bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DIVE: A REFERENCE-FREE STATISTICAL APPROACH TO DIVERSITY-GENERATING & MOBILE GENETIC ELEMENT DISCOVERY

A PREPRINT

Jordi Abante
Department of Biomedical Data Science
Center for Computational, Evolutionary and Human Genomics
Stanford University
Stanford, 94305, USA

Peter L. Wang
Department of Biomedical Data Science
Department of Biochemistry
Stanford University
Stanford, 94305, USA

Julia Salzman[1]
Department of Biomedical Data Science
Department of Biochemistry
Department of Statistics (by courtesy)
Stanford University
Stanford, 94305, USA
julia.salzman@stanford.edu

ABSTRACT

Mobile genetic elements (MGEs) and diversity-generating mechanisms (DGMs) are paramount in microbial and viral evolution. These elements provide evolutionary leaps, conferring novel phenotypes, including those detrimental to human health, such as antimicrobial resistance. Unfortunately, state-of-the-art algorithms to detect these elements have many limitations, including reliance on reference genomes, assemblers, and heuristics, resulting in computational bottlenecks and limiting the breadth of biological discoveries. Here we present DIVE, a novel statistical, reference-free paradigm for *de novo* discovery of MGEs and DGMs by identifying k-mer sequences associated with high rates of sequence diversification. We use DIVE to analyze hundreds of samples to rediscover thousands of known MGEs and DGMs, quantify their activity, study their within-element variability, and identify their preferential integration loci. Using DIVE, we rediscover CRISPR, identify previously unreported CRISPR direct repeats in *Escherichia coli*, discover a novel CRISPR direct repeat in *Ruminococcus bromii*, identify antisense RNA5 as a putative novel class of non-coding RNA prone to integrate MGEs and identify the boundaries of antibiotic resistance hotspots in SXT integrative and conjugative elements in *Vibrio cholerae*. DIVE also identifies thousands of sequences associated with high diversity rates that cannot be mapped to known MGEs or DGMs, pointing to a substantial gap in the characterization of the microbial sequence space.

Keywords horizontal gene transfer · mobile genetic elements · diversity-generating mechanisms · CRISPR · transposable elements · integrative and conjugative elements · non-coding RNA

Introduction

The microbial and viral sequence universe is vast: recent estimates predict the existence of 10^{30} bacterial cells on earth [1] and 1 trillion microbial species [2], with unknown numbers of strains, all of which are continually evolving ([3], [4], [5]). Mobile genetic elements (MGE) contribute to this vast diversity, and include plasmids, integrative and

[1]Corresponding author

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

conjugative elements, transposons, retrotransposons, and prophages. While point mutations lead to slow but continuous evolution, MGEs abruptly modify the structure and organization of genomes ([6], [7], [8], [9]), and serve as a vehicle to transfer antimicrobial-resistance genes ([10], [11], [12]). The rapid spread of these elements is causing a surge of multidrug-resistant microbial strains [13].

Despite the importance of identifying MGEs, state-of-the-art algorithms remain dependent on annotations, references, and heuristics, failing to capture the genetic complexity of MGEs themselves or their insertion sites within hosts ([14], [15], [16], [17]). This is not surprising: due to microbial and MGE sequence space diversity, reference-based approaches to identify MGEs are likely to miss many MGEs, including those driving microbial evolution [18]. Further, the complexity of the sequence space poses computational challenges for typical bioinformatic algorithms unless they are restrictive or impose ad hoc filtering steps.

Like MGEs, the acquisition of portions of phage or foreign genomes into bacterial hosts is a process that has been long known to diversify bacterial genomes. This phenomenon is most well known in the clustered regularly interspaced short palindromic repeats (CRISPR) systems. However, it has historically been observed that some MGE tend to integrate into specific loci, such as in genomic regions coding for transfer RNA ([19], [20], [21], [22]). Because of the continual acquisition of phage elements, it is impossible to represent their diversity in reference genomes. Since >40% of sequenced bacteria lack annotated CRISPR systems [23], it is of great interest to flag other hotspots of phage integrants to identify novel molecular pathways that could be similar in function to CRISPR-Cas. No published algorithm provides such a framework. Illustrating this, state of the art methods to discover CRISPR repeat arrays rely on heuristics and prior knowledge such as repeat composition or protein domains related to the Cas system ([24], [25], [26], [27], [28], [29], [30]), hampering discovery.

To address these challenges, we present DIVE, a purely statistical and completely annotation-free algorithm that proposes a new conceptual approach to discovering k-mer sequences associated with high rates of sequence diversification. DIVE is an efficient algorithm designed to identify sequences that may mechanistically cause sequence diversification (e.g., CRISPR repeat or transposon end) and the variable sequences near them, such as an insertion site (Fig. 1). The identified sequences are assigned statistical scores for biologists to prioritize them.

We ran DIVE on large collections of isolate sequencing from the best-studied human pathogens: *E. coli* and *V. cholerae*, as well as samples from the *N. gonorrhoeae* isolates and rotavirus infection metagenomic samples, to both evaluate DIVE's ability to rediscover known MGE and CRISPR direct repeats and show its discovery power. We show that DIVE identifies sequences enriched in true positives, such as known MGEs, and their insertion sites, and finds highly variable sequences interspersed in non-coding RNA (ncRNA) loci in *E. coli* and *V. cholerae*, which are known to be magnets of foreign genetic material. In addition, DIVE identifies putative cargo gene hotspots in SXT integrative and conjugative elements (ICEs) and antisense RNA as a potential new class of ncRNA subject to MGE insertion in *V. cholerae*. Furthermore, DIVE finds 23,125 unique sequences associated with high levels of sequence diversification in *E. coli* and *V. cholerae* that do not map to known elements (BLAST e>0.25), pointing to a potential for a currently unannotated universe of MGEs and phage integration hotspots even in well studied bacterial species.

Reference-free method for diversity-generating and mobile genetic elements discovery

DIVE is a statistical algorithm that operates directly on sequencing reads. It is built from the following simple observation: a mobile genetic element such as a transposon is defined by its bounding sequence or 'transposon arms' A. These arms mechanistically enable mobility, but they also define the element in an algorithmic sense: the constant sequence A will be flanked by a highly diverse set of sequences if A is indeed the arm of an MGE. Note this principle also applies to sequence elements that are magnets for phage integration, as well as other assemblies of constant and variable sequences such as CRISPR repeats. Thus, we reasoned that it is possible to identify MGE, phage insertion sites, and CRISPR repeats *de novo* by identifying sequences "A" with highly diverse (to be defined) neighboring sequences.

DIVE makes the preceding logic into a statistical algorithm. To explain the DIVE algorithm, we define "*anchor*" and "*target*," both k-mers of a predefined length k in a sequencing read. DIVE aims to find anchors that behave like transposon arms in that they neighbor statistically highly diverse sequences, the targets. In the transposon example, targets are sequences neighboring these arms (Fig. 1). DIVE takes as an input a FASTQ file and processes each read sequentially using a sliding window to construct *target* dictionaries for each k-mer (*anchor*) encountered in each read. DIVE then computes a score for each anchor to quantify levels of target diversity. Annotation and references are not used during any step of DIVE, except as a post-facto option for interpretations (see below).

DIVE operates symmetrically, forming upstream and downstream scores for each anchor; we describe the procedure here for downstream targets. For each anchor, DIVE generates a target dictionary with an online clustering method that collapses targets within "sequencing error" distance. It then models the number of clusters formed at each step

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

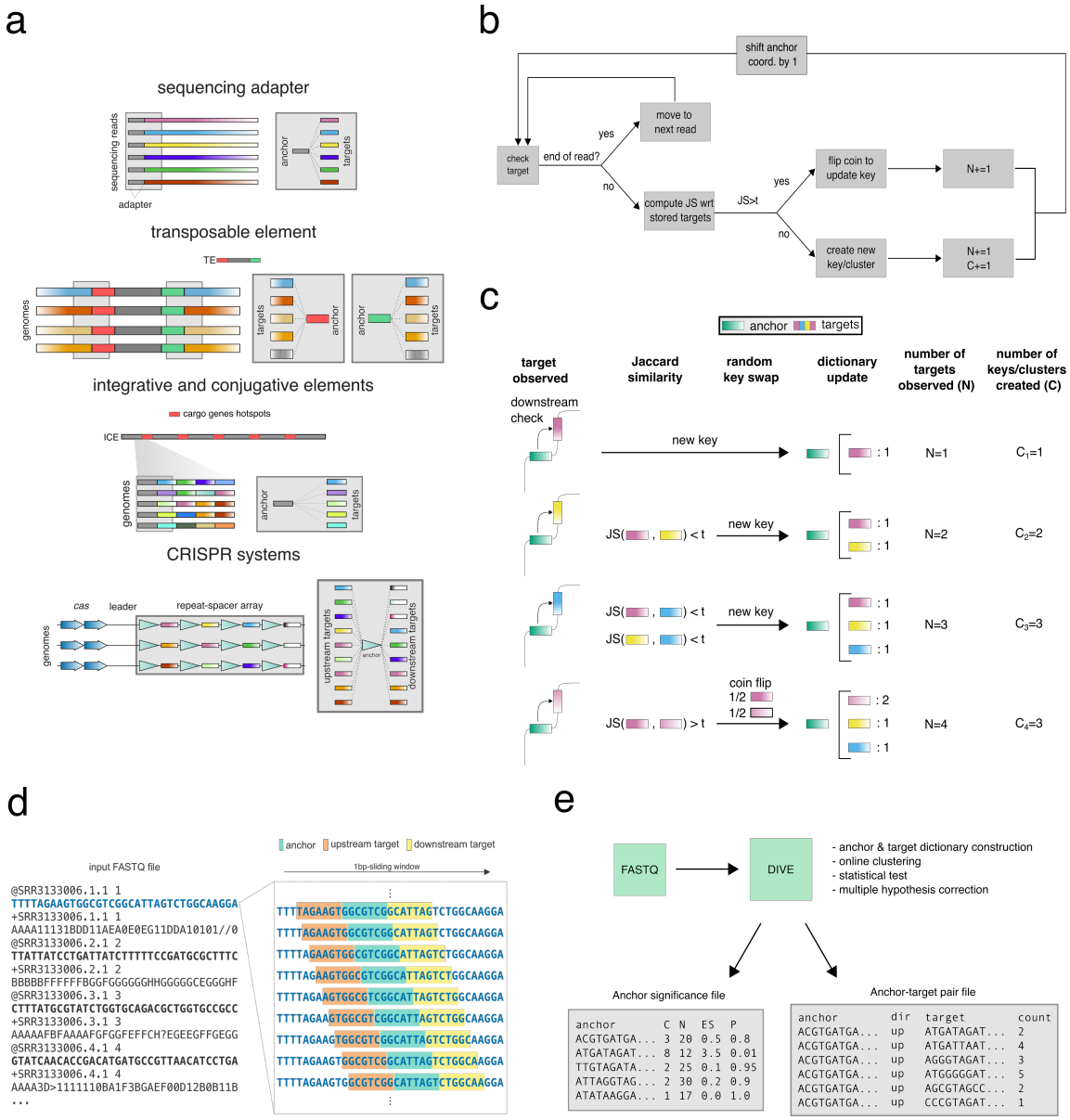


Figure 1: Motivating examples and DIVE algorithm. a. Sequencing adapters, transposable elements, clustered interspersed palindromic repeats (CRISPR) arrays, and sequences near cargo gene hotspot boundaries in integrative and conjugative elements (ICE) contain a diverse set of neighboring sequences. Sequencing adapters serve as technical positive controls. **b.** Flow diagram of DIVE's logic while parsing a FASTQ file. **c.** Cluster formation process based on Jaccard similarity (JS) for a given anchor. Each anchor has an upstream and a downstream target sequence dictionary. The example illustrates the cluster formation process for the downstream case. **d.** DIVE processes reads sequentially through a sliding window that moves along the sequencing reads, recording for each anchor k-mer the upstream and downstream k-mers, or target sequences, in the corresponding dictionary. **e.** DIVE takes a FASTQ file as input and stores the output in two output files. The anchor significance file contains information about each anchor, including the statistics and p-values, in addition to other metrics as nucleotide content. DIVE also produces an anchor-target pair file which contains information about the stored target sequences in the dictionary, the direction and the count for each anchor.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

using a Poisson-Binomial model (Fig. 1; Methods). An anchor must be observed sufficiently often ($> N_{\text{min}}$ times, see Methods) to have adequate statistical power to call its targets more diverse than expected by chance. To keep the memory footprint low, for every 100k reads, DIVE computes the probability that an anchor will be observed at least N_{min} times in the FASTQ file ($N_{\text{min}} = 25$). If this probability is <1%, DIVE does not accept newly observed anchors into the dictionary (Methods). After the FASTQ file is completely traversed, DIVE uses a Poisson-Binomial model to compute the probability $P_{H_0}(C_N \geq c_N \mid N)$ that the number of clusters C_N exceeds the observed value c_N under the null hypothesis that the observed target diversity is due to background variability (Methods). DIVE uses the Benjamini-Hochberg (BH) correction to control the false discovery rate (FDR) (Methods). Anchors producing significant target diversity are clustered using DBSCAN with the Levenshtein edit distance as a metric to remove redundancies. Within each anchor cluster, the anchor with the largest effect size in each direction is picked as the *representative anchor* (Methods). The subsequent analysis in this paper has been done using the representative anchors, which we will refer to as RAs from here on out. Representative anchors, the number of clusters (C), the number of times a target was observed with the anchor (N), the effect size ($\log_2(c_N/E[C_N \mid N]$), and corrected *p*-values are reported. All numbers described above, such as k, 1%, and 100k, are user-adjustable parameters. To facilitate interpretation, DIVE aligns the RAs (BLAST) to a set of databases provided by the user (Methods).

Artificial sequences such as sequencing adapters will exhibit statistical signatures like bona fide MGEs (highly variable downstream sequences) and serve as positive controls (Fig. 1). DIVE identified several such sequences, and we removed them before proceeding to the rest of the downstream analysis (Methods). We also discarded RAs with low nucleotide entropy ($H < 3$) to remove other potential technical artifacts (Methods).

DIVE is designed to quantify target diversity and evaluate its statistical significance (Fig. 1). The effect size computed by DIVE can be used to prioritize actively diversifying CRISPR arrays over less active arrays or highly mobilized over ancestral MGEs. We sought to compare DIVE to a naive algorithm that quantified properties of the RA sequence (e.g., abundance, repetitiveness) by correlating DIVE's effect size and the number of observations N, a proxy for the degeneracy of a k-mer (OLS $R^2 = 0.02$). The lack of correlation supports the point that k-mer prevalence alone cannot predict diversifying sequences (Fig. S1). To further show that DIVE captures information, we compared the proportion of RAs annotated as known MGE, CRISPR direct repeats, and CRISPR spacers to the fraction chosen randomly among all RAs detected in *E. coli* and *V. cholerae*. DIVE RAs are strongly enriched for mapping to annotated MGEs in both *E. coli* and *V. cholerae* (Table S3; $\text{OR}_{E.\ coli}$=4.87, $\text{OR}_{V.\ cholerae}$=3.15), whereas RAs annotated as CRISPR direct repeats showed the largest enrichment in both datasets ($\text{OR}_{E.\ coli}$=161.31, $\text{OR}_{V.\ cholerae}$=647.49). The vast presence of MGEs in bacterial genomes makes it plausible for a random k-mer to originate from such an element by chance. We also observed an enrichment of eukaryotic transposable elements (TEs). However, all the matches to the eukaryotic TE database (>80% identity) corresponded to tRNA genes present in the database and not actual TEs.

MGEfinder, a *de novo* MGE detection tool, outperformed state-of-the-art methods in a recent benchmark [17]. However, this tool requires a reference genome, the quality of which affects the algorithm's sensitivity. In addition, the algorithm relies on detecting target site duplication (TSD) events, which are not always present. Furthermore, MGEfinder does not detect within-MGE variability (e.g., cargo gene turnover in ICE). To test if the increased theoretical power of DIVE compared to MGEfinder is born out in real data, we ran DIVE on data highlighted in the MGEfinder study. We analyzed 200 *N. gonorrhoeae* isolates used in [17] and compared the detection power of DIVE to reported results by MGEfinder.

DIVE completed in <1 hour per sample, using less than 100Gb. DIVE reported 81,792 unique RAs mapping to 190 different TEs (BLAST *e<0.01*) compared to 28 TEs reported by MGEfinder in the 891 isolates analyzed in [17]. TEs detected by DIVE could constitute ancestral TEs present in the *N. gonorrhoeae* genome. To test this, we ran DIVE on unaligned reads after mapping (bowtie2) to the reference genome as in [17] (overall alignment rate 88%). DIVE called 868 unique RAs mapping to 107 different TEs (Table S4, Fig. 2). Tn3 transposons TnXc5, TnTsp1, and TnArsp6 were most prevalent, whereas TnXAj417 and Tn3434 showed the largest median effect size. The large effect sizes found in reads failing to align to the reference genome lead us to hypothesize that these elements are active TEs in *N. gonorrhoeae*. The reporting between DIVE and MGEfinder is not comparable since the latter does not cross-reference with any TE database.

DIVE rediscovers MGE and CRISPR systems in single-isolate sequencing and provides quantitative estimates of their activity

We ran DIVE on 514 *E. coli* isolates from environmental sources sequenced by the FDA (Methods) since this organism is well-studied and annotated, including its CRISPR systems. Each sample took approximately 40 minutes to process using less than 100 Gb of RAM.

30% of the total 166,335 unique RAs called by DIVE, mapped (BLAST *e<0.01*, Methods) to 3,412 known MGEs, 218 eukaryotic and 206 and prokaryotic TEs, 219 ICEs, 774 CRISPR spacers, 307 CRISPR direct repeats, and 627

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

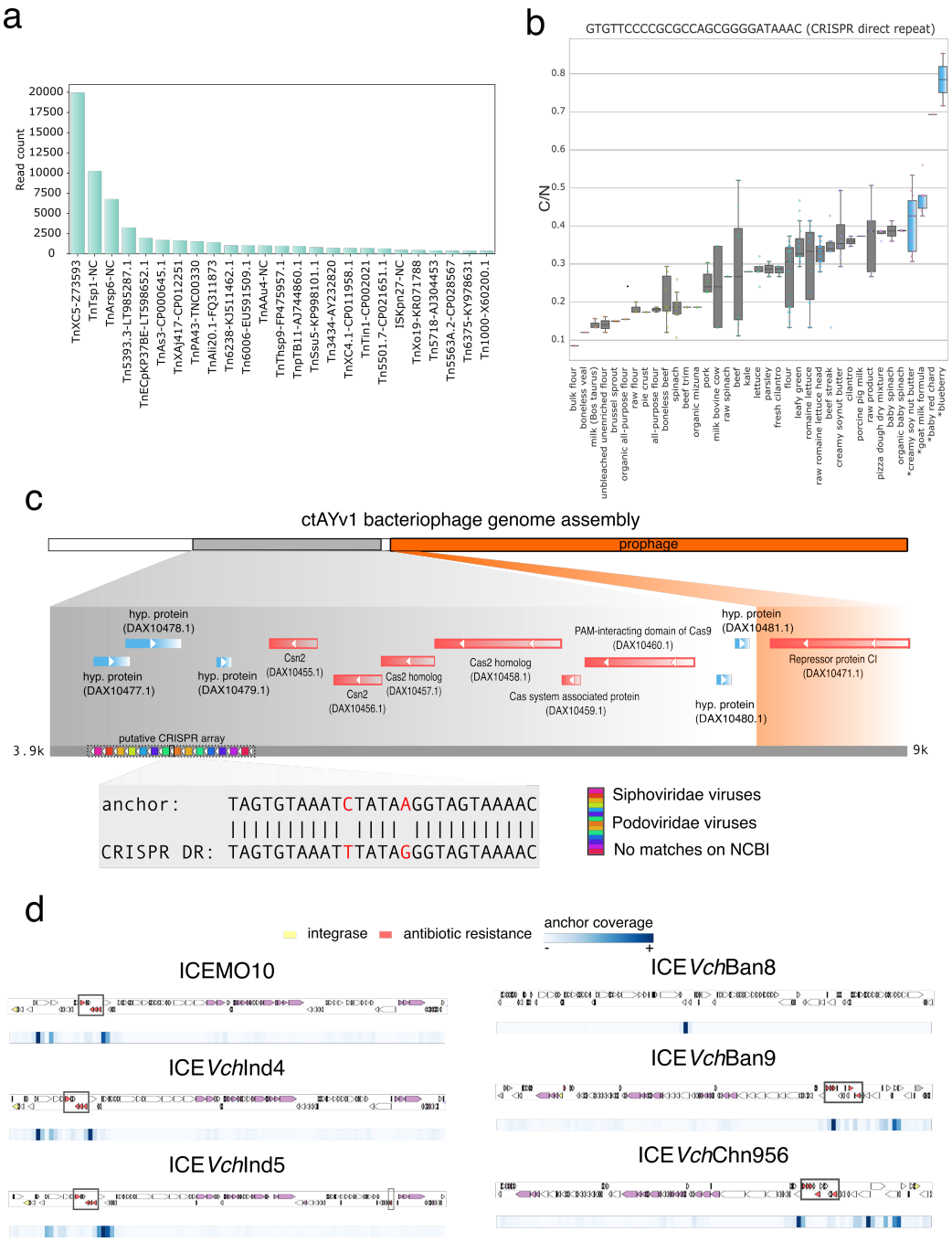


Figure 2: DIVE quantifies activity of MGE and DGMs. **a.** Number of reads containing an anchor mapping to each transposable element (TE) detected by DIVE among unaligned reads in *N. gonorrhoeae* data from Durrant et. al 2020. **b.** Boxplot of the ratio of number of clusters to the number of observation (C/N) of the anchor sequence GTGTTCCCCGCGCCAGCGGGGATAAAC called by DIVE in E. coli. The target diversity in four environments (shown in blue) is significantly larger than that of the rest of environments (Methods). **c.** Putative direct repeat TAGTGTAAATCTATAAGGTAGTAAAAC detected by DIVE in the Rotavirus dataset. The anchor reported by DIVE is two substitutions away of a known CRISPR direct repeat, and maps to the genome assembly of a newly characterized phage (ctAYv1) following the disposition of a canonical CRISPR array. **d.** Alignment of all significant anchors from the *V. cholerae* analysis to the six available sequences of integrative and conjugative elements (ICEs) in the ICEberg database. The annotated genes for each SXT ICE are shown above the plot (yellow genes: integrases; red genes: antibiotic resistance). Below the ICE, a heatmap shows the coverage of anchors with respect to the element. Anchors tended to cluster in the neighborhood of known antibiotic resistance genes in all but ICEVchBan8, where they overlap with a known transposase.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RNA genes from Rfam (Table S1, S2, Methods). 12% of RAs called by DIVE were unannotated (BLAST $e \geq 0.25$). Because CRISPR repeats arrays have a unique organization, we tested if DIVE could identify repeat arrays using purely statistical features.

In *E. coli*, RAs mapping to CRISPR DRs showed the largest median effect size (Fig. S1) and represented 301 known DRs. Their target sequences map to 774 known spacers (bowtie2). In addition, DIVE RAs perfectly matched 50/100 DRs previously reported in *E. coli*, and 8 DRs reported in Enterobacteria (7) and Proteobacteria (1) but not annotated as present in *E. coli* (Table S5). Note, however, that out of the 100 DRs previously reported in *E. coli*, only 91 are actually present in the dataset and only 59 had the minimum sample size required in at least one sample ($N_{\text{min}} = 25$). Notably, RAs mapping to CRISPR DRs were distinguishable from all other RAs. These consisted of representative RAs belonging to anchor clusters with bidirectional significance and appeared repeated in the sequence adjacent to the RA in at least a read. Imposing these requirements yields ten RAs perfectly mapping to six CRISPR DRs known in *E. coli* (Table S6), achieving 100% specificity for CRISPR DRs and suggesting a statistical method to identify un-annotated repeats *de novo*, including ones that lack annotated *Cas* proteins in *cis*.

To test if this were possible, we ran DIVE on metagenomic sequencing of a *Rotavirus* infection (Methods). DIVE called 133,528 unique RAs mapping to 2,991 MGEs and 575 CRISPR DRs. Imposing the above criteria revealed one RA did not perfectly match any known DR in the DR database but aligned 15 times (BLAST $e<0.01$) to a phage identified in metagenomic data [31](Fig. 2). The RA was not found in any other sequence on the NCBI nucleotide collection (nr/nt) or the DR database but has 2 mismatches from the closest DR. We note that BLAST of the RA to the NCBI WGS database restricted to the bacterium *Ruminococcus bromii* (taxid:40518) gives perfect matches in a locus with CRISPR-associated (Cas) genes very similar to that in the phage contig (other parts of the phage contig also match to *R. bromii*). 13/43 targets mapped (BLAST $e<0.01$) to *Siphoviridae (13/13), Podoviridae* (2/13), and *Myoviridae* (1/13), supporting it as a CRISPR system. Thus, we hypothesize this RA constitutes a novel functional CRISPR DR.

DIVE provides additional information about CRISPR repeats beyond discovering them *de novo*. Available CRISPR detection algorithms do not estimate the activity level of CRISPR systems in response to environmental changes. DIVE's effect size establishes a framework to estimate variation in spacer turnover in CRISPR arrays as a function of an isolate covariate (Methods). In the *E. coli* dataset, five RAs mapped to *E. coli* CRISPR DRs had an effect size that was a statistically significant function of the isolation source (Fig. 2, Fig. S4). For example, for DR GTGTTCCC-CGCGCCAGCGGGGATAAAC, the target diversity varies 10-fold between isolates obtained from blueberry and bulk flour (Fig. 2). Beyond CRISPR, 77 RAs mapping to lambdoid phages, ICEs, TEs, and RNA had target diversity that is a statistically significant function of the isolation source (Fig. S2, S5, S6).

Non-coding RNA loci show large target sequence hyper-variability

Non-coding RNAs (ncRNA) loci are known magnets for foreign genetic material ([19], [20], [21], [22]). In tRNA, MGEs contain a primer binding site presenting complementarity with the corresponding primer tRNA [20]. Thus, we decided to investigate whether DIVE would rediscover this known phenomenon and potentially identify new primers used by MGEs. Indeed, DIVE found strong target diversity among RAs mapping to ncRNAs, including several examples where the target variability was a function of available covariates (Fig. S6).

In *E. coli*, DIVE called 712 unique RAs that showed target hyper-variability mapping to 84 different Rfam accessions annotated as tRNA genes (BLAST $e<0.01$). The target diversity of seven RAs, mapping to four unique tRNA genes, was significantly associated with the isolation source (Table S7; Fig. S6). These included Met-tRNA, a known target of retrotransposons [32], and a valU operon encoding three tRNA genes. The RAs tended to align to the ends of the gene and accrued up to 3,000 unique targets resulting in 1,400 clusters (Methods). However, no RA mapped within an individual genome on the NCBI nucleotide database more than eleven times (BLAST $e<0.01$), suggesting the observed target variability cannot be explained alone by the within-genome variability. Only 2% of the targets mapped tRNA genes (bowtie2), suggesting the observed variability was not due to subtle modifications in the tRNA sequence.

Among ncRNA genes showing a significant association with the isolation source, we identified an *E. coli* nucleoid-associated noncoding RNA 4 (naRNA4) gene (U00096.3/900752-900832). naRNA4 genes are encoded in repetitive extragenic palindrome (REP) regions, previously observed at the recombination junctions of lambda phages [33] and described as hotspots for transposition events [34]. 823 unique RAs mapped (BLAST $e<0.01$) to 82 Rfam accessions annotated as distinct naRNA4 genes, concentrating at the ends of the gene (Table S8). The maximum number of times a RA aligned within an individual genome on NCBI was 131 (BLAST $e<0.01$), whereas RAs had up to 140,000 distinct targets, comprising over 500 sequence clusters (Methods). Furthermore, over 90% of targets did not align to any naRNA4 gene (bowtie2), suggesting the observed variability was not due to subtle modifications to the naRNA4 sequence.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DIVE produced similar results for four tRNA genes in the *V. cholera* isolate data (Methods). DIVE also identified 264 RAs mapping (BLAST *e*<0.01) to 26 Rfam accessions annotated as antisense RNA 5 (asRNA5), including four RAs mapping to two *V. cholerae* O1 biovar el Tor asRNA5 genes showing a target diversity significantly associated with available covariates (Table S9). RAs showed no bias in their alignment position. Although no RA mapped within an individual genome on NCBI more than 51 times (BLAST *e*<0.01), some RAs reached 160 target clusters. Furthermore, over 75% of targets did not map to any asRNA5 gene (bowtie2), suggesting the observed target diversity was not due to subtle modifications to the asRNA5 sequence. To our knowledge, no type of antisense RNA has been previously characterized as an insertion site for MGEs.

DIVE identifies preferential insertion sites of transposons near cargo gene hotspots in *V. cholerae*

We next sought to study whether DIVE could identify within-MGE variability. Integrative and conjugative elements (ICEs) confer various properties to their hosts, such as phage and antibiotic resistance, through cargo genes [35]. In *V. cholerae*, SXT ICEs determine phage resistance, and deletion of hotspot five in SXT ICEs leads to susceptibility to ICP1 phage infection [36]. SXT ICE variants share a set of core genes and differ in the cargo genes found in hotspots. Thus, we hypothesized that DIVE could recognize the boundaries of the hotspots.

We ran DIVE on 247 *V. cholerae* isolates and aligned the resulting RAs to six known SXT ICEs. RAs accumulated nearby antibiotic resistance genes (Fig. 2), which define cargo gene hotspots. In particular, RAs clustered near known hotspots in SXT ICE variants ICE*Vch*Ind4, ICE*Vch*Ind5, ICE*Vch*MO10, ICE*Vch*Ban9, and ICE*Vch*Chn956. In addition, we found a unique RA cluster in ICE*Vch*Ban8 in a different location, overlapping a known transposase. Furthermore, when we aligned the corresponding set of targets to a transposable element (TE) database (TnCentral), we found alignment rates of 10% (bowtie2), consistent with the fact that cargo genes mobilize via transposons [37]. Only targets derived from RAs aligning to ICE*Vch*Ban8, a variant lacking annotation information on the database, produced an alignment rate of 0.02%. Nevertheless, in 19% of the targets of RAs aligning to this SXT variant, we observed $[A]^n$ or $[T]^n$ homopolymers, with $n \geq 3$, at the ends of the targets, consistent with the insertion signature of non-LTR transposons. Thus, we hypothesize these RAs point to a putative hotspot in ICE*Vch*Ban8. The percentage of targets among the other SXT variants enriched in A/T homopolymers ranged from 25 to 31%.

Unexplained genetic diversity in *E. coli* and *V. cholerae*

Given the potential of DIVE to discover unannotated genetic hyper-variability, we sought to identify interesting examples that could be the subject of future work. To that end, we filtered the unannotated RAs to keep the most promising examples with at least a median effect size of two and a sample prevalence of at least 5%, resulting in a shortlist of 5 anchors among *E. coli* isolates and 29 among *V. cholerae* isolates (Table S10).

In *E. coli*, RA CCGCCATATCACCTCCGTGATGGTTGC showed the largest median effect (ES=4.22, prevalence=5.28%). The RA produced a single match in at least one hundred NCBI accessions, mainly of *E. coli* strains. In at least three references, the RA lies within the coding sequence of the KdsD gene (arabinose-5-phosphate isomerase), approximately 60 bp from the stop codon (Fig. 3). Only two targets were observed upstream (n=514), but 1,653 different targets (697 clusters) were observed downstream. We found strong agreement upstream of the RA across reads (>99% nucleotide identity) and a significant decay downstream in nucleotide identity across reads, with some positions showing less than 35% agreement (Fig. 3, Supplementary Material). This diversity was not well represented on the reference genomes on NCBI, with only 21 targets (1% of different target sequences) producing hits on the database (BLAST *e<0.5)*. In addition, sequence hyper-variability extended well beyond the range of the target sequence, spanning over 100 nt, and the variation included substitutions and indels altering the protein sequence of KdsD.

In *V. cholerae*, RA CGACTGGTTAAACCACAACCAAATGCA showed the largest median effect size (ES=3.47, prevalence=5.17%). The RA produced a single match in at least one hundred NCBI accessions, all *V. cholerae* strains, mapping to the intergenic region between tRNA-Cys and the artP gene (arginine ABC transporter ATPase), with the former being in the direction of more target diversity (Fig. 3). The RA only had two upstream targets but 396 different targets (95 clusters) downstream. We found strong agreement upstream of the RA across reads (>99% nucleotide identity) and a significant decay downstream in nucleotide identity across reads, with some positions showing less than 45% agreement (Fig. 3, Supplementary Material). This diversity was not well represented on the reference genomes on NCBI, with only 61 targets (15% of different target sequences) producing hits on the database (BLAST *e<0.5)*. In addition, sequence hyper-variability extended well beyond the range of the target sequence, spanning over 100 nt, and the variation included substitutions and indels (Fig. 3).

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

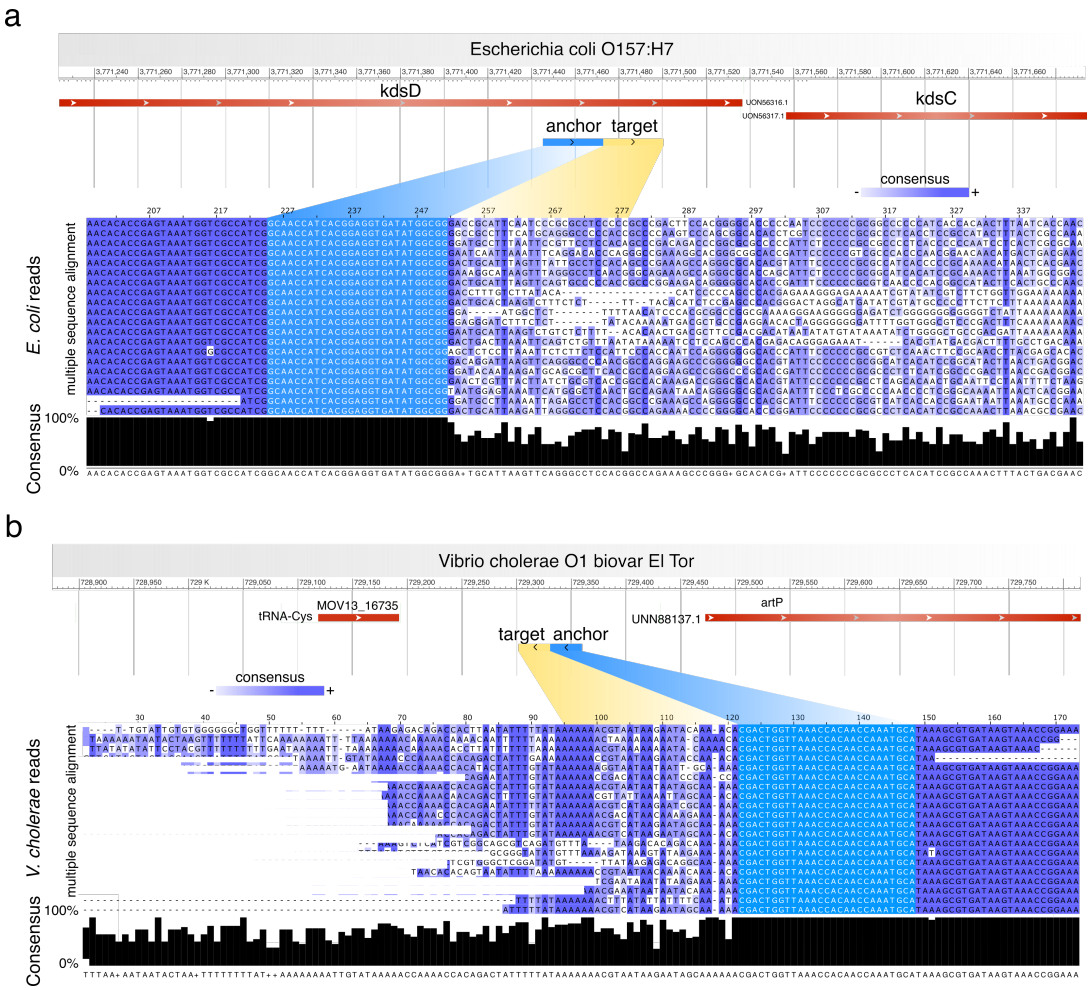


Figure 3: DIVE discovers unexplained hyper-variability in *E. coli* and *V. cholerae*. a. The unannotated anchor CCGCCATATCACCTCCGTGATGGTTGC showed the largest median effect among unannotated anchors with over 5% prevalence in *E. coli* (effect size=4.22, prevalence=5.28%). The anchor produces a unique match in *E. coli* O157 overlapping isomerase Arabinose 5-phosphate isomerase (kdsD). The multiple sequence alignment (CLUSTALW) of a random selection of 20 reads containing the unannotated anchor is shown below. The reads were selected at random from the samples where the anchors were deemed to be significantly associated with target hyper-variability. The sequence consensus (black bars) is approximately 100% upstream of the anchor (light blue), whereas downstream of the anchor the sequences found across the reads diverge significantly and introduce substitutions and indels. **b.** The unannotated anchor CCGCCATATCACCTCCGTGATGGTTGC showed the largest median effect size among unannotated anchors with over 5% prevalence in *V. cholerae* (ES=3.47, prevalence=5.17%). The anchor produced a single match in at least one hundred NCBI accessions, all *V. cholerae* strains, mapping to the intergenic region between tRNA-Cys and the artP gene (arginine ABC transporter ATPase), with the former being in the direction of more target diversity. The multiple sequence alignment (CLUSTALW) of a random selection of 20 reads containing the unannotated anchor is shown below. The reads were selected at random from the samples where the anchors were deemed to be significantly associated with target hyper-variability. The sequence consensus (black bars) is approximately 100% upstream of the anchor (light blue), whereas downstream of the anchor the sequences found across the reads diverge significantly and introduce substitutions and indels.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

DIVE is a new reference-free paradigm for discovering sequences associated with statistically high rates of sequence diversification. We have implemented DIVE in a user-friendly python package, which allows the user to perform analysis by writing a few lines of code. To the best of our knowledge, DIVE is the first algorithm that detects MGEs and DGMs from FASTQ files alone, without relying on a reference genome or assembly during the detection step.

DIVE identified hundreds of technical and biological positive controls, including thousands of mobile genetic elements and hundreds of CRISPR direct repeats. DIVE also identifies large amounts of unexplained genetic diversity (over 23k unannotated representative anchors), pointing to a substantial gap in characterizing microbial sequence space. Here we highlighted the two examples showing the largest median effect size in *E. coli* and *V. cholerae*. This gap is likely growing daily as new and highly active elements will evade assemblers and reference-based attempts to keep pace.

Using DIVE, CRISPR direct repeats can be found very specifically and fast, imposing a simple criterion on DIVE's output. In doing so, DIVE identified a novel direct repeat in a locus that has the hallmarks of a functional CRISPR locus. DIVE also enables the discovery of CRISPR micro-arrays since the algorithm does not rely on repetitiveness patterns, unlike most state-of-the-art algorithms.

We have also shown how DIVE enables quantification of the activity level of a CRISPR array and that using the output, different conditions or time points can be compared. Using DIVE, we discovered several known CRISPR direct repeats in *E. coli* showing up to a 10-fold difference in the target variability depending on the isolation source. We also show that this analysis can be done for other elements, such as MGEs or their corresponding insertion sites, highlighting examples involving lambdoid phages and ncRNA (tRNA, naRNA4, asRNA5).

We further validated DIVE by identifying anchors associated with target hyper-variability mapping to ncRNA genes, which are known magnets of foreign genetic material. DIVE found at least three types of ncRNAs associated with hyper-variability and, in some instances, correlating with available covariates such as time-points or environment, providing further evidence of functional relevance. In addition to tRNA and naRNA4, we hypothesize *Vibrio* asRNA5 could constitute another ncRNA class prone to accept MGEs.

Studying cargo genes within ICEs is paramount in addressing the growing antibiotic resistance among bacteria. Using DIVE, we identified putative insertion sites for transposable elements in SXT-like ICEs in *V. cholera* near loci containing antibiotic resistance genes. This discovery shows the potential of DIVE for detecting not only insertions of MGEs into a genome but also for studying genetic variability within elements. Importantly, these sites are of particular relevance since they commonly introduce cargo genes conferring the host with antibiotic resistance and other phenotypes.

Overall, DIVE offers a new reference-free paradigm for studying genetic bacterial, archaeal, and viral evolution, enabling the discovery and quantification of the activity level of MGEs and DGMs. More importantly, given the generality of the principle used, DIVE also enables the discovery of completely novel biological mechanisms. In future work, we plan to reduce the memory requirements of DIVE and to introduce online hyper-variability calling to improve the efficiency of the method. The proposed paradigm could have other applications, such as immune receptor recombination and extensive alternative splicing.

Methods

DIVE configuration. Here we defined the k-mer size k to be 27, striking a balance between the memory requirements and the specificity of the sequence. DIVE also allows the user to introduce a gap in between the anchor and the targets of an arbitrary distance g, which here we set at $g = 0$. We defined the minimum and the maximum number of targets per anchor to be $N_{\text{min}} = 25$ and $N_{\text{max}} = 75$, respectively. Finally, to strike a balance between power and computational resources, we decided to limit the number of FASTQ records processed per sample to 2.5M.

Clustering targets on the fly. Upon observing a target k-mer, DIVE checks whether the k-mer produces a Jaccard similarity (JS) larger than a given threshold (0.2) with any of the previously observed targets. If so, the counter of that key is increased by one, and, with a probability of 50%, the key is replaced by the newly observed target sequence. If no such target sequence exists in the dictionary, a new key is created, and its counter is initialized at one. To calculate the JS between two targets of length $k = 27$, DIVE splits each one into smaller k-mers of length seven and computes the JS similarity between the two using

$$\text{JS}\left(\mathcal{X}_1, \mathcal{X}_2\right) = \frac{\mathcal{X}_1 \cap \mathcal{X}_2}{\mathcal{X}_1 \cup \mathcal{X}_2} \tag{1}$$

where $\mathcal{X}_i$ is the set of 7-mers in the i-th target sequence.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Minimum and maximum N. For each anchor, DIVE requires a minimum number of targets observed N_{min} to proceed with the downstream statistical analysis. In addition, since it might not be necessary to keep all the targets observed exhaustively, DIVE also imposes a maximum number of targets N_{max} to be observed for each anchor. This allows the algorithm to skip anchors for which it has already observed N_{max} targets in each direction and move faster along the FASTQ file. Here we set these parameters to $N_{\text{min}} = 25$ and $N_{\text{max}} = 75$.

Minimum sample size prediction. To manage the memory burden and make the algorithm more efficient, we impose a minimum number N_{min} of targets to be observed by the end of the FASTQ file. We let the total number of FASTQ records in the file be L and the number of observed records so far be l. Then, letting $p_{\text{min}=N_{\text{min}}/L}$ be the probability that we will observe a target sequence for any given anchor in a single read at least once, we compute the probability that, after l records observed, we have observed zero instances using

$$\mathrm{P}[N = 0; l] = (1 - p_{\text{min}})^l \tag{2}$$

When this probability becomes smaller than 0.01 this implies that the rate we are observing targets for the anchor is too slow for us to observe N_{min} by the end with 99% probability. Thus, DIVE does not accept new anchors into the anchor dictionary DA past this point. This condition allows us to discard new anchors that appear late in the FASTQ file, which will likely not produce enough data to test for hyper-variability, reducing an unnecessary burden for dictionary $\mathcal{D}_A$. Past the point where DIVE does not accept new anchors, DIVE also anticipates the number of targets that will be observed for each anchor remaining in the dictionary to decide whether a given anchor should remain in the dictionary $\mathcal{D}_A$. More precisely, the oracle used by DIVE computes the probability that we will observe at least N_{min} targets by the end of the FASTQ file. Letting L be the total number of FASTQ reads, x the number of targets observed, and l the number of FASTQ reads processed so far, we can compute

$$\mathrm{P}\left[N \geq N_{\text{min}}; x, l\right] = \sum_{n=N_{\text{min}}}^{L} \left(\begin{array}{c} L \\ n \end{array} \right) \left(\frac{x}{l}\right)^n \left(1 - \left(\frac{x}{l}\right)\right)^{L-n} \tag{3}$$

where *x/l* is an empirical estimate of the probability of observing a target sequence for the given anchor at least once in a single read. However, we use the normal approximation to the binomial distribution for efficiency to compute this probability. The anchor is conserved in the dictionary $\mathcal{D}_A$ if this probability is larger than 50%. Otherwise, the anchor is removed to reduce the size of the dictionary and speed up the algorithm.

Target sequence hyper-variability. We let X_n be a binary random variable indicating whether we created a new key in the target sequence dictionary upon observing the n-th target sequence for a given anchor and direction. Let C_n be the total number of keys target sequence clusters (keys) formed after observing the n-th target sequence. Then, assuming that X_n is independent and distributed according to a Bernoulli distribution with success probability p_n, we have that $C_N \sim \text{PoissBin}(p_1, \ldots, p_N)$. Upon observing a set of trajectories $\{C_1, C_2, \ldots, C_M\}$, the success probabilities $\{p_n\}$ can be estimated from the data by simply computing

$$\hat{p}_n = \frac{\sum_{m=1}^{M} x_{m,n}}{\sum_{m=1}^{M} 1\left\{N_m \geq n\right\}} \tag{4}$$

This follows from the fact that we can assume that most of the anchors will not be nearby hyper-variable regions. Nevertheless, the following test will be more conservative if we happen to include anchors adjacent to hyper-variable regions by chance. For a given $C_N = c$, we can compute the probability under the null hypothesis that the observed value c can be explained by the background biological (e.g., point mutations) and technical variability (e.g., sequencing errors) using

$$P_{H_0}\left[C_N \geq c\right] = \sum_{k=C_N}^{N} \frac{1}{N+1} \sum_{l=0}^{N} R^{-lk} \prod_{n=1}^{N} \left(1 + \left(R^l - 1\right) \hat{p}_n\right) \tag{5}$$

The resulting p-values are corrected using the Benjamini-Hochberg (BH) correction, resulting in a list of q-values that allow us to perform false discovery rate (FDR) control.

Efficient Benjamini-Hochberg correction. To correct for multiple hypothesis testing, we DIVE implements a memory-efficient version of the Benjamini-Hochberg (BH) correction. DIVE records the total number of statistical tests performed m and drops non-significant cases ($p > 0.1$). Then, it adjusts the remaining m' p-values by using the

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

following procedure. First, it ranks the remaining m' p-values $\{p_i\}$ in ascending order $\{p^{(1)}, p^{(2)}, \ldots, p^{(m')}\}$. Then, it computes the adjusted p-value using $q^{(i)} = p^{(i)}m/i$, and it reports the anchors for which $q < 0.1$. Note that p-values computed for overlapping k-mers will be positively correlated. Nevertheless, the BH correction can still control the FDR under that form of dependence [38].

Clustering of anchors. Once the BH correction is applied to the computed p-values, DIVE uses DBSCAN (Levenshtein edit distance) to cluster the anchor sequences to avoid redundancies in the output. This is done using the function DBSCAN from the python package sklearn with parameters `eps=2` and `min_samples=1`. For each anchor cluster, a representative anchor is chosen for each direction (upstream and downstream) by picking the anchor that maximizes the effect size.

Effect size computation. The effect size is quantified using the log-fold change between the observed number of clusters c_N and the expected number of clusters for a given number of observed targets, as quantified by N, and it is given by

$$\mathrm{ES}\left(c_N\right) = \log_2 \frac{c_N}{E\left[C_N \mid N\right]} \tag{6}$$

This quantity can be used to rank the results in terms of effect size and allows the user to filter out positives with small effect size and thus little biological interest.

Annotation of DIVE's output. To validate our results, we used blastn-short 2.11.0 [39] to align the detected k-mers to a set of sequence databases stored in FASTA format comprised of CRISPR direct repeats (CRISPR-Cas++ [30]), transposable elements (Dfam [40]; TnCentral [41]), mobile genetic elements (ACLAME [42]), internal transcribed spacers 1 and 2 (ITSoneDB [43], ITS2 [44]), integrative and conjugative elements (ICEberg [45]), and RNA sequence families of structural RNAs (Rfam [46]). Furthermore, we cross-referenced our results with Illumina adapter (obtained from TrimGalore) and UniVec sequences (NCBI) to remove technical artifacts. We classified each k-mer as unannotated if no BLAST produced $e \leq 0.25$, questionable if $0.05 \leq e < 0.25$, and annotated if the lowest $e \leq 0.05$ with the corresponding annotation. Here we chose this comprehensive set of annotation files. Nevertheless, DIVE can take an arbitrary number of annotation files as input, and thus this step can be adjusted depending on the application.

Removal of low entropy anchors. To prevent technical artifacts, we remove anchors with low entropy for downstream analysis. To that end, we compute an empirical estimate of the entropy based on 5-mers using

$$H = \sum_{x \in \mathcal{S}_5} p(x) \log p(x) \tag{7}$$

where $p(x)$ is estimated from the anchor by partitioning it into overlapping 5-mers and where we take the convention that $0 \log 0 = 0$. We require $H > 3$ for the anchor to be considered in the downstream analysis.

Enrichment analysis of positive controls. A total of 100,000 27-mers were chosen at random among the entire *E. coli* dataset, allowing for repetition. We used BLAST to align this set of random anchors to the set of annotation files we used during our analysis. Then, we deemed anchors as unannotated and annotated if their BLAST e-values were $e > 0.25$ and $e < 0.01$, respectively. Then, we used the function fisher_exact python's scipy package to perform a two-sided Fisher's exact test comparing the proportions of annotated to unannotated between this set of random anchors and the set of DIVE anchors.

Binomial regression. To find associations between the target variability and a set of covariates given by $\mathbf{x}$, we use a binomial regression model such that $C \sim \mathrm{Binomial}(N; p(\mathbf{x}))$, such that

$$p(\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{x}}} \tag{8}$$

For *E. coli* isolates, we used the isolation source as a covariate, whereas for *V. cholerae* isolates we used the environment, date, country, as well as first-order interactions.

A posteriori clustering of targets. To cluster the targets across samples *a posteriori*, we used a greedy clustering approach. We rank the targets based on their total count across the entire dataset. Starting from the most prevalent target, we recruit all targets for which the JS>0.2 and eliminate these from the list. We proceed with the most prevalent target sequence remaining in the list the same way, and we repeat this procedure until no sequences are left. The number of times we repeat this procedure defines the number of across-sample target sequence clusters.

Data availability. The data sets analyzed in this paper are publicly available and published. The *E. coli* isolate sequencing data used was generated by the GenomeTrakr Project, US Food and Drug Administration (PRJNA230969).

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We used 514 samples encompassing 62 different isolation sources (Table S11). The *V. cholerae* sequencing data was downloaded from SRA (PRJNA723557). We used all the 247 *V. cholerae* isolates sequenced as part of this project (Table S12). The rotavirus metagenomics sequencing data was downloaded from SRA (PRJNA729919). We used 102 in our analysis (Table S13). Finally, the *N. gonorrhoeae* data was downloaded from SRA (PRJNA298332), from which we used 200 samples in our analysis (Table S14).

Code availability. The code used in this work is available at https://github.com/jordiabante/biodive.

Acknowledgments. We would like to thank Ruth Timme, David Lipman, Andrew Fire, Kaitlin Chuang, Roozbeh Dehghannasiri, Elisabeth Meyer, and Ivan Zheludev for their comments and suggestions. We would also like to thank Kaitlin Chuang for her help processing part of the data used in this paper. J.A. is partially supported by the Stanford Center for Computational, Evolutionary and Human Genomics Postdoctoral Fellowship. J.S. is supported by the Stanford University Discovery Innovation Award, National Institute of General Medical Sciences grant nos. R35 GM139517 and the National Science Federation Faculty Early Career Development Program Award no. MCB1552196.

References

[1] Hans-Curt Flemming and Stefan Wuertz. Bacteria and archaea on earth and their abundance in biofilms. *Nature Reviews Microbiology*, 17(4):247–260, 2 2019. ISSN 1740-1526. doi:10.1038/s41579-019-0158-9. URL `http://dx.doi.org/10.1038/s41579-019-0158-9`.

[2] Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016.

[3] Novella IS, Domingo E, and Holland JJ. Rapid viral quasispecies evolution: implications for vaccine and drug strategies. *Molecular medicine today*, 1:17607887, Aug 1995. ISSN 1357-4310. doi:10.1016/s1357-4310(95)91551-6. URL `https://dx.doi.org/10.1016/s1357-4310(95)91551-6`.

[4] Ruiting Lan and Peter R. Reeves. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends in Microbiology*, 8(9):396–401, 9 2000. ISSN 0966-842X. doi:10.1016/s0966-842x(00)01791-1. URL `http://dx.doi.org/10.1016/s0966-842x(00)01791-1`.

[5] Siobain Duffy. Why are RNA virus mutation rates so damn high? *PLOS Biology*, 16(8):e3000003, 8 2018. ISSN 1545-7885. doi:10.1371/journal.pbio.3000003. URL `http://dx.doi.org/10.1371/journal.pbio.3000003`.

[6] Chitra Dutta and Munmun Sarkar. Horizontal gene transfer and bacterial diversity. In *Encyclopedia of Metagenomics*, pages 251–257. Springer US, 2015. doi:10.1007/978-1-4899-7478-5_225. URL `http://dx.doi.org/10.1007/978-1-4899-7478-5_225`.

[7] W Ford Doolittle. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, 14(8):307–311, 12 1998. ISSN 0168-9525. doi:10.1016/s0168-9525(98)01494-2. URL `http://dx.doi.org/10.1016/s0168-9525(98)01494-2`.

[8] Yuri I. Wolf, L. Aravind, and Eugene V. Koonin. Rickettsiae and chlamydiae: evidence of horizontal gene transfer and gene exchange. *Trends in Genetics*, 15(5):173–175, 5 1999. ISSN 0168-9525. doi:10.1016/s0168-9525(99)01704-7. URL `http://dx.doi.org/10.1016/s0168-9525(99)01704-7`.

[9] Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, and Fraser CM. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature*, 399:10360571, May 1999. ISSN 0028-0836. doi:10.1038/20601. URL `https://dx.doi.org/10.1038/20601`.

[10] John Penders, Ellen E. Stobberingh, Paul H. M. Savelkoul, and Petra F. G. Wolffs. The human microbiome as a reservoir of antimicrobial resistance. *Frontiers in Microbiology*, 4, 2013. ISSN 1664-302X. doi:10.3389/fmicb.2013.00087. URL `http://dx.doi.org/10.3389/fmicb.2013.00087`.

[11] Rachel A. F. Wozniak, Derrick E. Fouts, Matteo Spagnoletti, Mauro M. Colombo, Daniela Ceccarelli, Geneviève Garriss, Christine Déry, Vincent Burrus, and Matthew K. Waldor. Comparative ICE genomics: Insights into the evolution of the SXT/R391 family of ices. *PLoS Genetics*, 5(12):e1000786, 12 2009. ISSN 1553-7404. doi:10.1371/journal.pgen.1000786. URL `http://dx.doi.org/10.1371/journal.pgen.1000786`.

[12] Xinyue Li, Yu Du, Pengcheng Du, Hang Dai, Yujie Fang, Zhenpeng Li, Na Lv, Baoli Zhu, Biao Kan, and Duochun Wang. SXT/R391 integrative and conjugative elements in proteus species reveal abundant genetic diversity and multidrug resistance. *Scientific Reports*, 6(1), 11 2016. ISSN 2045-2322. doi:10.1038/srep37372. URL `http://dx.doi.org/10.1038/srep37372`.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[13] Smith RA, M'ikanatha NM, and Read AF. Antibiotic resistance: a primer and call to action. *Health communication*, 30:25121990, 2015. ISSN 1041-0236. doi:10.1080/10410236.2014.943634. URL https://dx.doi.org/10.1080/10410236.2014.943634.

[14] Aaron E. Darling, Bob Mau, and Nicole T. Perna. progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6):e11147, 6 2010. ISSN 1932-6203. doi:10.1371/journal.pone.0011147. URL http://dx.doi.org/10.1371/journal.pone.0011147.

[15] Choumouss Kamoun, Thibaut Payen, Aurélie Hua-Van, and Jonathan Filée. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics*, 14(1), 10 2013. ISSN 1471-2164. doi:10.1186/1471-2164-14-700. URL http://dx.doi.org/10.1186/1471-2164-14-700.

[16] Treepong P, Guyeux C, Meunier A, Couchoud C, Hocquet D, and Valot B. panisa: ab initio detection of insertion sequences in bacterial genomes from short read sequence data. *Bioinformatics (Oxford, England)*, 34:29931098, Nov 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty479. URL https://dx.doi.org/10.1093/bioinformatics/bty479.

[17] Durrant MG, Li MM, Siranosian BA, Montgomery SB, and Bhatt AS. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell host & microbe*, 27:31862382, Jan 2020. ISSN 1931-3128. doi:10.1016/j.chom.2019.10.022. URL https://dx.doi.org/10.1016/j.chom.2019.10.022.

[18] Izaak Coleman and Tal Korem. Embracing metagenomic complexity with a genome-free approach. *mSystems*, 6(4), 8 2021. ISSN 2379-5077. doi:10.1128/msystems.00816-21. URL http://dx.doi.org/10.1128/msystems.00816-21.

[19] Marschalek R, Brechner T, Amon-Böhm E, and Dingermann T. Transfer RNA genes: landmarks for integration of mobile genetic elements in dictyostelium discoideum. *Science (New York, N.Y.)*, 244:2567533, Jun 1989. ISSN 0036-8075. doi:10.1126/science.2567533. URL https://dx.doi.org/10.1126/science.2567533.

[20] R. Marquet, C. Isel, C. Ehresmann, and B. Ehresmann. trnas as primer of reverse transcriptases. *Biochimie*, 77 (1-2):113–124, 1 1995. ISSN 0300-9084. doi:10.1016/0300-9084(96)88114-4. URL http://dx.doi.org/10.1016/0300-9084(96)88114-4.

[21] Jin M. Kim, Swathi Vanguri, Jef D. Boeke, Abram Gabriel, and Daniel F. Voytas. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete saccharomyces cerevisiae genome sequence. *Genome Research*, 8(5):464–478, 5 1998. ISSN 1088-9051. doi:10.1101/gr.8.5.464. URL http://dx.doi.org/10.1101/gr.8.5.464.

[22] Thomas H Eickbush and Danna G Eickbush. Finely orchestrated movements: Evolution of the ribosomal RNA genes. *Genetics*, 175(2):477–485, 2 2007. ISSN 1943-2631. doi:10.1534/genetics.107.071399. URL http://dx.doi.org/10.1534/genetics.107.071399.

[23] Rotem Sorek, Victor Kunin, and Philip Hugenholtz. CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology*, 6(3):181–186, 3 2008. ISSN 1740-1526. doi:10.1038/nrmicro1793. URL http://dx.doi.org/10.1038/nrmicro1793.

[24] I. Grissa, G. Vergnaud, and C. Pourcel. Crisprfinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, 35(Web Server):W52–W57, 5 2007. ISSN 0305-1048. doi:10.1093/nar/gkm360. URL http://dx.doi.org/10.1093/nar/gkm360.

[25] Robert C Edgar. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, 8(1), 1 2007. ISSN 1471-2105. doi:10.1186/1471-2105-8-18. URL http://dx.doi.org/10.1186/1471-2105-8-18.

[26] Mina Rho, Yu-Wei Wu, Haixu Tang, Thomas G. Doak, and Yuzhen Ye. Diverse crisprs evolving in human microbiomes. *PLoS Genetics*, 8(6):e1002441, 6 2012. ISSN 1553-7404. doi:10.1371/journal.pgen.1002441. URL http://dx.doi.org/10.1371/journal.pgen.1002441.

[27] Connor T. Skennerton, Michael Imelfort, and Gene W. Tyson. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Research*, 41(10):e105–e105, 3 2013. ISSN 1362-4962. doi:10.1093/nar/gkt183. URL http://dx.doi.org/10.1093/nar/gkt183.

[28] Yuzhen Ye. Identification of diversity-generating retroelements in human microbiomes. *International Journal of Molecular Sciences*, 15(8):14234–14246, 8 2014. ISSN 1422-0067. doi:10.3390/ijms150814234. URL http://dx.doi.org/10.3390/ijms150814234.

[29] Moller AG and Liang C. Metacrast: reference-guided extraction of CRISPR spacers from unassembled metagenomes. *PeerJ*, 5:28894651, 2017. doi:10.7717/peerj.3788. URL https://dx.doi.org/10.7717/peerj.3788.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[30] David Couvin, Aude Bernheim, Claire Toffano-Nioche, Marie Touchon, Juraj Michalik, Bertrand Néron, Eduardo P C Rocha, Gilles Vergnaud, Daniel Gautheret, and Christine Pourcel. Crisprcasfinder, an update of crisrfinder, includes a portable version, enhanced performance and integrates search for cas proteins. *Nucleic Acids Research*, 46(W1):W246–W251, 5 2018. ISSN 0305-1048. doi:10.1093/nar/gky425. URL http://dx.doi.org/10.1093/nar/gky425.

[31] Michael J. Tisza and Christopher B. Buck. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences*, 118 (23), 6 2021. ISSN 0027-8424. doi:10.1073/pnas.2023202118. URL http://dx.doi.org/10.1073/pnas.2023202118.

[32] German Martinez. trnas as primers and inhibitors of retrotransposons. *Mobile Genetic Elements*, 7(5):1–6, 9 2017. ISSN 2159-256X. doi:10.1080/2159256x.2017.1393490. URL http://dx.doi.org/10.1080/2159256x.2017.1393490.

[33] Michiyo Kumagai and Hideo Ikeda. Molecular analysis of the recombination junctions of λ bio transducing phases. *Molecular and General Genetics MGG*, 230(1-2):60–64, 11 1991. ISSN 0026-8925. doi:10.1007/bf00290651. URL http://dx.doi.org/10.1007/bf00290651.

[34] Raquel Tobes and Eduardo Pareja. Bacterial repetitive extragenic palindromic sequences are DNA targets for insertion sequence elements. *BMC Genomics*, 7(1), 3 2006. ISSN 1471-2164. doi:10.1186/1471-2164-7-62. URL http://dx.doi.org/10.1186/1471-2164-7-62.

[35] Rachel A. F. Wozniak and Matthew K. Waldor. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology*, 8(8):552–563, 7 2010. ISSN 1740-1526. doi:10.1038/nrmicro2382. URL http://dx.doi.org/10.1038/nrmicro2382.

[36] LeGault KN, Hays SG, Angermeyer A, McKitterick AC, Johura FT, Sultana M, Ahmed T, Alam M, and Seed KD. Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science (New York, N.Y.)*, 373: 34326207, Jul 2021. ISSN 0036-8075. doi:10.1126/science.abg2166. URL https://dx.doi.org/10.1126/science.abg2166.

[37] Sean Benler, Guilhem Faure, Han-Altae Tran, Sergey Shmakov, Feng Zheng, and Eugene Koonin. Cargo genes of tn7-like transposons comprise an enormous diversity of defense systems, mobile genetic elements and antibiotic resistance genes. *Mbio*, 8 2021. doi:10.1101/2021.08.23.457393. URL http://dx.doi.org/10.1101/2021.08.23.457393.

[38] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), 8 2001. ISSN 0090-5364. doi:10.1214/aos/1013699998. URL http://dx.doi.org/10.1214/aos/1013699998.

[39] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), 12 2009. ISSN 1471-2105. doi:10.1186/1471-2105-10-421. URL http://dx.doi.org/10.1186/1471-2105-10-421.

[40] Jessica Storer, Robert Hubley, Jeb Rosen, Travis J. Wheeler, and Arian F. Smit. The dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, 12(1), 1 2021. ISSN 1759-8753. doi:10.1186/s13100-020-00230-y. URL http://dx.doi.org/10.1186/s13100-020-00230-y.

[41] Karen Ross, Alessandro M Varani, Erik Snesrud, Hongzhan Huang, Danillo Oliveira Alvarenga, Jian Zhang, Cathy Wu, Patrick McGann, and Mick Chandler. Tncentral: a prokaryotic transposable element database and web portal for transposon analysis. *MBio*, 12(5):e02060–21, 2021.

[42] R. Leplae. ACLAME: A classification of mobile genetic elements. *Nucleic Acids Research*, 32(90001):45D–49, 1 2004. ISSN 1362-4962. doi:10.1093/nar/gkh084. URL http://dx.doi.org/10.1093/nar/gkh084.

[43] Monica Santamaria, Bruno Fosso, Flavio Licciulli, Bachir Balech, Ilaria Larini, Giorgio Grillo, Giorgio De Caro, Sabino Liuni, and Graziano Pesole. Itsonedb: a comprehensive collection of eukaryotic ribosomal RNA internal transcribed spacer 1 (ITS1) sequences. *Nucleic Acids Research*, 46(D1):D127–D132, 9 2017. ISSN 0305-1048. doi:10.1093/nar/gkx855. URL http://dx.doi.org/10.1093/nar/gkx855.

[44] Selig C, Wolf M, Müller T, Dandekar T, and Schultz J. The ITS2 database II: homology modelling RNA structure for molecular systematics. *Nucleic acids research*, 36:17933769, Jan 2008. ISSN 0305-1048. doi:10.1093/nar/gkm827. URL https://dx.doi.org/10.1093/nar/gkm827.

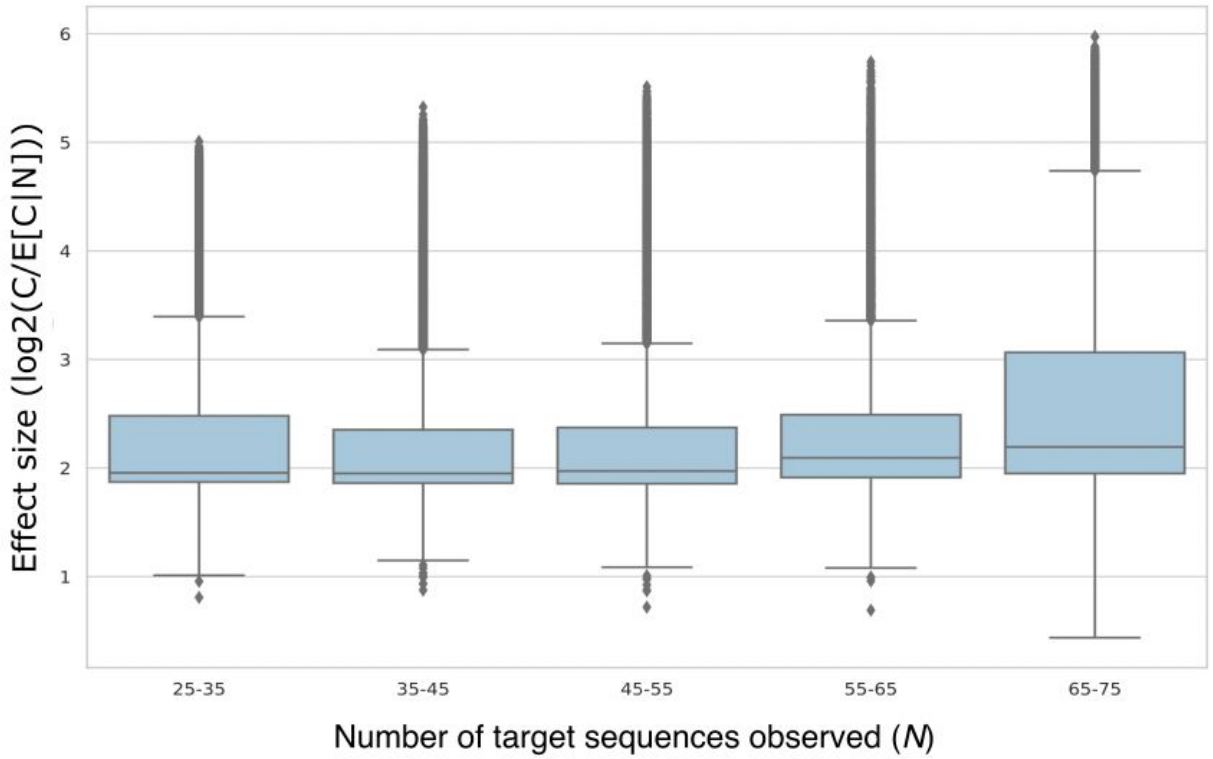
[45] Dexi Bi, Zhen Xu, Ewan M. Harrison, Cui Tai, Yiqing Wei, Xinyi He, Shiru Jia, Zixin Deng, Kumar Rajakumar, and Hong-Yu Ou. Iceberg: a web-based resource for integrative and conjugative elements found in bacteria. *Nucleic Acids Research*, 40(D1):D621–D626, 10 2011. ISSN 1362-4962. doi:10.1093/nar/gkr846. URL http://dx.doi.org/10.1093/nar/gkr846.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[46] Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, Elena Rivas, Sean R Eddy, Robert D Finn, Alex Bateman, and Anton I Petrov. Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic Acids Research*, 49(D1):D192–D200, 11 2020. ISSN 0305-1048. doi:10.1093/nar/gkaa1047. URL `http://dx.doi.org/10.1093/nar/gkaa1047`.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a.



b.

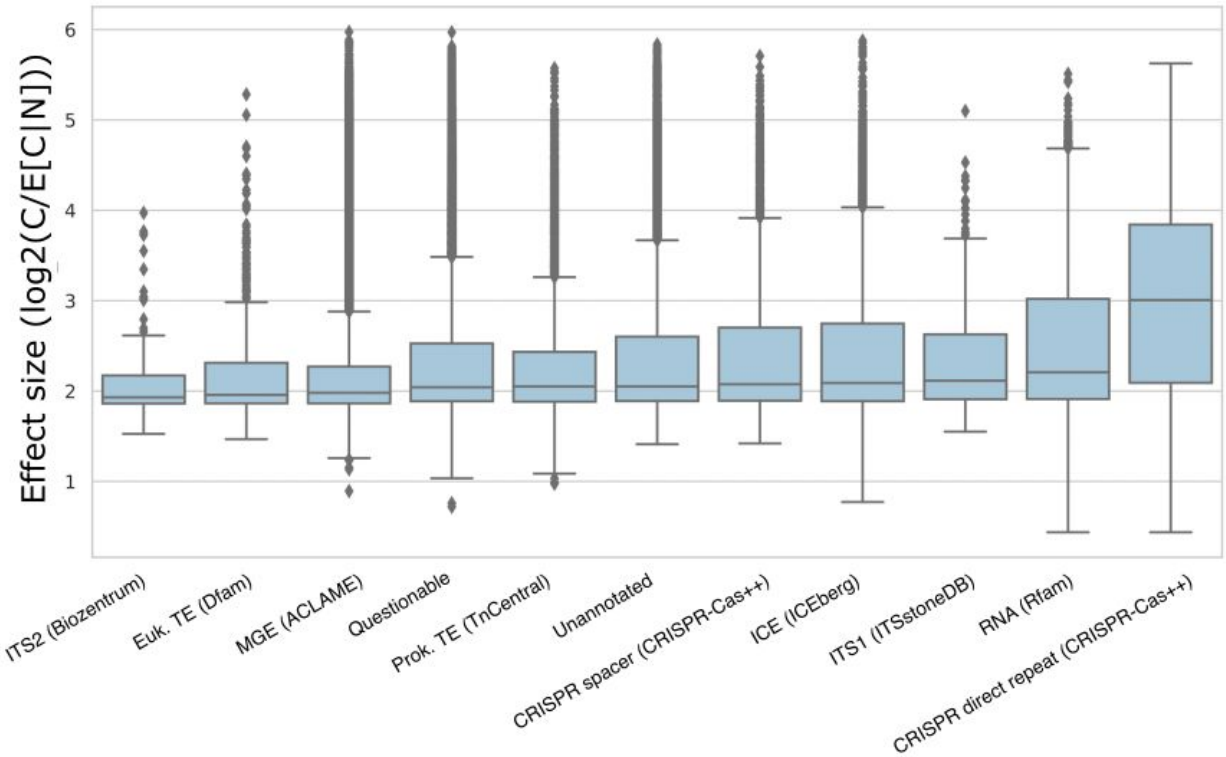
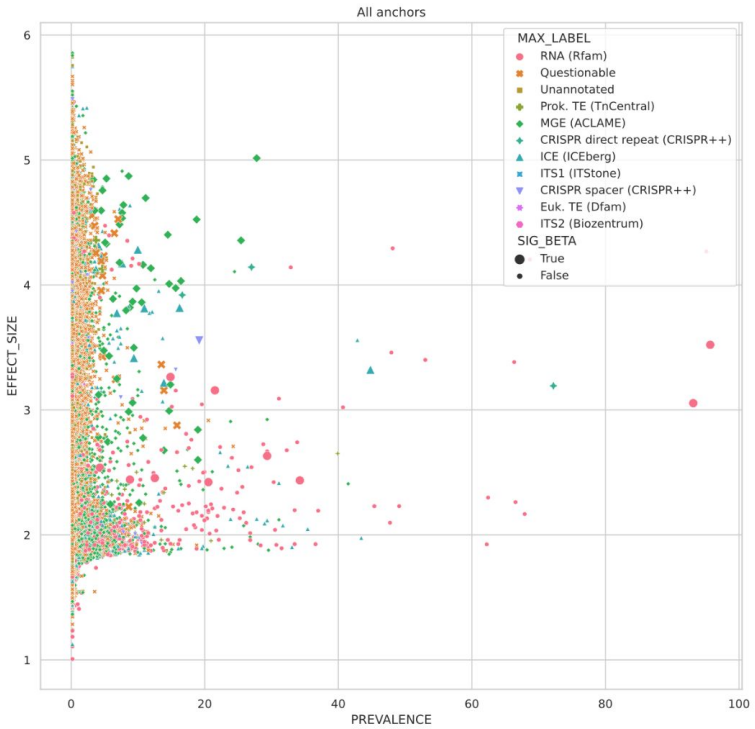


Figure S1

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a.



b.

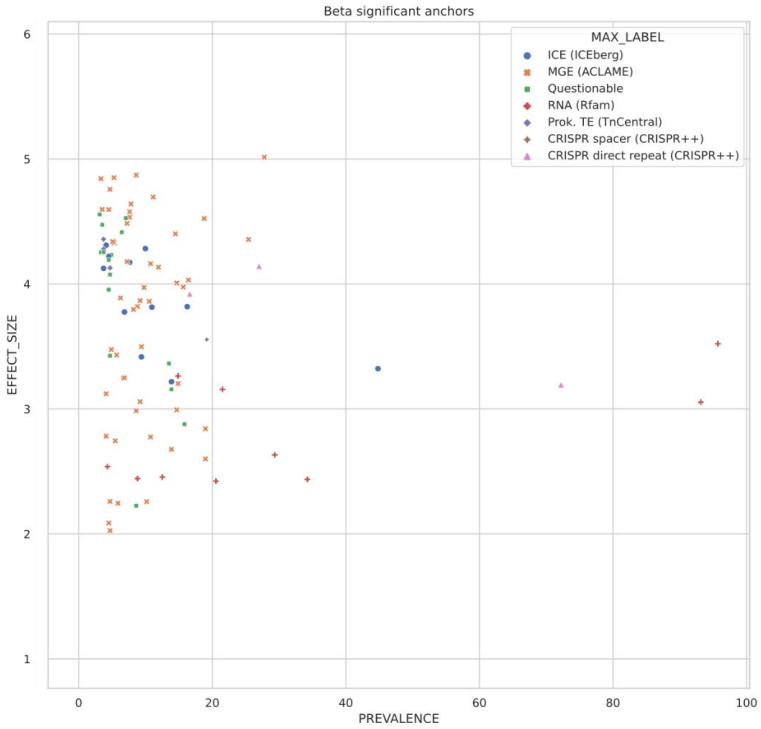
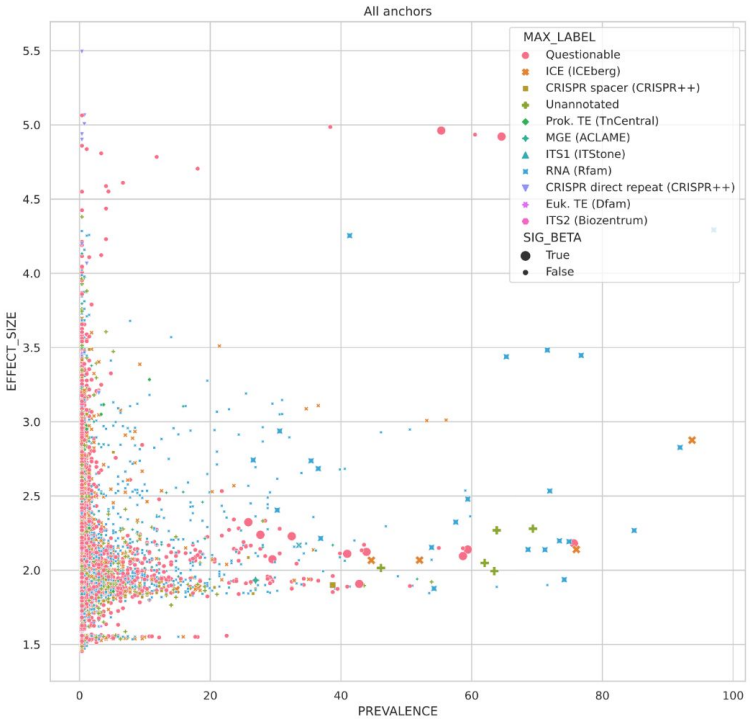


Figure S2

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a.



b.

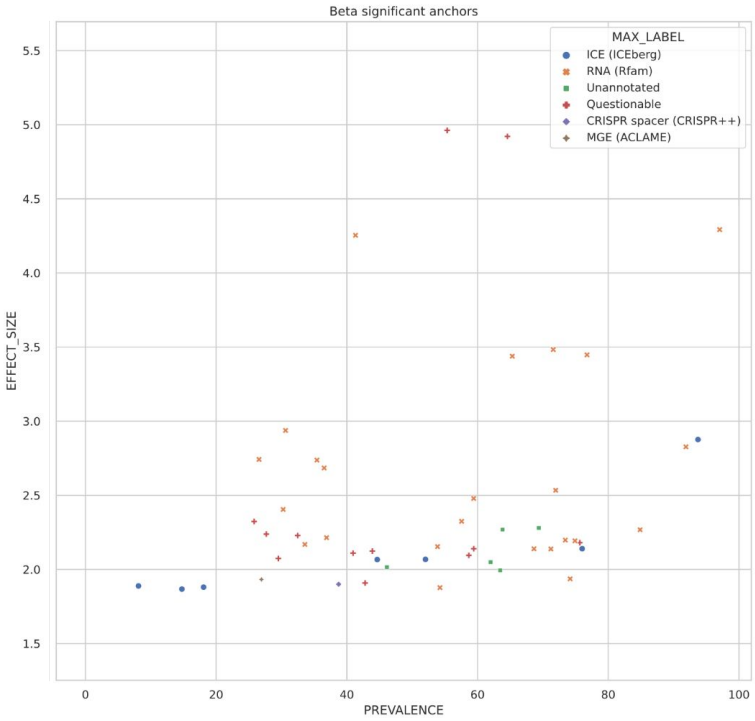


Figure S3

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

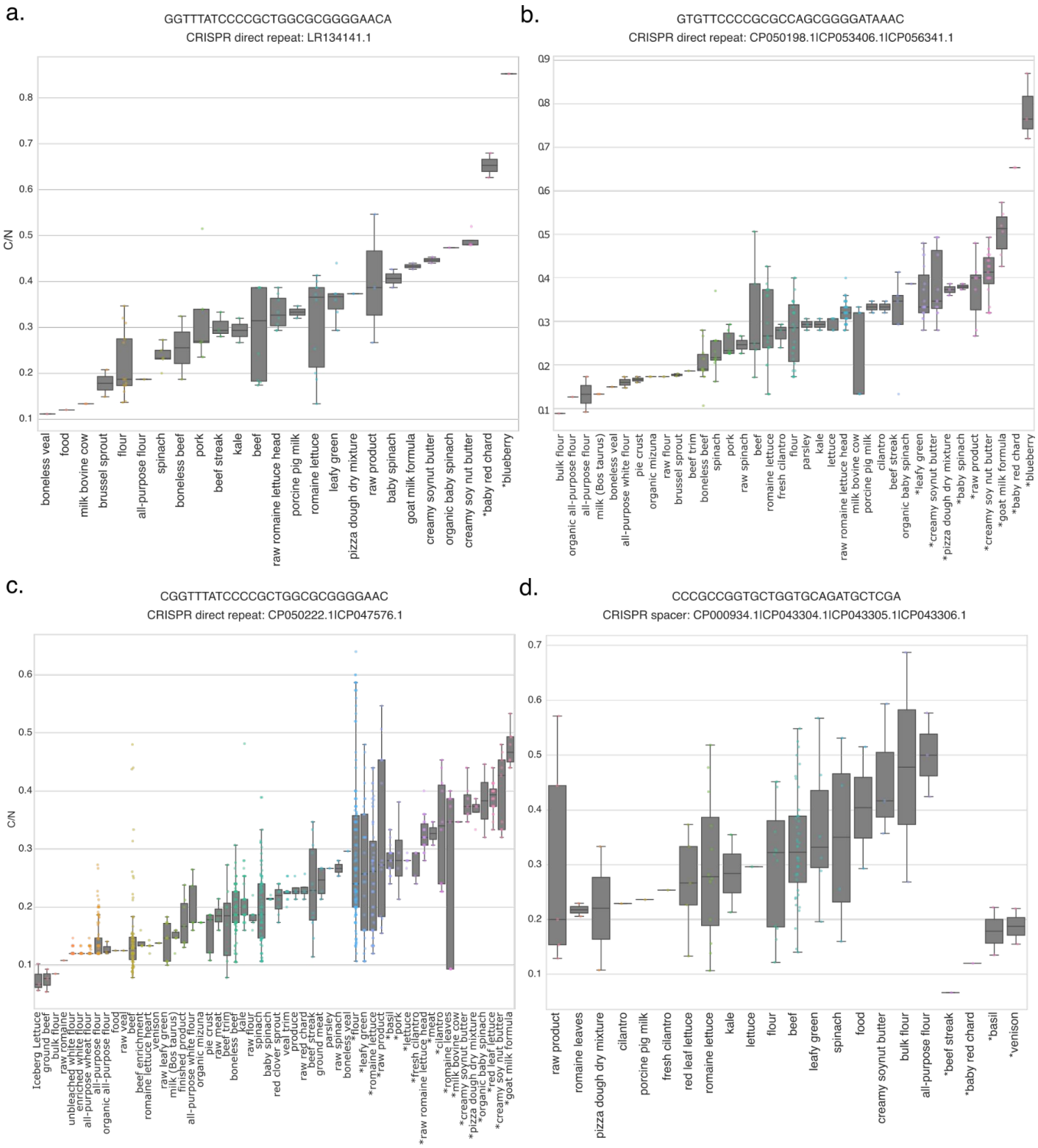


Figure S4

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

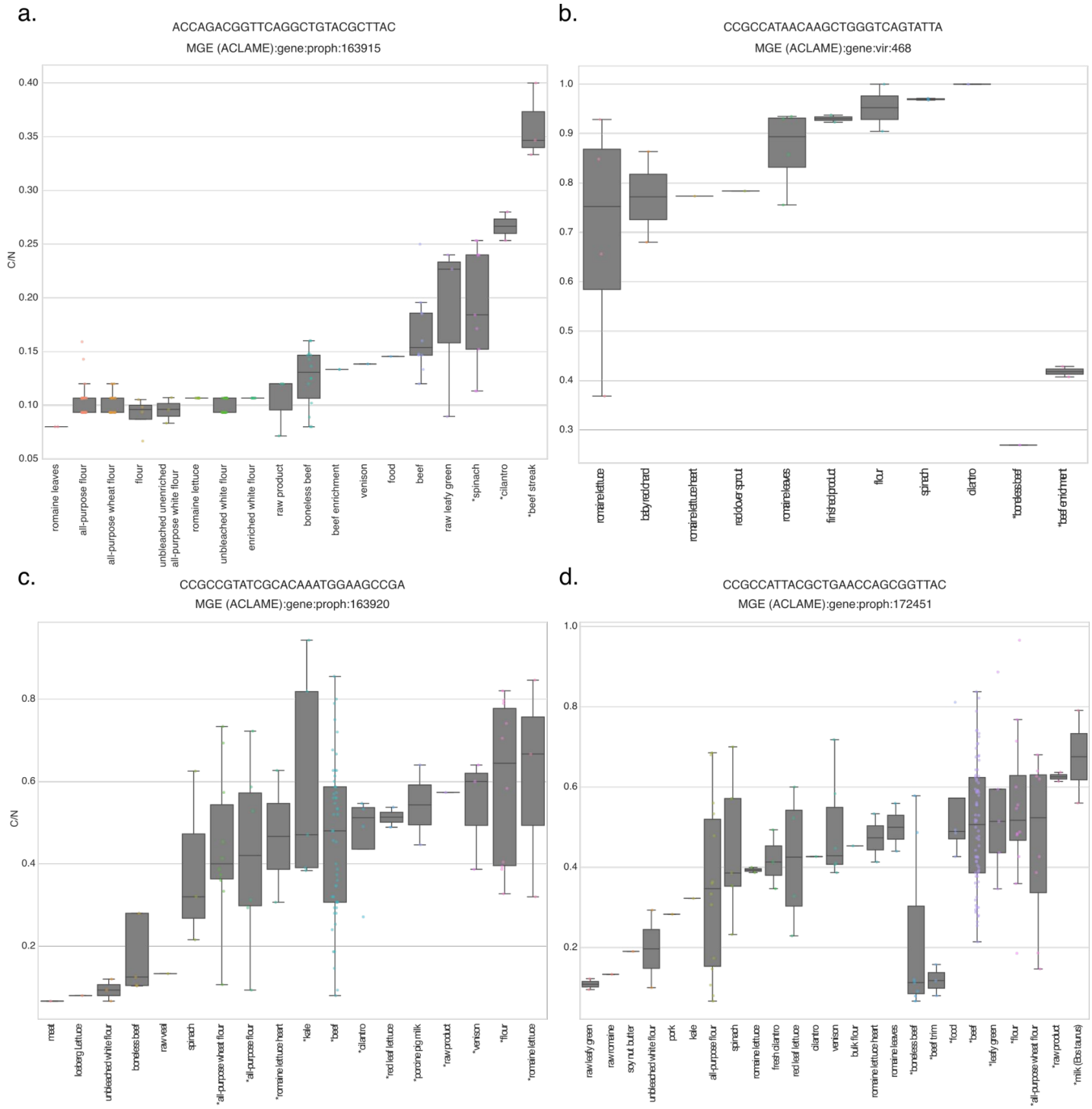


Figure S5

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.13.495703; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

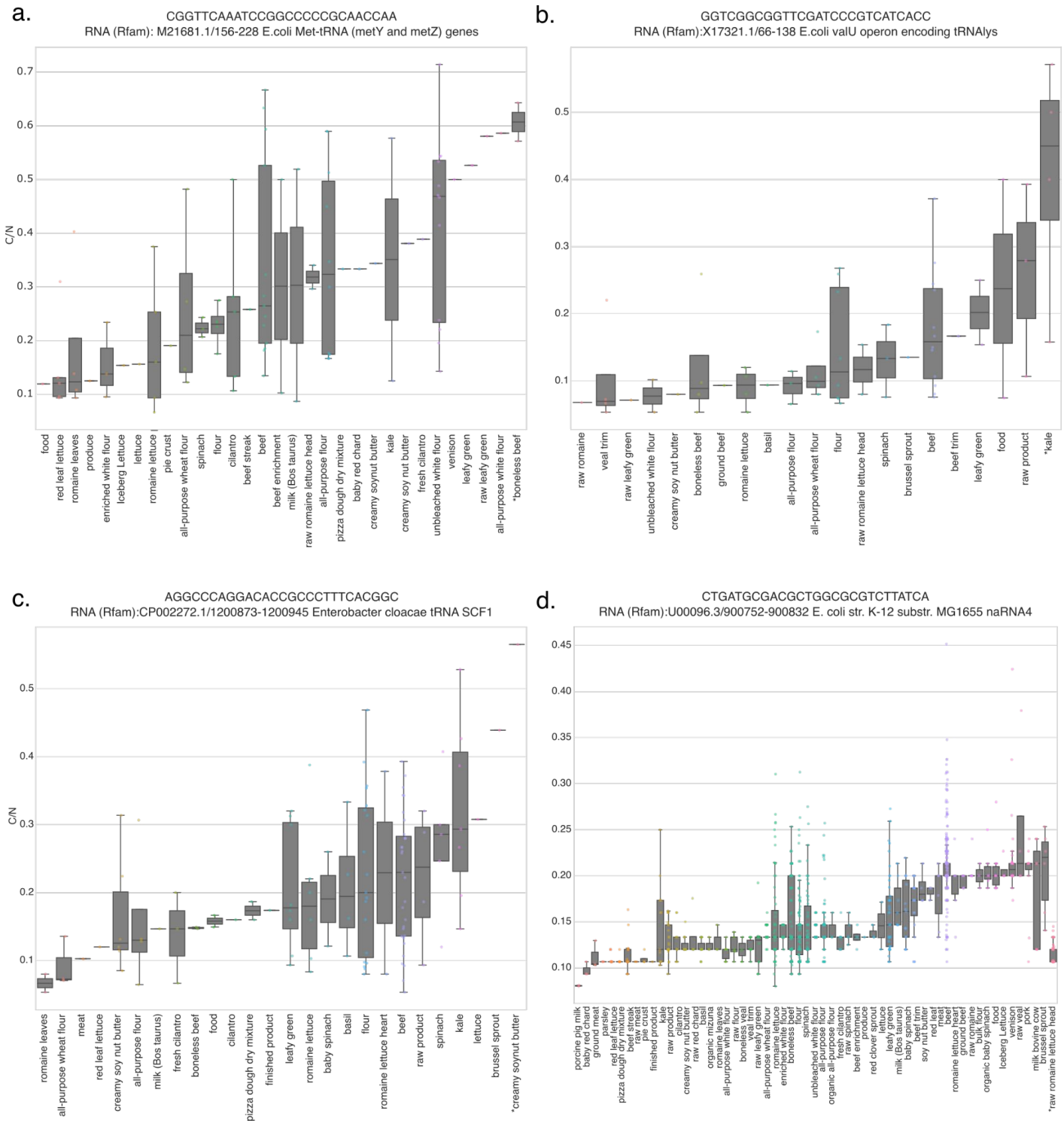


Figure S6