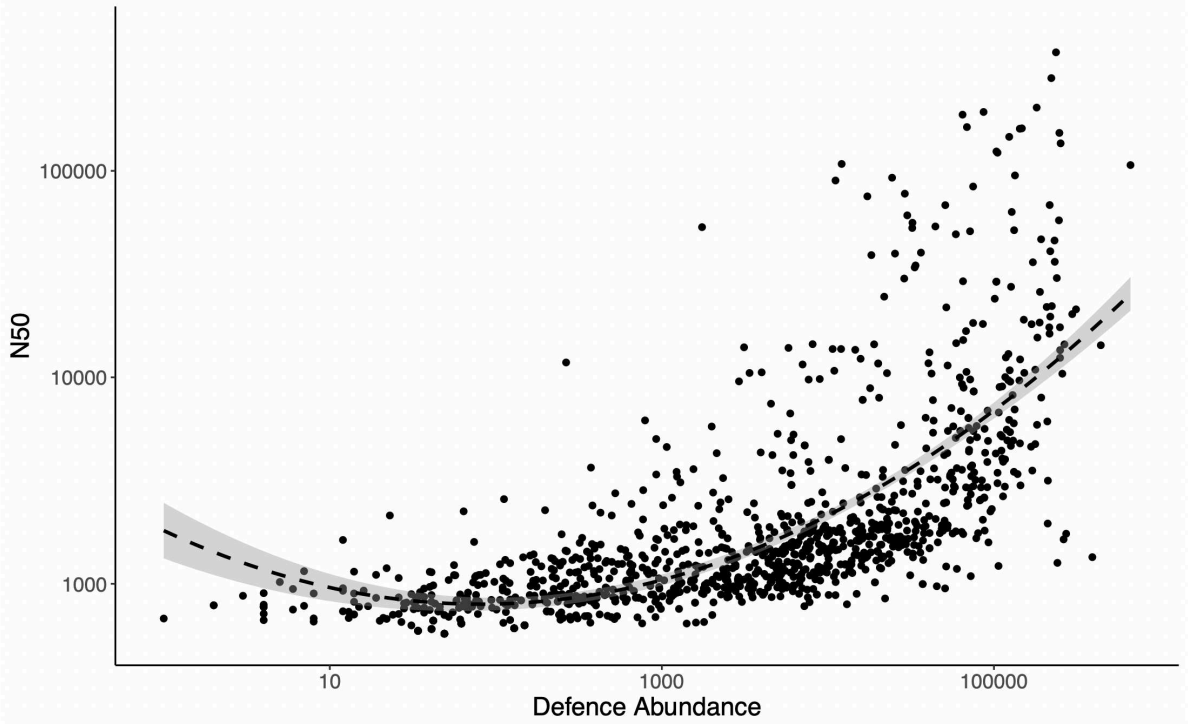
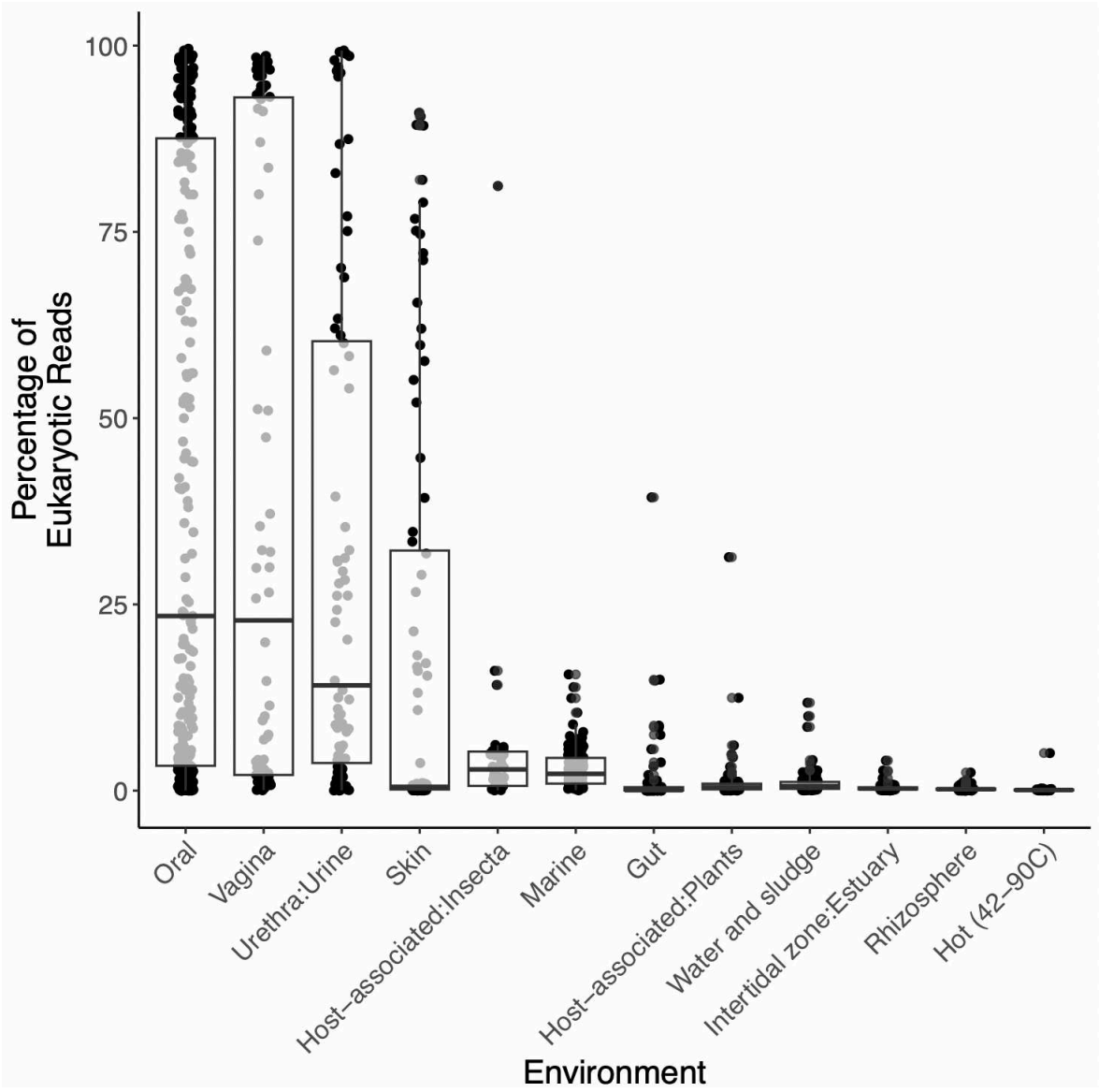


Fig. S2. Metagenomes with more contiguous assemblies (greater N50 value) recover more defence systems

Correlation between the N50 value of each assembly and the count of reads that map to defence carrying contigs. N50 values represent the shortest contig length, when ranked by length, to be included to cover 50% of the genome and we use here as a proxy for assembly fragmentation. Points represent metagenomic assemblies and dashed line represents a polynomial linear model, with shaded areas showing 95% confidence intervals.



752

753 **Fig. S3. Percentage of reads classified as of human host origin.**

754 Boxplots showing the percentage of reads classified as eukaryotic from the Kraken2
755 classifier. Samples are grouped by environmental classification and ordered based
756 on the median value. Each point represents an individual metagenome. Samples
757 with >50% of reads being classed as eukaryotic were excluded from downstream
758 analysis. Note that the standard Kraken2 database includes microbial genomes
759 (bacterial, archaeal, plasmid, viral, UniVec vectors and the human reference
760 genome) but not plant, fungal or protozoa.

761

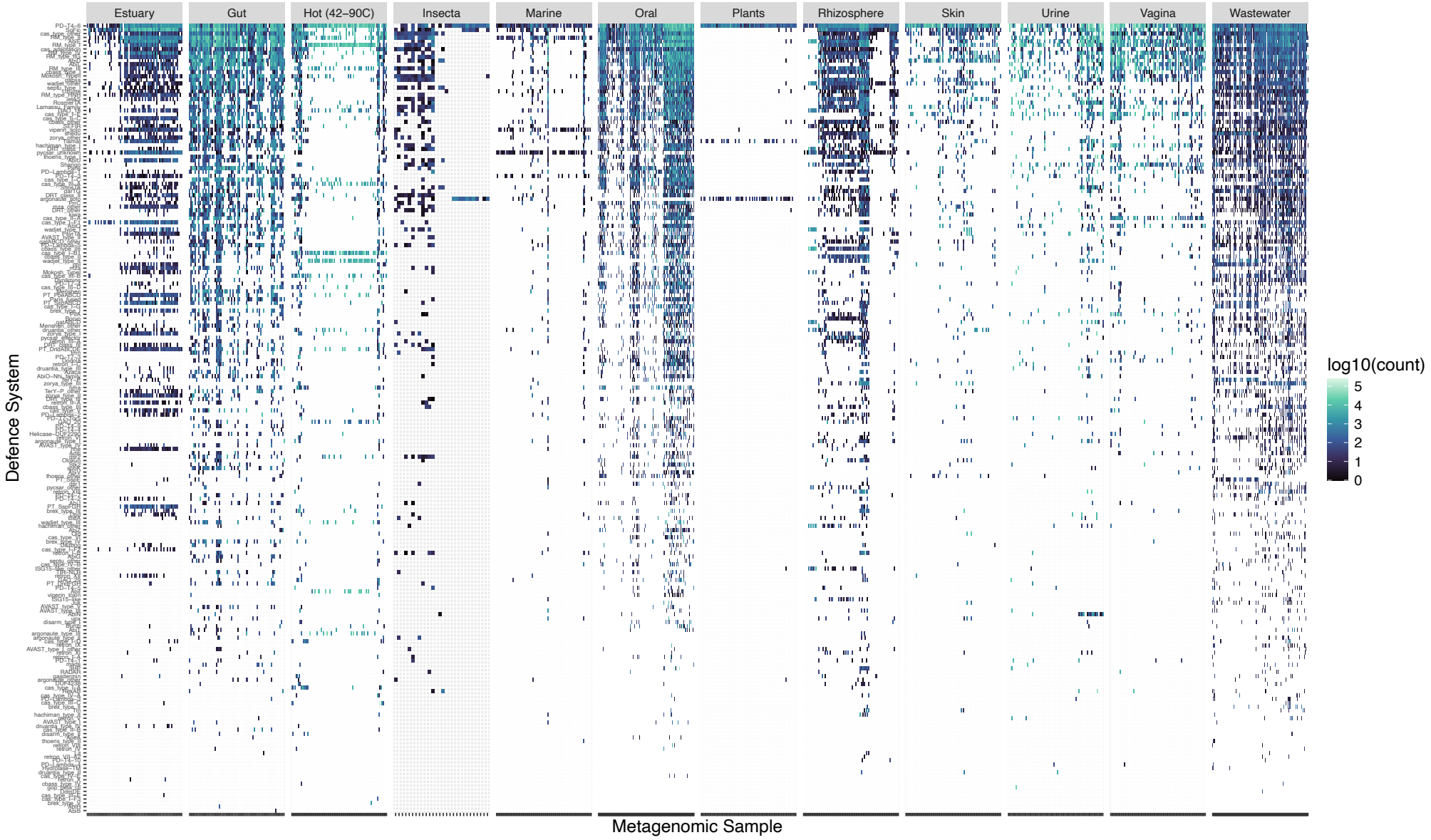

Estuary
Gut
Hot (42–90C)
Insecta
Marine
Oral
Plants
Rhizosphere
Skin
Urine
Vagina
Wastewater
Defence System
Metagenomic Sample
log10(count)
5
4
3
2
1
0

Fig. S4. Variation in defence system distributions across microbial ecoystems. Individual columns refer to metagenomic assemblies, with groupings representing the microbial environments samples were collected from. Rows represent individual defence systems and grid colours represent the sum of counts of reads recruited to contigs carrying each defence system from based on a standardised subsample of 1 million reads per sample.

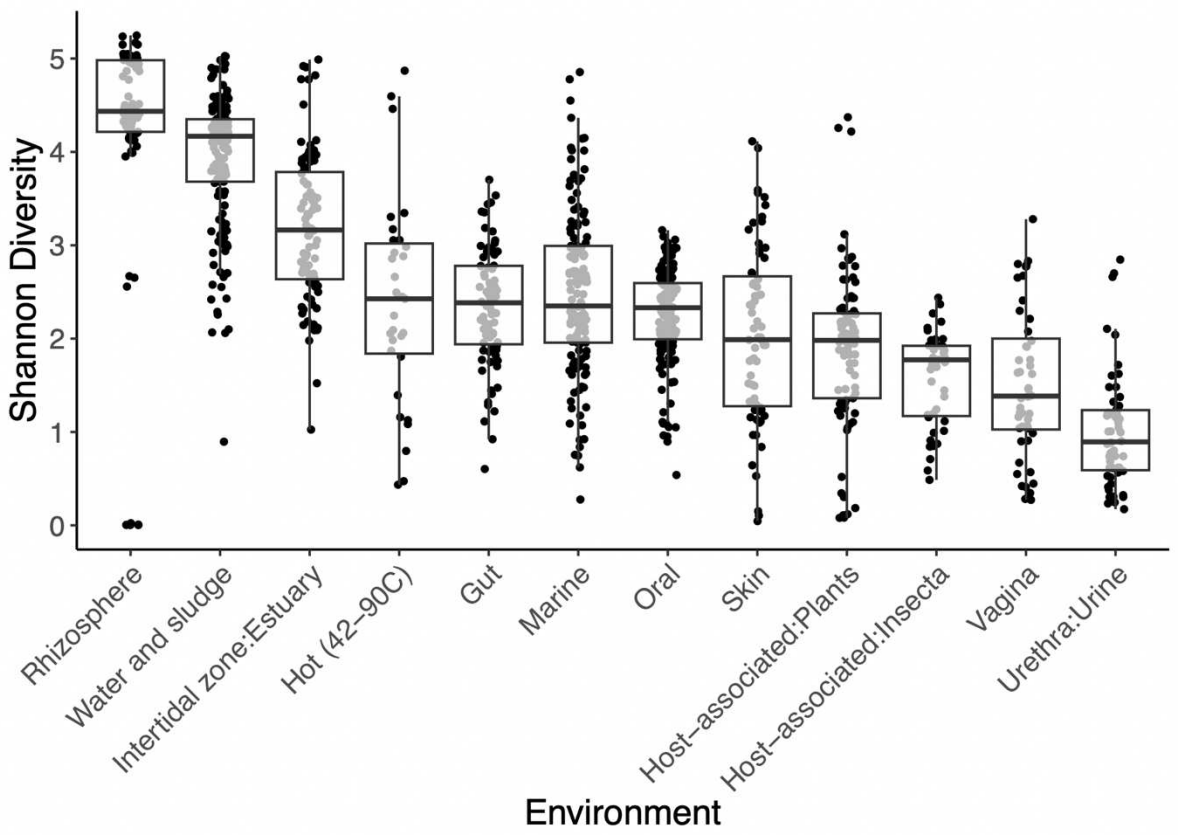


Fig. S5. Microbial community diversity varies by environment

Boxplots of the Shannon Diversity values based on the genera identified in each sample from Kraken2 analysis. A subsample of 1 million reads were classified with Kraken2 and the frequency of each genus collected. Categories are ordered by the median value per environment. Note these counts include all identified genera, but largely reflect bacterial and archaeal diversity.

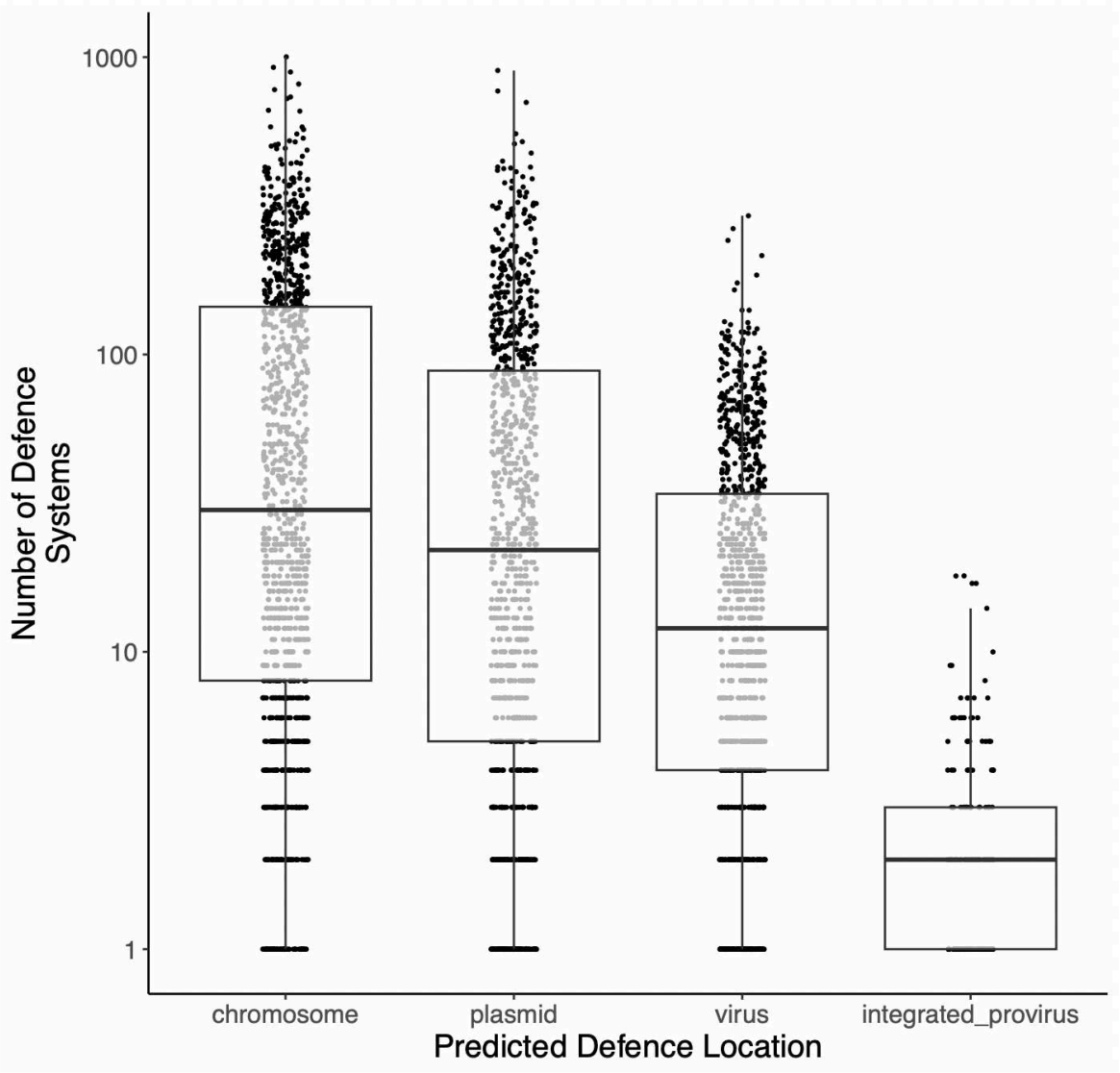


Fig. S6. Predicted locations of defence systems

All defence system carrying contigs were classified with the geNomad pipeline and assigned as chromosomal, plasmid, viral or integrated proviral based on the maximum score assigned by geNomad. For those contigs carrying an an integrated provirus, the genomic coordinates of the provirus was used to assign the defence as proviral or chromosomally encoded. Note y-axis here represents the frequency of defence systems, not abundance. We also note that many of the viral contigs likely represent fragments of prophages and the low number of integrated prophages results from the scarcity of fully intact prophage genomes complete with chromosomal flanks due to fragmented metagenomic assemblies.

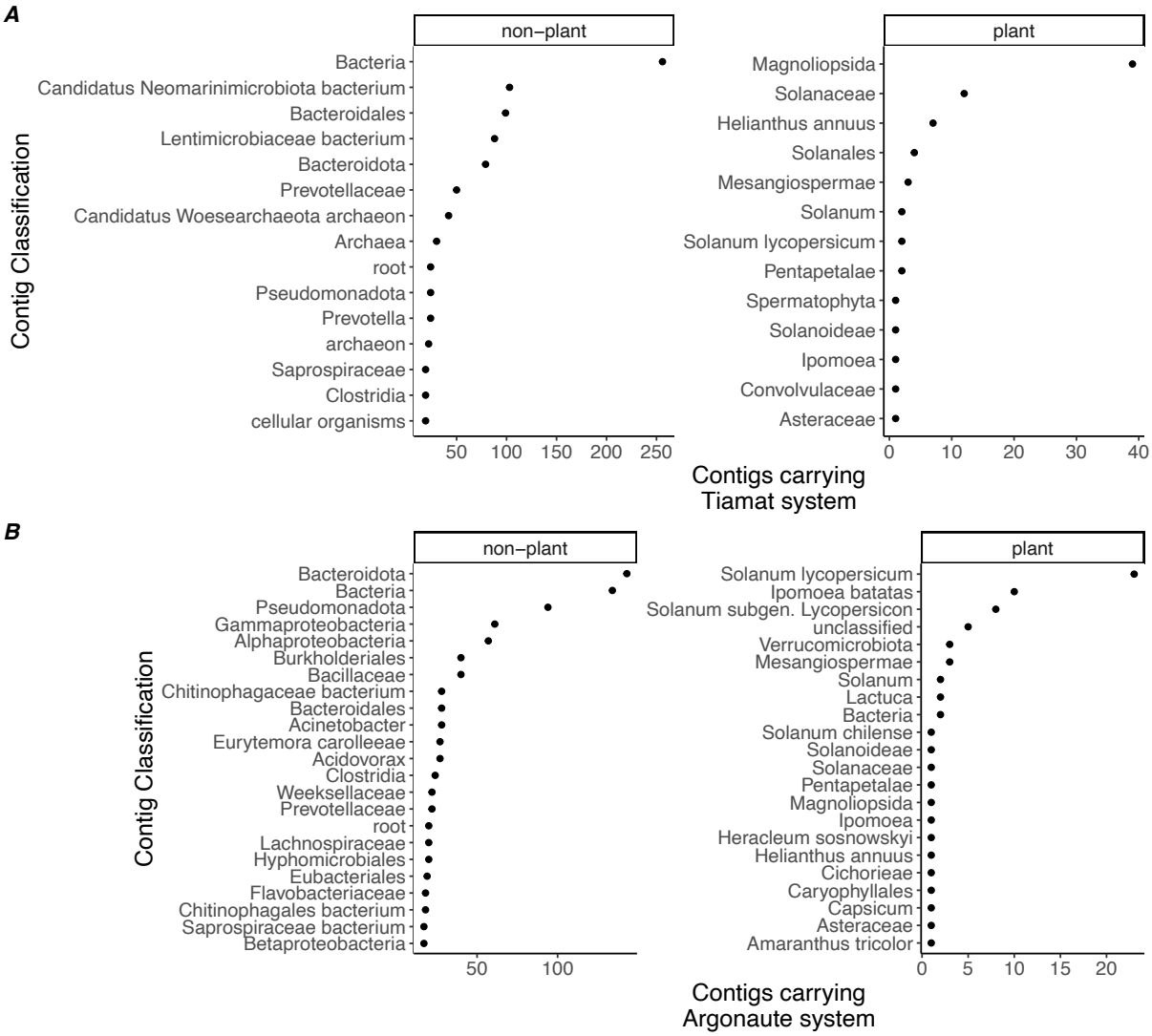


Fig. S7. Classifications of contigs carrying Tiamat or Argonaute systems

Counts of contigs extracted from the metagenome assemblies that were identified as carrying a Tiamat (A) or Argonaute (B) system. These contigs were classified using MMseqs2 against the NCBI nr nucleotide database. Counts of contigs are grouped by each last common ancestor (LCA) prediction. Samples were pooled according to plant or non-plant origin. Data shown encompass all classified contigs from the plant samples, but only the corresponding number of ranked LCAs for the non-plant samples.

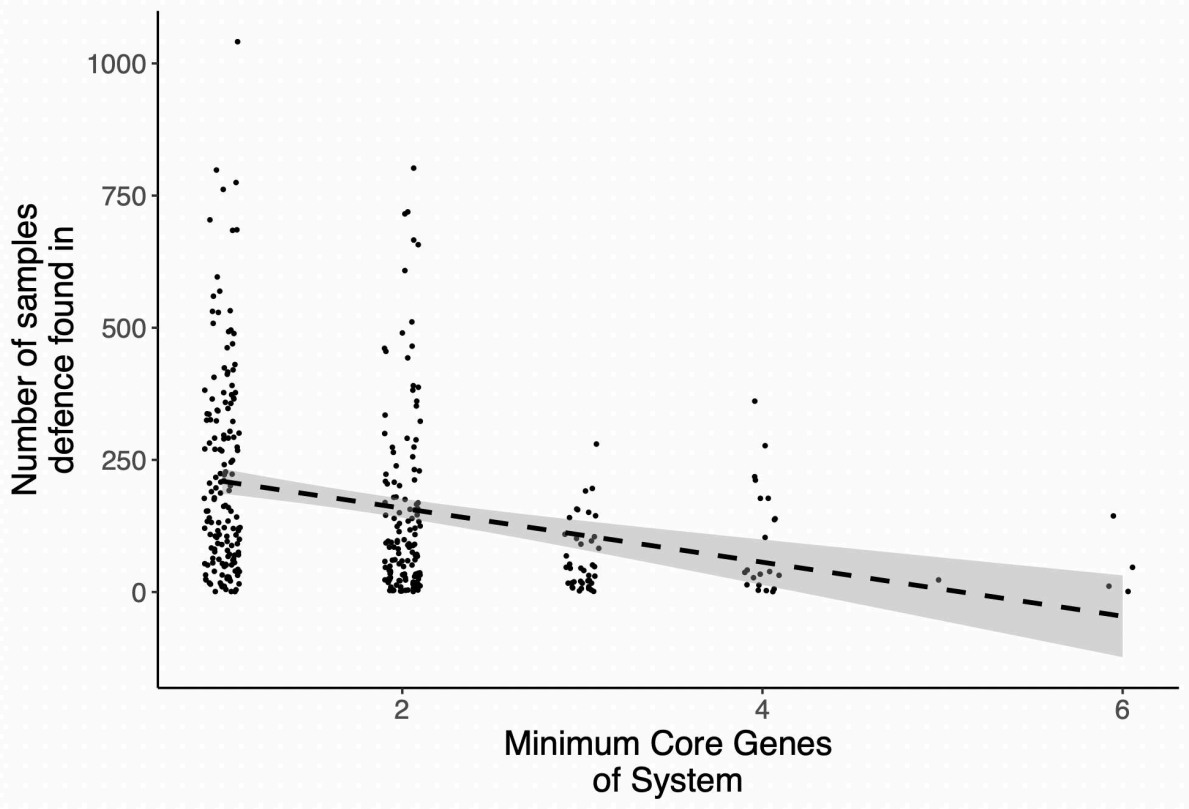


Fig. S8. Prevalence of defence system and number of minimum core genes

The minimum number of core genes required for each system was extracted from the PADLOC database system files. These values were plotted against the prevalence of each system i.e. the number of samples in which we detected that defence system. Dashed line represents a linear model fit and shaded areas represent 95% confidence intervals.