The ISME Journal (2019) 13:2589–2602
https://doi.org/10.1038/s41396-019-0411-2

ARTICLE

Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy

JL Weissman[1] · Rohan M. R. Laljani[1] · William F. Fagan[1] · Philip L. F. Johnson[1]

Received: 4 October 2018 / Revised: 15 March 2019 / Accepted: 24 March 2019 / Published online: 25 June 2019

Abstract

Bacteria and archaea are locked in a near-constant battle with their viral pathogens. Despite previous mechanistic characterization of numerous prokaryotic defense strategies, the underlying ecological drivers of different strategies remain largely unknown and predicting which species will take which strategies remains a challenge. Here, we focus on the CRISPR immune strategy and develop a phylogenetically-corrected machine learning approach to build a predictive model of CRISPR incidence using data on over 100 traits across over 2600 species. We discover a strong but hitherto-unknown negative interaction between CRISPR and aerobicity, which we hypothesize may result from interference between CRISPR-associated proteins and non-homologous end-joining DNA repair due to oxidative stress. Our predictive model also quantitatively confirms previous observations of an association between CRISPR and temperature. Finally, we contrast the environmental associations of different CRISPR system types (I, II, III) and restriction modification systems, all of which act as intracellular immune systems.

Introduction

In the world of prokaryotes, infection by viruses poses a constant threat to continued existence (e.g. [1]). In order to evade viral predation, bacteria and archaea employ a range of defense mechanisms that interfere with one or more stages of the viral life-cycle. Modifications to the host's cell surface can prevent viral entry in the first place. Alternatively, if a virus is able to enter the host cell, then intracellular immune systems, such as the clustered regularly inter-spaced short palindromic repeat (CRISPR) adaptive immune system or restriction-modification (RM) innate immune systems, may degrade viral genetic material and thus prevent replication [2–7]. Despite our increasingly in-depth understanding of the mechanisms behind each of these defenses, we lack a comprehensive understanding of the factors that cause selection to favor one defense strategy over another.

Here we focus on the CRISPR adaptive immune system, which is a particularly interesting case study due to its uneven distribution across prokaryotic taxa and environments. Previous analyses have shown that bacterial thermophiles and archaea (both mesophilic and thermophilic) frequently have CRISPR systems (~90%), whereas less than half of mesophilic bacteria have CRISPR (~40%; [8–12]). Environmental samples have revealed that many uncultured bacterial lineages have few or no representatives with CRISPR systems, and that the apparent lack of CRISPR in these lineages may be linked to an obligately symbiotic lifestyle and/or a highly reduced genome [13]. Nevertheless, no systematic exploration of the ecological conditions that favor the evolution and maintenance of CRISPR immunity has been made. Additionally, though these previous results appear broadly to be true [14], no explicit accounting has been made for the potentially confounding effects of phylogeny in linking CRISPR incidence to particular traits.

What mechanisms might shape the distribution of CRISPR systems across microbes? Some researchers have emphasized the role of the local viral community, suggesting that when viral diversity and abundance is high CRISPR will fail, and thus be selected against [11, 12, 15]. Others have focused on the tradeoff between constitutively

Supplementary information The online version of this article (https://doi.org/10.1038/s41396-019-0411-2) contains supplementary material, which is available to authorized users.

✉ Philip L. F. Johnson
plfj@umd.edu

[1] Department of Biology, University of Maryland, College Park, MD, USA

expressed defenses like membrane modification and inducible defenses such as CRISPR [15]. Yet others have noted that hot, and possibly other extreme environments can constrain membrane evolution, necessitating the evolution of intracellular defenses like CRISPR or RM systems [16–18]. Many have observed that since CRISPR prevents horizontal gene transfer, it may be selected against when such transfers are beneficial (e.g. [19, 20]). More recently it has been shown that at least one CRISPR-associated (Cas) protein can suppress non-homologous end-joining (NHEJ) DNA repair, which may lead to selection against having CRISPR in some taxa [21]. In order to determine the relative importances of these different mechanisms, we must first identify the habitats and microbial lifestyles associated with CRISPR immunity.

Here we aim to expand on previous analyses of CRISPR incidence in three ways: (1) by drastically expanding the number of environmental and lifestyle traits considered as predictors using the combination of a large prokaryotic trait database and machine learning approaches, (2) by incorporating appropriate statistical corrections for non-independence among taxa due to shared evolutionary history, which has not always been done, and (3) by simultaneously looking for patterns in RM systems, which will help us untangle the difference between environments that specifically favor CRISPR adaptive immunity versus DNA-degrading intracellular immune systems in general (RM and CRISPR).

Methods

Data

For a schematic outlining the entire data compilation process Fig. S15. For a list of all visualizations, predictive models, and statistical tests see Text S7.

Trait data

We downloaded the ProTraits microbial traits database [22], which describes 424 traits in 3046 microbial species. These traits include metabolic phenotypes, preferred habitats, and specific behaviors like motility, among many others. ProTraits was built using a semi-supervised text mining approach, drawing from several online databases and the literature. All traits are binary, with categorical traits split up into dummy variables (e.g. oxygen requirement listed as “aerobic”, “anaerobic”, and “facultative”). For each trait in each species, two “confidence scores” in the range [0, 1], are given, corresponding to the confidence of the text mining approach that a particular species does (c_+) or does not (c_-) have a particular trait.

We derived a single score (p) that captured the confidences both that a species does and does not have a particular trait. Assuming we want our score to lay in the interval [0,1], such a score should be zero when we are completely confident that a species does not have a trait, one when we are completely confident that a species has a trait, and 0.5 when we are completely uncertain whether or not a species has a trait (i.e., equally confident that it does and does not have the trait). In the following formula, $\frac{c_+}{c_+ + c_-}$ captures the relative confidence that a species does rather than does not have a trait, which we then scale by the overall maximal confidence (so that as overall confidence decreases the score shrinks toward 0.5)

$$p = \frac{1}{2} + \left(\frac{c_+}{c_+ + c_-} - \frac{1}{2}\right) \times \max(c_+, c_-). \quad (1)$$

Many of the scores are missing for particular species-trait combinations (18%), indicating situations in which the text mining approach was unable to make a trait prediction. Our downstream analyses do not tolerate missing data, and so we imputed missing values using a random forest approach (R package missForest; [23]). There is a set of summary traits in the ProTraits dataset that were created de-novo using a machine learning approach, as well as a number of traits describing the growth substrates a particular species can use. We removed both summary and substrate traits from the dataset for increased interpretability (post-imputation; 174 traits remaining).

We note that the authors of ProTraits also used genomic data to help them infer trait scores, though we found that the exclusion of this data does not affect our overall outcome (Text S6 and Fig. S9).

Genomic data and immune systems

For each species listed in the ProTraits dataset we downloaded a single genome from NCBI’s RefSeq database, with a preference for completely assembled reference or representative genomes. See Text S2 and Fig. S21 for a confirmation that our results are robust to the resampling of genomes. A number of species (333) had no genomes available in RefSeq, or only had genomes that had been suppressed since submission, and we discarded these species from the ProTraits dataset.

CRISPR incidence in each genome was determined using CRISPRDetect [24]. Additionally, data on the number of CRISPR arrays found among all available RefSeq genomes from a species were taken from Weissman et al. [25])

We downloaded the REBASE Gold database of experimentally verified RM proteins and performed blastx searches of our genomes against this database [26, 27]. The distribution of E-values we observed was bimodal, providing a natural cutoff ($E < 10^{-19}$).

To assess the ability of a microbe to perform non-homologous end-joining (NHEJ) DNA repair we used hmmsearch to search the HMM profile of the Ku protein implicated in NHEJ against all RefSeq genomes (*E*-value cutoff of 10^{-2}/# genomes; Pfam PF02735; [28–30]). We also used the annotated number of 16S rRNA genes in each downloaded RefSeq genome as a proxy for growth rate and the annotated *cas3*, *cas9*, and *cas10* genes as indicators of system type [31]. Where available as meta-data from NCBI, we also downloaded the oxygen (1949 records) and temperature requirements (1094 records) for the biosample record associated with each RefSeq genome. The NCBI trait data was used exclusively for building Fig. 4 and the analyses implicating Ku in the CRISPR versus oxygen association.

Phylogeny

We used PhyloSift to locate and align a large set of marker genes (738) found broadly across microbes, generally as a single copy [32, 33]. Of these marker genes, 67 were found in at least 500 of our genomes, and we limited our analysis to just this set. Additionally, eight genomes had few (<20) representatives of any marker genes and were excluded from further analysis. We concatenated the alignments for these 67 marker genes and used FastTree (general-time reversible and CAT options [34]); to build a phylogeny (Fig. S16). In order to analyze the effect of tree uncertainty on our phylogenetic regressions, we bootstrapped our dataset using seqboot and built a new tree from each replicate.

Visualizing CRISPR/RM incidence

The size of the ProTraits dataset, both in terms of number of species and number of traits, and the probable complicated interactions between variables necessitate techniques that can handle complex, large scale data. To visualize the structure of microbial trait space and the distribution of immune strategies within that space we made use of two unsupervised machine learning techniques, principal component analysis (PCA, prcomp() function in R) and *t*-distributed stochastic neighbor embedding (t-SNE, perplexity = 50 and 5000 iterations using Rtsne() function in Rtsne R package, otherwise default parameters, perplexity varied in Fig. S17; [35, 36]).

PCA is a well-used technique in ecology that allows us to reduce the dimensionality of a dataset for effective visualization in two-dimensional space. Essentially, we collapse our trait dataset into two or three composite traits and observe whether species with a particular immune strategy tend to vary systematically in terms of where they fall in this “trait space”. A newer variant of this approach, t-SNE, performs a similar process, but unlike PCA allows for non-linear transformations of trait space. Therefore, local structure and non-linear interactions between traits in high-dimensional space are preserved by t-SNE but often not captured by PCA [35]. On the other hand, t-SNE axes are less easily interpreted precisely because they represent non-linear rather than linear combinations of variables.

CRISPR/RM prediction from ProTraits

In order to predict the distribution of CRISPR and RM systems, we applied a number of supervised machine learning approaches to our dataset (see Fig. S18 for a flowchart describing the logic behind our model choices). In order to obtain accurate estimates of model performance, we initially set aside a portion of the data as a test set to be used exclusively in model assessment after all models were constructed (no fitting to this set). Because of the underlying evolutionary relationships in the data, we chose a test set that is phylogenetically independent of our training set. Alternatively, if we were to draw a test set at random from the microbial species we would risk underestimating our prediction errors due to non-independence of the training and test sets [37]. We chose the Proteobacteria as a test set because they are well-represented in the dataset (1139 species), ecologically diverse, and highly heterogeneous in terms of CRISPR incidence (Fig. S19). The remaining phyla were used to train our models.

First we built a series of linear models to classify species by immune strategy (CRISPR present or absent) using logistic regression. We had a large number of predictor variables (100+), which necessitated a model-selection approach in order to build a reasonably (and optimally) sized model. We used a forward selection algorithm to select the optimal set of predictors for each model size, with mean squared error under cross validation (CV) as our optimality criterion. We then selected model size by comparing BIC among these optimal models (i.e., selecting the model with the lowest score).

Similar to choosing a test set, care must be taken when performing CV on phylogenetically-structured data. CV assumes that when the data is partitioned into folds, each of these folds is independent of the others. If we draw species at random from a phylogeny, this assumption is violated, since the same hierarchical tree-structure will underlay each fold. Therefore, it is better to perform “blocked” CV than random CV [37], wherein folds are chosen based on divergent groups on the tree (e.g. phyla). If each group has diverged far enough in the past from the others, we can consider these folds to be essentially evolutionarily independent in terms of trait evolution (Fig. S20 for a conceptual example). Therefore blocked CV is essentially a non-parametric method (i.e., no explicit evolutionary model) to account for the non-independence arising from the shared evolutionary history between species. We use both random and blocked CV to build models. We clustered the data into blocked folds using the pairwise distances between tips on our tree (partitioning around mediods, pam() function in R package cluster, five folds so that $k = 5$; [38, 39]). A key assumption we make here is that our folds can be taken as independent from one another (i.e. no effect of shared evolutionary history). Since these clusters correspond roughly to Phylum-level splits, and since CRISPR and other prokaryotic immune systems are rapidly gained and lost over evolutionary time [40], we are comfortable making this assumption. We also repeated this analysis using phylogenetic logistic regression to more formally correct for phylogeny (R package phylolm; [41, 42]). Phylogenetic logistic regression is a more powerful method since it fits an explicit model of trait evolution, although it relies on the assumption that traits evolve according to the chosen model and can give misleading results otherwise.

Stepwise methods for variable selection, such as those used above (i.e., forward subset selection), are simple, computationally feasible, and easy to implement and interpret, but perform poorly when variables in the dataset covary with one another (i.e. multicollinearity; [43, 44]). As it so happens, the trait data used here exhibit strong multicollinearity (R package mctest; [45, 46]). Therefore, we sought out methods that deal well with this type of data, specifically partial least squares regression (PLS; [43]). Briefly, PLS combines features of PCA and linear regression to find the linear combination of predictors that maximizes the variance of the data in the space of outcome variables. We use a variant of PLS, sparse partial least squares discriminant analysis (sPLS-DA), where the "sparse" refers to a built-in variable selection process in the model-fitting algorithm and "discriminant analysis" refers to the fact that we are focused on a classification problem (i.e., presence vs. absence of a particular immune strategy; we used tune.splsda() perform five-fold cross validation, repeated 50 times, to select the optimal number of components n to include and splsda() to perform variable selection and model selection simultaneously given n as an input; functions in R package mixOmics; [47, 48]).

We also attempt to ameliorate the effects of shared evolutionary history on our PLS model by using a philosophically similar approach to our blocked CV method above. Multivariate integrative (MINT) sPLS-DA is a variant of PLS that can account for systematic variation between groups of data when those groupings are known (e.g., our phylogenetically blocked folds from above). It was originally developed for use in situations where multiple experiments testing the same hypothesis could show systematic biases from one another. In our case, the history of prokaryotic evolution is our experiment, and deep branching lineages are our replicates. We apply MINT sPLS-DA to the data, using the same blocked folds we used for CV (we used tune.mint.splsda() to perform five-fold blocked cross validation to select the optimal number of components n to include and mint.splsda() to perform variable selection and model selection simultaneously given n as an input; functions in R package mixOmics [48, 49]).

While regression provides easily interpretable trait weights and is computational efficient, in order to capture higher-order relationships between microbial traits we needed more powerful methods. Random forests (RF) are an attractive choice for our aims since they produce a readily-interpretable output and can incorporate nonlinear relationships between predictor variables [50]. We built an RF classifier on our training data from 5000 trees (otherwise default settings in R package randomForest so that the number of variables tried at each split is the square root of the total number of predictors [51]). To prevent fitting to phylogeny, we took an ensemble approach which was similar in philosophy to our blocked CV and MINT sPLS-DA approaches above. Using the phylogenetically blocked folds defined above we fit five individual forests, each leaving out one of the five folds. We then weighted these forests by their relative predictive ability on the respective fold excluded during the fitting process (measured as Cohen's κ, [52]). We predicted using our ensemble of forests by choosing the predicted outcome with the greatest total weight.

Results

Below, we associate specific microbial immune strategies with a diverse list of microbial traits. The traits span a range of scales including aspects of habitat (e.g. "aquatic"), morphology (e.g., "coccus"), and physiology (e.g., "heterotroph") [22]. While this variety of scales poses a modeling challenge to traditional approaches including linear regression, machine learning algorithms provide an elegant means of integrating such multi-scale traits in a statistically rigorous predictive framework. In particular, we apply algorithms that excel at identifying both linear and nonlinear combinations of traits with high predictive ability. For a systematic comparison of the output of our predictive models, discussed individually below, please see Fig. S1.

Visualizing CRISPR incidence in trait space

We visualized CRISPR incidence in microbial trait space using two unsupervised algorithms to collapse high-dimensional data (174 binary traits assessed in 2679 species; see Methods) into fewer dimensions. Both methods revealed clear differences between the placement of

Table 1 Top 10 variable loadings on the first three principal components of the PCA performed on the microbial traits dataset, shown in Fig. 1 and S2

PC1	Weight	PC2	Weight	PC3	Weight
ecosystemcategory_human	−0.16	temperaturerange_mesophilic	0.19	growth_in_groups	−0.24
specificecosystem_sediment	0.16	temperaturerange_thermophilic	−0.19	gram_stain_positive	−0.24
ecosystem_environmental	0.16	oxygenreq_strictanaero	−0.19	cellarrangement_singles	0.21
knownhabitats_host	−0.15	temperaturerange_hyperthermophilic	−0.18	cellarrangement_filaments	−0.20
ecosystemsubtype_intertidalzone	0.15	knownhabitats_hotspring	−0.17	sporulation	−0.20
ecosystem_hostassociated	−0.15	exosystemtype_rhizoplane	0.17	energysource_chemoorganotroph	−0.19
habitat_hostassociated	−0.15	habitat_specialized	−0.16	cellarrangement_clusters	−0.18
habitat_freeliving	0.15	metabolism_methanogen	−0.16	shape_tailed	−0.18
ecosystemtype_digestivesystem	−0.14	ecosystemcategory_plants	0.15	habitat_terrestrial	−0.18
specificecosystem_fecal	0.14	ecosystemtype_thermalsprings	−0.15	motility	0.17

These three components explain 17%, 10%, and 7% of the total variance, respectively

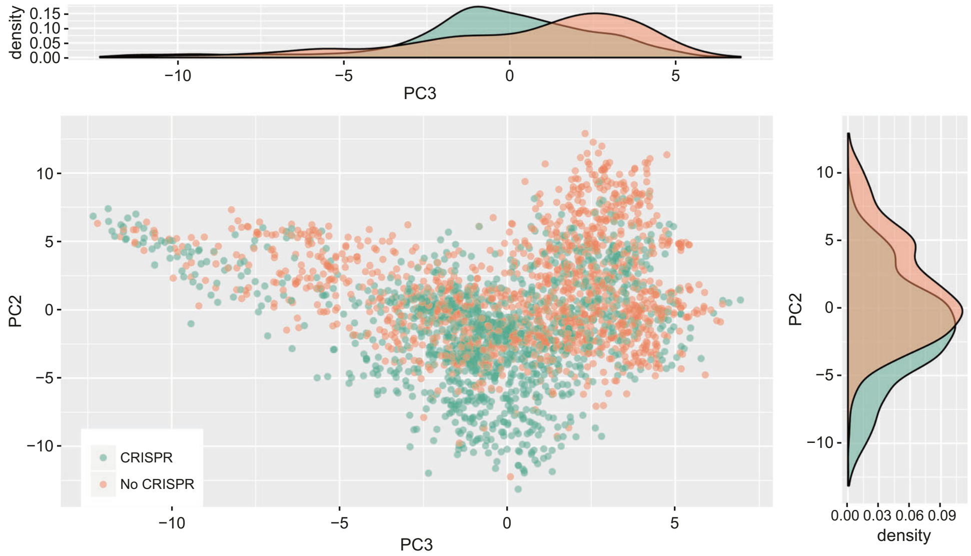

Fig. 1 Organisms with CRISPR separate from those without in trait space. The second and third components from a PCA of the microbial traits dataset are shown, where each point is a single species. CRISPR incidence is indicated by color (green with, orange without), but was not included when constructing the PCA. Notice the separation of organisms with and without CRISPR along both components. Marginal densities along each component are shown to facilitate interpretation. See Fig. S2 for the first component

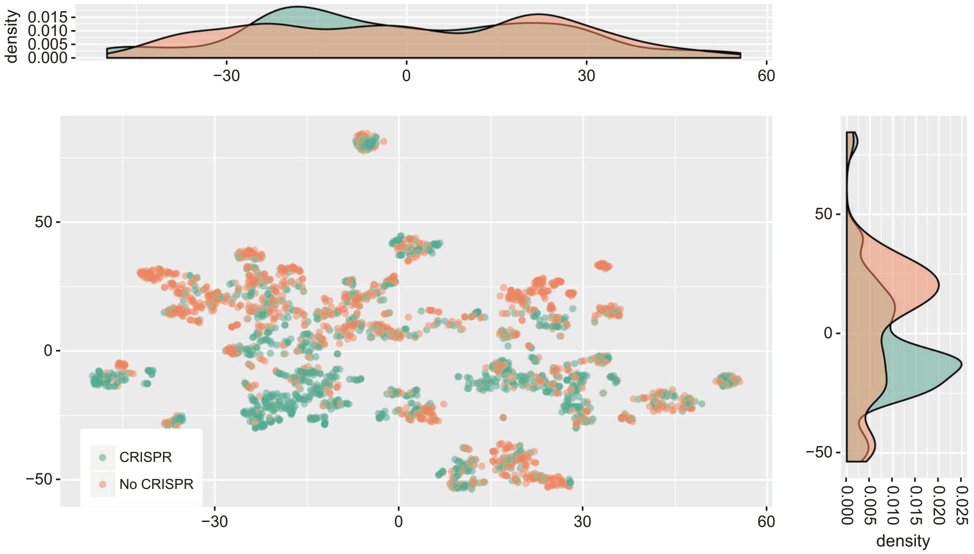

Fig. 2 Organisms with CRISPR partially cluster in trait space away from those without. Two dimensional output of t-SNE dimension reduction of the microbial traits dataset are shown, where each point is a single species (same dataset as in Fig. 1). CRISPR incidence is indicated by color (green with, orange without), but was not included when performing dimension reduction. The axes of t-SNE plots have no clear interpretation due to the non-linearity of the transformation

Table 2 Predictive ability of models of CRISPR incidence on the Proteobacteria test set

Model type	Phylogenetic correction		Performance			
	Non-parametric	Parametric	Model size	Accuracy (%)	κ	TPR
Log. Reg.	No	No	18	66.1	0.152	0.233
Log. Reg.	Yes	No	9	67.5	0.168	0.209
Log. Reg.	No	Yes	10	67.7	0.188	0.246
Log. Reg.	Yes	Yes	6	67.4	0.160	0.294
sPLS-DA	No	No	[7, 159, 4, 169, 50] (5 comp.)	68.4	0.190	0.219
MINT sPLS-DA	Yes	No	32 (1 comp.)	60.5	0.173	0.538
RF	No	No	–	68.8	0.241	0.327
RF Ensemble	Yes	No	–	68.6	0.240	0.332

Model size refers to number of variables chosen overall, or per-component in the case of the partial least squares models. Accuracy is measured as the total number of correct predictions over the total attempted and κ is Cohen's κ, which corrects for uneven class counts that can inflate accuracy even if discriminative ability is low. Roughly, κ expresses how much better the model predicts the data than one that simply knows the frequency of different classes ($\kappa = 0$ being no better, $\kappa > 0$ indicating improved predictive ability). The true positive rate (TPR) is the number of correctly identified genomes having CRISPR divided by the total number of genomes having CRISPR in the test set. The non-parametric correction for phylogeny refers to our phylogenetically blocked folds, whereas the parametric correction refers to our use of phylogenetic logistic regression [41]. Observe that the RF model appears to perform best at prediction in general

CRISPR-encoding and CRISPR-lacking organisms in trait space, despite the fact that no explicit information about CRISPR was included.

First, principal components analysis (PCA) of the trait data reveals several previously recognized patterns of microbial lifestyle choice and CRISPR incidence. The first principal component (17% variance explained) corresponds broadly to an axis running from host-associated to free-living microbes (Table 1), as observed by others [53, 54]. CRISPR-encoding and CRISPR-lacking microbes are not differentiated along this axis (Fig. S2). We see CRISPR-encoding and CRISPR-lacking organisms beginning to separate along the second (10% variance explained) and third (7% variance explained) principal components (Fig. 1). The second component roughly represents a split between extremophilic species typically living in low-productivity environments and mesophilic, plant-associated species (Table 1). Optimal growth temperature appears to be an important predictor of CRISPR incidence, as previously noted by others [11, 12]. The third component is not as easy to interpret, but appears to indicate a spectrum from group-living microbes (e.g. biofilms) to microbes that tend to live as lone, motile cells (Table 1). That CRISPR is possibly favored in group-living microbes is not entirely surprising, considering the increased risk of viral outbreak at high population density, and that some species up-regulate CRISPR during biofilm formation [55].

Second, we visualized the trait data using *t*-distributed stochastic neighbor embedding (t-SNE), which is a non-linear method that can often detect more subtle relationships in a dataset (Fig. 2[35]). This method reveals a clustering of CRISPR-encoding microbes in trait space, further emphasizing that microbial immune strategy is influenced by ecological conditions. Because the axes of t-SNE plots are not easily interpretable, we mapped the top weighted traits from the PCA above (Table 1) onto the t-SNE reduced data (Fig. S3). Surprisingly, the most clearly aligned trait with CRISPR-incidence is having an obligately anaerobic metabolism.

Predicting CRISPR incidence

The above unsupervised approaches (i.e. uninformed about the outcome variable, CRISPR) revealed that CRISPR incidence appears to be impacted by other microbial traits. In order to more formally characterize these patterns, and exploit them for their predictive ability, we applied several supervised prediction methods (i.e. trained with information about CRISPR incidence) methods to the complete trait dataset.

Unlike traditional statistical techniques focused on assigning *p*-values to particular input variables, with our machine learning approach we assessed model performance in terms of predictive ability. For unbiased error estimates, we chose an independent "test" set to withhold during the model fitting process and to be used only during model assessment. We consider effective prediction of CRISPR incidence in this independent dataset as support that our model encodes real information about how different microbial traits influence the ecological advantages of the CRISPR system. We then examined the structure of these models, and which variables play an outsize role in their

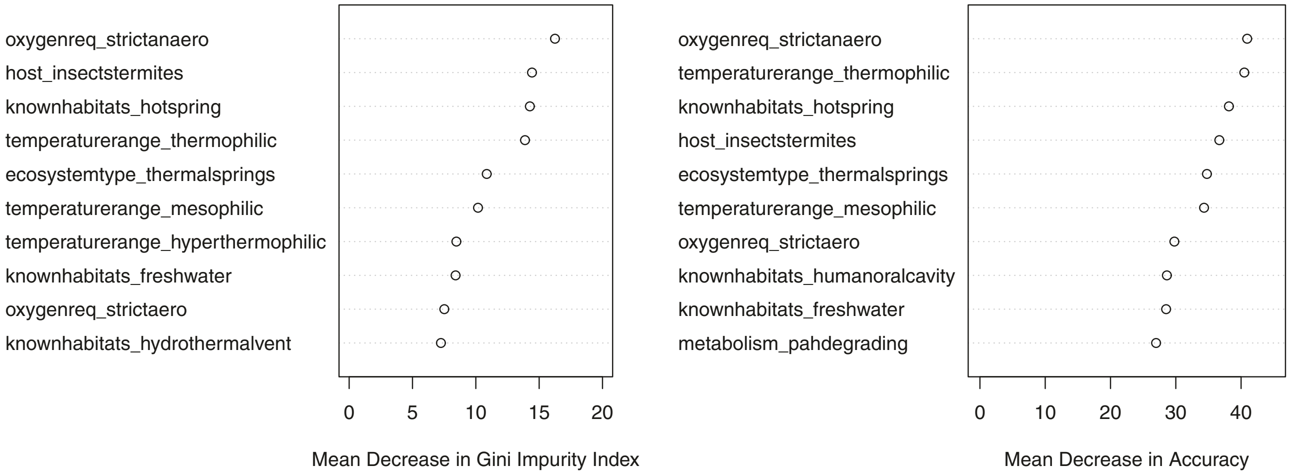

Fig. 3 Importance of top ten predictors in the RF model of CRISPR incidence using the ProTraits predictors. The mean decrease in accuracy measures the reduction in model accuracy when a variable is randomly permuted in the dataset. The Gini impurity index is a common score used to measure the performance of decision-tree based models (e.g. RF models). Briefly, when a decision tree is built the Gini impurity index measures how well separated the different classes of outcome variable are at the terminal nodes of the tree (i.e., how "pure" each of the nodes is). The mean decrease in Gini impurity measures the estimated reduction in impurity (increase in purity) when a given variable is added to the model. These importance scores are useful to rank variables as candidates for further study, but in themselves should not be taken as statistical support or effect sizes similar to those seen in linear regression. RF models may include non-linear combinations of variables, and therefore the contribution of any one variable is not as easily interpreted as with a linear model, a drawback of this approach. See Fig. S7 for all predictor importances

performance, in order to select candidate traits associated with CRISPR incidence. Importantly, we chose the Proteobacteria as our test set because they represent a phylogenetically independent group from our training set (see Methods).

All models we implemented showed improved predictive ability over a null model only accounting for the relative frequency of CRISPR among species (Cohen's $\kappa > 0$; Table 2), indicating that there is some ecological signal in CRISPR incidence, though overall predictive performance was not overwhelming. Of these models the random forest (RF) model ranked highest, and did reasonably well ($\kappa = 0.241$). The percent incidences of CRISPR in the training (56%) and test sets (36%) are considerably different, which may have been difficult for these models to overcome. It is also possible that the Proteobacteria vary systematically from other phyla in terms of ecology and immune strategy, making them a particularly difficult (and thus conservative) test set. Nevertheless, the trait data clearly held some information about CRISPR incidence. We will primarily focus here on the RF model since it performed best, but see Text S1 for further discussion of the performance of our other models.

While each of our models revealed a distinct set of top predictors of CRISPR incidence, there was broad agreement overall (Table S1, Fig. 3, S4 and S5). Keywords indicating a thermophilic lifestyle (e.g. thermophilic, hot springs, hyperthermophilic, thermal springs) appeared across all models as either the most important or second most important predictor of CRISPR incidence. Keywords relating to oxygen requirement (e.g. anaerobic, aerobic) also appeared across nearly all models as top predictors, excluding only the two worst performing models (Table S1). In the case of the RF and sPLS-DA models, oxygen requirement was always one of the top three predictors, and often the top predictor of CRISPR incidence (Fig. 3, S4, S5 and S6). Other predictors that frequently appeared across model types included termite hosts (host_insectstermites), the degradation of polycyclic aromatic hydrocarbons (PAH; metabolism_pahdegrading), freshwater habitat (knownhabitats_freshwater), and growth as filaments (shape_filamentous). In general, the sPLS-DA, MINT sPLS-DA, RF, and RF ensemble models agreed with each other rather closely. Finally, we built an RF model using only traits related to temperature range, oxygen requirement, and thermophilic lifestyle (hot springs, thermal springs, hydrothermal vents). This temperature- and oxygen-only RF model outperformed all non-RF models ($\kappa = 0.191$). These traits alone appear to hold the majority of information about CRISPR incidence in the dataset.

As an additional check that these candidate traits versus CRISPR associations are real and not due to some irregularity in our dataset, we downloaded meta-data available from NCBI. We were able to reproduce the result that thermophiles strongly prefer CRISPR (92% with CRISPR as opposed to 49% in mesophiles, Fig. 4a [11, 12]). Though we have too few genomes categorized as psychrotolerant (35) or psychrophilic (14) to make any strong claims, these genomes seem to lack CRISPR most of the time, suggesting that CRISPR incidence decreases continuously as

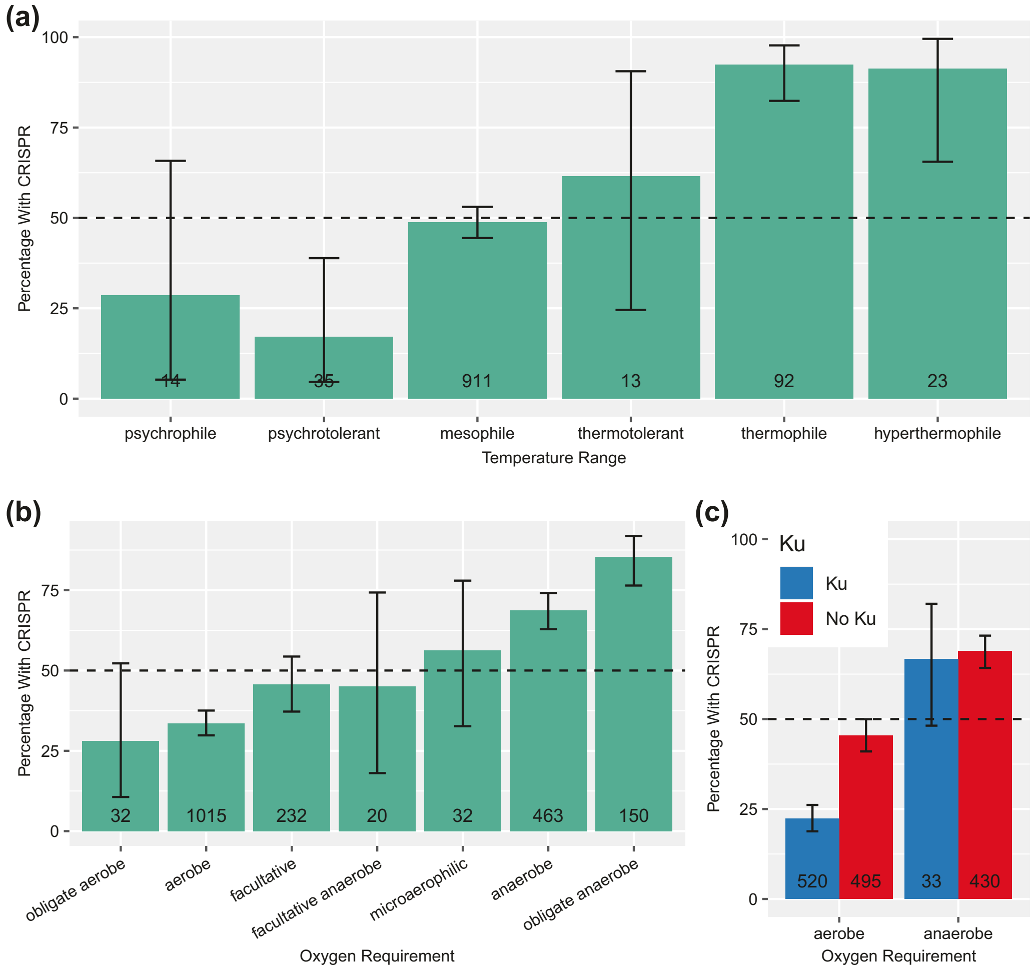

Fig. 4 Temperature range and oxygen requirement are strong predictors of CRISPR incidence. Trait data taken from NCBI. **a** Thermophiles strongly favor CRISPR immunity, while mesophiles appear ambivalent. **b** Anaerobes favor CRISPR immunity, while aerobes tend to lack CRISPR and facultative species fall somewhere in between. **c** CRISPR and the Ku protein are negatively associated in aerobes but not anaerobes. Error bars are 99% binomial confidence intervals (non-overlapping intervals can be taken as evidence for a statistically significant difference at the $p < 0.01$ level). Total number of genomes in each trait category shown at the bottom of each bar. Categories represented by fewer than 10 genomes were omitted

environmental temperatures decrease [10]. We were also able to confirm that, in agreement with our visualizations and predictive modeling, aerobes disfavor CRISPR immunity (34% with CRISPR) while anaerobes favor CRISPR immunity (67% with CRISPR, Fig. 4b). This is true independent of growth temperature, with mesophiles showing a similarly strong oxygen-CRISPR link (Fig. S8). Overall, both oxygen ($\chi^2 = 254.04$, $p < 2.2 \times 10^{-16}$, categories with <10 observations excluded) and temperature ($\chi^2 = 98.86$, $p < 2.2 \times 10^{-16}$, categories with <10 observations excluded) had significant effects on incidence (for breakdown see Fig. 4).

Following previous suggestions that CRISPR incidence might be negatively associated with host population density and growth rate [11, 12, 15], and that this could be driving the link between CRISPR incidence and optimal temperature range, we sought to determine if growth rate was a major determinant of CRISPR incidence. The number of 16S rRNA genes in a genome is an oft used, if imperfect, proxy for microbial growth rates and an indicator of copiotrophic lifestyle in general [56–58]. While CRISPR-encoding genomes had slightly more 16S genes than CRISPR-lacking ones (3.1 and 2.9 on average, respectively), the 16S rRNA gene count in a genome was not a significant predictor of CRISPR incidence (logistic regression, $p = 0.05248$), although when correcting for phylogeny 16S gene count does seem to be significantly positively associated with CRISPR incidence (phylogenetic logistic regression, $m = 0.06277$, $p = 6.651 \times 10^{-5}$), the opposite of what we would expect if growth rate were driving the CRISPR-temperature relationship (though the effect was not consistent across bootstrapped trees; Table S2).

As a secondary confirmation of the link between oxygen and CRISPR, we examined metagenomic data from the

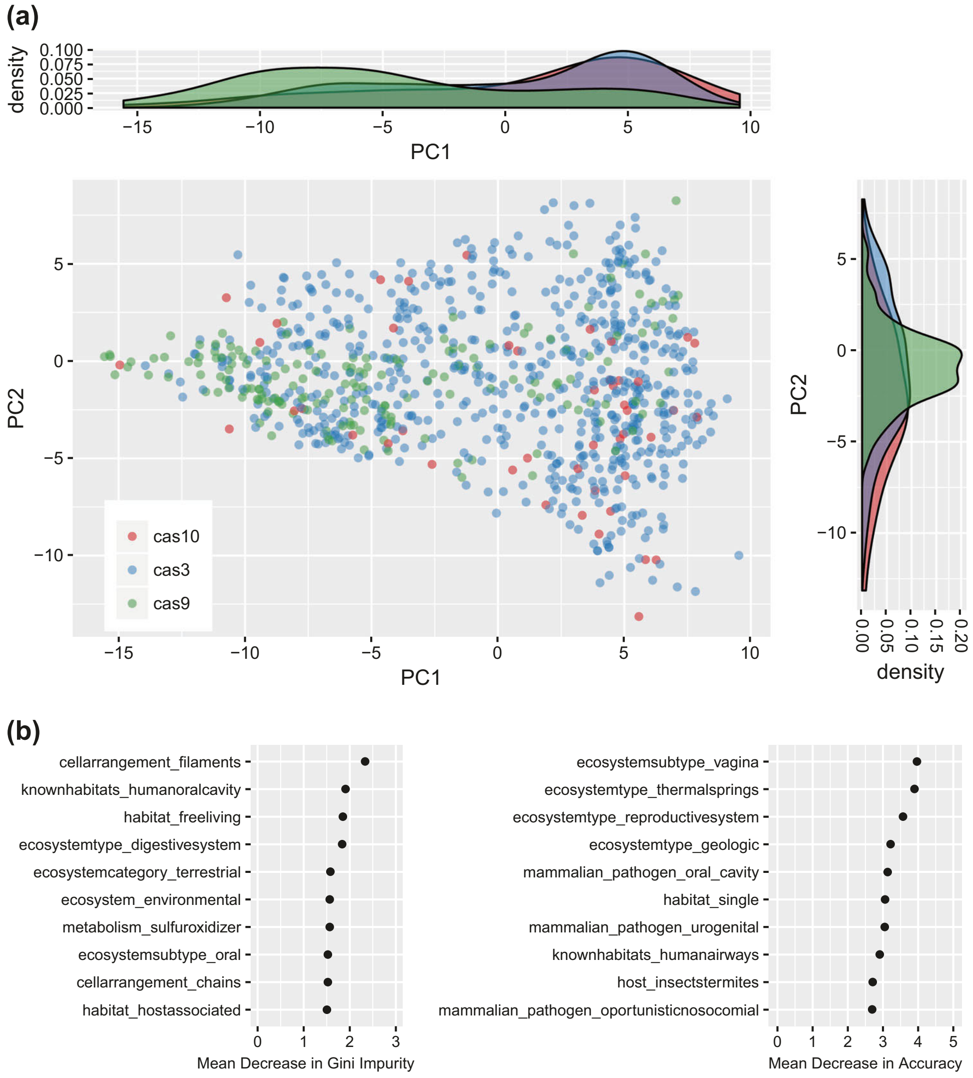

Fig. 5 Type II CRISPR systems appear to be more prevalent in host-associated microbes. **a** The cas targeting genes associated with type I, type II, and type III systems (*cas3*, *cas9*, and *cas10*, respectively) mapped onto the PCA in Fig. S2. Organisms without any targeting genes were omitted from the plot for readability. Recall from Table 1 that PC1 roughly corresponds to a spectrum running from host-associated to free-living microbes. **b** A variable importance plot from an RF model of *cas9* incidence. Observe that keywords related to a host-associated lifestyle appear many times

Tara Oceans Project [59], and found that across a large set of ocean metagenome samples CRISPR prevalence was inversely related to environmental oxygen concentration (Text S3 and Fig. S22).

We also attempted to predict the number of CRISPR arrays in a genome given that that genome had at least one array, though this attempt was entirely unsuccessful (Text S4).

Predicting CRISPR type

Each CRISPR system type is associated with a signature *cas* targeting gene unique to that type (*cas3*, *cas9*, and *cas10* for type I, II, and III systems, respectively). There are many species in the dataset with *cas3* (605), but relatively few with *cas9* (160) and *cas10* (222), suggesting that the traits correlated with CRISPR incidence probably correspond primarily to type I systems (the dominance of type I systems has been noted previously [60]). We mapped the incidence of each of these genes onto the PCA we constructed earlier (see Fig. S2 and Table 1), and found that *cas9* separates from *cas3* and *cas10* along the first component (Fig. 5a). Broadly, this indicates that type II systems are more commonly found in host-associated than free-living microbes, the opposite of the other two system types.

We built an RF model of *cas9* incidence, with the Proteobacteria as the test set. Because our training set had so few cases of *cas9* incidence (10% of set), we performed stratified sampling during the RF construction process to ensure representative samples of organisms with and without *cas9*. Surprisingly, despite the extremely small number of organisms with *cas9* in the training and test sets (160 and 58, respectively), this model was accurately able to predict type II CRISPR incidence and had some discriminative ability (Accuracy = 93.0%, $\kappa = 0.164$), though it missed many of the positive cases (TPR = 0.172). This model also suggested that a host-associated lifestyle seems to be a major factor influencing the incidence of type II systems, with many of the top-ranking variables in terms of importance corresponding to keywords having to do with the split between host-associated and free-living organisms (Fig. 5b).

NHEJ, CRISPR, and oxygen

Recently, Bernheim et al. [21] demonstrated that the type II-A CRISPR system interferes with the NHEJ DNA repair pathway, leading to an inverse relationship between the presence of type II-A systems and the NHEJ pathway in microbial genomes. We hypothesized that this negative relationship between CRISPR and NHEJ might be more widespread across system types. We also hypothesized that this could explain the negative relationship between CRISPR and aerobicity we observe, since reactive oxygen species produced during aerobic respiration can induce double-strand breaks, thus selecting for the presence of NHEJ repair in aerobic organisms [61, 62]. We use the presence of Ku protein as a proxy for the NHEJ pathway, since this protein is central to the pathway.

There was a clear interaction between the presence of Ku and aerobicity on the incidence of CRISPR (Fig. 4c, using aerobicity meta-data from NCBI for this and below analyses). Using our full set of RefSeq genomes, we found a weak negative association between CRISPR and Ku incidence overall (Pearson's correlation, $\rho = -0.012$; $\chi^2 = 15.015$, $p = 1.067 \times 10^{-4}$), but restricting only to aerobes the negative association between Ku and CRISPR was much stronger (Pearson's correlation, $\rho = -0.250$, $p = 9.109 \times 10^{-16}$), whereas in anaerobes it was nonexistent ($\rho = -0.023$, $p = 0.704$). This pattern was consistent when correcting for phylogeny (Text S5 and Table S4), and was true for both type I and III systems individually, though was not significant for type II systems of which there were fewer in the dataset Fig. S12.

Similar to our CRISPR analysis, we used PCA and an RF model to find if and where Ku-possessing organisms clustered in trait space. We found that the NHEJ pathway clusters strongly in trait space (Fig. S10), and is favored in soil-dwelling, spore-forming, aerobic microbes, consistent with expectations of where NHEJ will be most important [61, 62] (Fig. S11).

Predicting RM incidence

So far, our analyses have not distinguished if temperature and oxygen predict whether a microbe has an intracellular immune system that degrades DNA in general, or whether these traits are specific to CRISPR adaptive immunity. We tested these two possibilities by building an RF model of restriction enzyme incidence using the same stratified sampling approach that we used for CRISPR system type. This model showed decent predictive ability ($\kappa = 0.317$). However, the correlation between variable importance scores for the CRISPR and restriction enzyme RF models was low (Fig. 3 vs Fig. S14; Pearson's correlation, $\rho = 0.169$ for mean decrease in Gini Impurity Index, $\rho = -0.0487$ for mean decrease in accuracy; also Fig. S1). This result implies that RM systems have different traits determining their incidence than do CRISPR systems (also note PCA plot, Fig. S13). When we directly tested for an association with temperature and oxygen we also found that the number of restriction enzymes was, unlike CRISPR incidence, negatively associated with an anaerobic lifestyle ($m = -4.53877$, $p = 2 \times 10^{-16}$, phylogenetic linear regression), and only marginally significantly associated with a thermophilic lifestyle ($m = 1.51063$, $p = 0.03779$, phylogenetic linear regression). These results were consistent across bootstrapped trees (Table S3).

Discussion

We detected a clear association between microbial traits and the incidence of the CRISPR immune system across species. We found that two predictors were especially important for predicting CRISPR incidence, thermophilicity and aerobicity. The links between these two traits and CRISPR were confirmed with annotations from NCBI, and in the case of aerobicity with metagenomic data from the Tara Oceans Project (Text S3 [59]). The relationship between temperature and CRISPR is well known [8–10], but we lend further support here by formally correcting for shared evolutionary history in our statistical analyses using both parametric and non-parametric approaches.

Previous theoretical models predict that CRISPR will be selected against in environments with dense and diverse viral communities [11, 12], since hosts are less likely to repeatedly encounter the same virus in such environments. These models in turn predict that in high-density host communities CRISPR will not be adaptive, since high host density leads to high viral diversity [11, 12], and that this might explain why potentially slow-growing thermophiles favor CRISPR immunity (as opposed to copiotrophic mesophiles). Our results show a marginal positive association between growth rate and CRISPR incidence, and that group-living microbes seem to favor CRISPR immunity, calling these prior viral diversity and density based explanations into question. Additionally, our analysis suggests that psychrophilic and psychrotolerant species disfavor CRISPR more strongly than mesophiles, which is not clearly explained or predicted by hypotheses based on host density.

We suspect that another factor could be affecting the degree of viral diversity that a host encounters, so that viral diversity is high in colder environments and low in hotter ones. Differences in dispersal limitation among viruses could lead to lower immigration rates in hot environments, as viral decay rates may be low at lower temperatures and high at higher temperatures [63], though this is highly speculative. We note that host dispersal rates are unlikely to affect the viral diversity seen by a host on average unless most of the host population is dispersing, an unrealistic expectation.

Surprisingly, we find that oxygen requirement appears to be just as important of a predictor of CRISPR incidence as temperature, and that this pattern is independent of any effect of temperature. Possibly, this association can be explained by inhibitory effects of CRISPR on NHEJ DNA repair. Type II-A CRISPR systems have been shown to directly interfere with the action of the NHEJ DNA repair pathway in prokaryotes [21]. Reactive oxygen species are produced during aerobic metabolism and can cause DNA damage [61], making NHEJ potentially particularly important in aerobes. Thus, if CRISPR interferes with the NHEJ repair pathway, and this pathway is important in aerobes, we would expect CRISPR incidence to be inversely related to the presence of oxygen. Our data showed a clear interaction between aerobicity and the NHEJ machinery in determining CRISPR incidence that suggests that the link between CRISPR and aerobicity may be mediated by the presence of the NHEJ pathway (Fig. 4c). The Cas proteins share many structural similarities with proteins implicated in DNA repair, and in some cases prefer to associate with DSBs, and it is perhaps unsurprising that they appear to broadly inhibit the NHEJ pathway whose proteins may be competing for substrate [64]. Nevertheless, the evidence supporting this hypothesis is only preliminary. The negative interaction between CRISPR and Ku should be experimentally confirmed in type I and type III systems. Additionally, our repair versus immunity tradeoff hypothesis could be tested using an experimental evolution setup in which organisms with CRISPR are exposed to DNA damage.

The link that we propose between aerobic metabolism and NHEJ repair is somewhat tenuous. Reactive oxygen species are thought to directly produce single strand breaks which are most often converted to double-strand breaks during cell growth, the precise time when repair may be possible via homologous recombination due to the presence of multiple genome copies. That being said, reactive oxygen species can lead to double-strand breaks during stationary phase when damage is spatially clustered on the genome [65, 66], when cells experience specific types of starvation that lead to vulnerable single-stranded DNA gaps [67, 68], or when ROS occurs in conjunction with other damaging agents including cyanide [69] and irradiation [70–72]. Furthermore, while NHEJ certainly will be important during stationary phase, its relevance during growth is unknown. The pathway itself does appear to be more prevalent in environments with oxygen (Figs. S10 and S11). Nevertheless, we have no ability to assess causality presently, and the strong interaction between Ku and aerobicity on CRISPR incidence we observed could be the result of some other, as yet unrevealed driver. For example, NHEJ is thought to be important for desiccation resistance [73, 74], and many organisms facing this specific threat are likely to be aerobic.

As an alternative to our NHEJ hypothesis, could patterns in viral diversity explain the relationship between aerobicity and CRISPR incidence? The viral-decay hypothesis we proposed to explain the enrichment of thermophiles with CRISPR does not make sense in this context, since we might expect viruses to decay more readily in the presence of oxygen rather than under anoxic conditions. It is unclear to us why the viruses of anaerobes would be more dispersal limited. Nevertheless, if the viral communities infecting anaerobes were shown to be less diverse than those infecting aerobes this could also explain the increased incidence of CRISPR among these organisms.

We found no strong link between the incidence or number of RM systems on a genome and a thermophilic or anaerobic lifestyle, suggesting that the major drivers of CRISPR incidence are indeed CRISPR specific, consistent with our viral diversity and NHEJ-inhibition hypotheses.

We were also able to show that CRISPR types vary in in terms of the environments they are found in, with type II systems appearing primarily in host-associated microbes. This phenomenon could be due in part to phylogenetic biases in the dataset, but our use of a phylogenetically independent test set lends credence to the overall trend. We have no clear mechanistic understanding of why *cas9* containing microbes tend to favor a host-associated lifestyle. Nevertheless this result may have practical implications for CRISPR genome editing, since it has recently been found that humans frequently have a preexisting adaptive immune response to variants of the Cas9 protein [75]. We note that type I and III systems do not appear to have a strong link to host-associated lifestyles.

While our dataset spanned a broad phylogenetic range (with some notable exceptions such as the Candidate Phyla Radiation [76]), we had a limited number of microbial traits, which may have obscured some important CRISPR-trait associations. With the number of microbial genomes in public databases constantly expanding, so too should efforts to provide metadata about each of the organisms represented by those genomes. At least part of the problem lies in the lack of a universally accepted controlled vocabulary for microbial traits (similar to that provided by the Gene Ontology Consortium [77]), although some admirable attempts have been made [78, 79]. This would both facilitate the construction of more expansive trait databases, and would help deal with the issue of comparing traits that span many different scales.

The ecological drivers of microbial immune strategy are likely as diverse as the ever-increasing number of known prokaryotic defense systems [80, 81]. The exploratory, database-centered approach we take here can be complemented by targeted studies examining shifts in immune strategy across environmental gradients (e.g., Text S3) to provide a more fine-grained understanding of how microbial populations adapt to their local pathogenic and abiotic environments. Ultimately, experimental manipulations will provide the power to fully validate proposed mechanisms behind ecological patterns in immune strategy.

Acknowledgements JLW was supported by a GAANN Fellowship from the U.S. Department of Education and the University of Maryland. WFF was partially supported the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant W911NF-14-1-0490. PLFJ was supported in part by NIH R00 GM104158.

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Munson-McGee JH, Peng S, Dewerff S, Stepanauskas R, Whitaker RJ, Weitz JS, et al. A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. ISME J. 2018;12:1.
2. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005;151: 2551–61.
3. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol. 2005;60: 174–82.
4. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315:1709–12.
5. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8:317–27.
6. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013;41:4360–77.
7. Sv Houte, Buckling A, Westra ER. Evolutionary ecology of prokaryotic immune mechanisms. Microbiol Mol Biol Rev. 2016;80:745–63.
8. Mojica Francisco JM, Cesar Díez-Villaseñor, Elena Soria, Guadalupe Juez. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol Microbiol. 2002;36:244–6.
9. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct. 2006;1:7.
10. Anderson RE, Brazelton WJ, Baross JA. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol. 2011;77:120–33.
11. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV. Viral diversity threshold for adaptive immunity in prokaryotes. MBio. 2012;3:e00456–12.
12. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. J Bacteriol. 2013;195: 3834–44.
13. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat Commun. 2016;7:10613.
14. Makarova KS, Wolf YI, Koonin EV. The basic building blocks and evolution of CRISPR–Cas systems. Biochem Soc Trans. 2013;41:1392–1400.
15. Westra ER, van Houte S, Oyesiku-Blakemore S, Makin B, Broniewski JM, Best A, et al. Parasite exposure drives selective evolution of constitutive versus inducible defense. Curr Biol. 2015;25:1043–9.
16. Chung YJ, Krueger C, Metzgar D, Saier MH. Size comparisons among integral membrane transport protein homologues in Bacteria, Archaea, and Eucarya. J Bacteriol. 2001;183:1012–21.
17. Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. 2005;33:3390–3400.
18. Ledford H. Five big mysteries about CRISPR's origins. Nat News. 2017;541:280.
19. Bikard D, Hatoum-Aslan A, Mucida D, Marraffini LA. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. Cell Host Microbe. 2012;12:177–86.
20. Jiang W, Maniv I, Arain F, Wang Y, Levin BR, Marraffini LA. Dealing with the evolutionary downside of crispr immunity: bacteria and beneficial plasmids. PLoS Genet. 2013;9:e1003844.
21. Bernheim A, Calvo-Villamañán A, Basier C, Cui L, EPC Rocha, Touchon M, et al. Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. Nat Commun. 2017;8:2094.
22. Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. 2016;44:10074–90.
23. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28: 112–8.
24. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics. 2016;17:356.
25. Weissman JL, Fagan WF, Johnson PL. Selective maintenance of multiple CRISPR arrays across prokaryotes. CRISPR J. 2018;1: 405–13.

26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
27. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2010;38:D234–D236.
28. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14:755–63.
29. Doherty Aidan J, Jackson Stephen P, Weller Geoffrey R. Identification of bacterial homologues of the Ku DNA repair proteins. FEBS Lett. 2001;500:186–8.
30. Aravind L, Koonin EV. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. Genome Res. 2001;11:1365–74.
31. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44:6614–24.
32. Lang JM, Darling AE, Eisen JA. Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices. PLOS ONE. 2013;8:e62510.
33. Darling AE, Jospin G, Lowe E, Iv FAM, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243.
34. Price MN, Dehal PS, Arkin AP. FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010;5:e9490.
35. Maaten Lvd, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
36. Krijthe JH Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation; 2015. R package version 0.15. https://github.com/jkrijthe/Rtsne.
37. Roberts David R, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40:913–29.
38. Reynolds AP, Richards G, Bdl Iglesia, Rayward-Smith VJ. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. J Math Model Algorithms. 2006;5:475–504.
39. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K cluster: Cluster Analysis Basics and Extensions; 2018. R package version 2.0.7-1.
40. Puigbò P, Makarova KS, Kristensen DM, Wolf YI, Koonin EV. Reconstruction of the evolution of microbial defense systems. BMC Evol Biol. 2017;17:94.
41. Ives AR, Garland T. Phylogenetic Logistic Regression for Binary Dependent Variables. Syst Biol. 2010;59:9–26.
42. LsT Ho, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst Biol. 2014;63:397–408.
43. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. Chemometr Intell Lab Syst. 2005;78:103–12.
44. Morozova O, Levina O, Uusküla A, Heimer R. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Medical Res Methodol. 2015;15:71.
45. Farrar DE, Glauber RR. Multicollinearity in Regression Analysis: The Problem Revisited. Rev Econ Stat. 1967;49:92–107.
46. Imdadullah M, Aslam M, Altaf S mctest: An R Package for Detection of Collinearity among Regressors. R Journal. 2016;8.
47. Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011;12:253.
48. Rohart F, Gautier B, Singh A, Cao KAL. mixOmics: An R package for ‘omics feature selection and multiple data integration. PLOS Comput Biol. 2017;13:e1005752.
49. Rohart F, Eslami A, Matigian N, Bougeard S, Lê Cao KA. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics. 2017;18:128.
50. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
51. Liaw A, Wiener M, others. Classification and regression by randomForest. R News. 2002;2.
52. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960;20:37–46.
53. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol. 2008;6:776–88.
54. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature. 2017;551:457–63.
55. Patterson AG, Jackson SA, Taylor C, Evans GB, Salmond GPC, Przybilski R, et al. Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas Systems. Mol Cell. 2016;64:1102–8.
56. Condon C, Liveris D, Squires C, Schwartz I, Squires CL. rRNA operon multiplicity in Escherichia coli and the physiological implications of rrn inactivation. J Bacteriol. 1995;177:4152–6.
57. Vieira-Silva S, Rocha EPC. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. PLOS Genet. 2010;6:e1000808.
58. Roller BRK, Stoddard SF, Schmidt TM. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nat Microbiol. 2016;1:16160.
59. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.
60. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015;13:722–36.
61. Karanjawala ZE, Murphy N, Hinton DR, Hsieh CL, Lieber MR. Oxygen Metabolism Causes Chromosome Breaks and Is Associated with the Neuronal Apoptosis Observed in DNA Double-Strand Break Repair Mutants. Curr Biol. 2002;12:397–402.
62. Pitcher RS, Brissett NC, Doherty AJ. Nonhomologous end-joining in bacteria: a microbial perspective. Ann Rev Microbiol. 2007;61:259–82.
63. Jończyk E, Kłak M, MiĘdzybrodzki R, Górski A. The influence of external factors on bacteriophages. Folia Microbiol. 2011;56:191–200.
64. Faure G, Makarova KS, Koonin EV CRISPR-Cas: Complex functional networks and multiple roles beyond adaptive immunity. J Mol Biol. 2018.
65. Dianov GL, Timehenko TV, Sinitsina OI, Kuzminov AV, Medvedev OA, Salganik RI. Repair of uracil residues closely spaced on the opposite strands of plasmid DNA results in double-strand break and deletion formation. Mol Gen Genet. 1991;225:448–52.
66. Kozmin SG, Sedletska Y, Reynaud-Angelin A, Gasparutto D, Sage E. The formation of double-strand breaks at multiply damaged sites is driven by the kinetics of excision/incision at base damage in eukaryotic cells. Nucleic Acids Res. 2009;37:1767–77.
67. Hong Y, Li L, Luan G, Drlica K, Zhao X. Contribution of reactive oxygen species to thymineless death in Escherichia coli. Nat Microbiol. 2017;2:1667.
68. Henrikus SS, Henry C, McDonald JP, Hellmich Y, Wood EA, Woodgate R, et al. DNA double-strand breaks induced by reactive oxygen species promote DNA polymerase IV activity in Escherichia coli. bioRxiv. 2019;p. 533422.
69. Mahaseth T, Kuzminov A. Prompt repair of hydrogen peroxide-induced DNA lesions prevents catastrophic chromosomal fragmentation. DNA Repair. 2016;41:42–53.

70. Bonura T, Town CD, Smith KC, Kaplan HS. The influence of oxygen on the yield of DNA double-strand breaks in X-irradiated Escherichia coli K-12. Radiat Res. 1975;63:567–77.

71. Tilby MJ, Loverock PS. Measurements of DNA double-strand break yields in E. coli after rapid irradiation and cell inactivation: the effects of inactivation technique and anoxic radiosensitizers. Radiat Res. 1983;96:309–21.

72. Van der Schans G, Blok J. The influence of oxygen and sulphhydryl compounds on the production of breaks in bacteriophage DNA by gamma-rays. Int J Radiat Biol Relat Stud Phys Chem Med. 1970;17:25–38.

73. Pitcher RS, Green AJ, Brzostek A, Korycka-Machala M, Dziadek J, Doherty AJ. NHEJ protects mycobacteria in stationary phase against the harmful effects of desiccation. DNA Repair. 2007;6:1271–6.

74. Dupuy P, Gourion B, Sauviac L, Bruand C. DNA double-strand break repair is involved in desiccation resistance of Sinorhizobium meliloti, but is not essential for its symbiotic interaction with Medicago truncatula. Microbiology. 2017;163:333–42.

75. Charlesworth CT, Deshpande PS, Dever DP, Dejene B, Gomez-Ospina N, Mantri S, et al. Identification of Pre-Existing Adaptive Immunity to Cas9 Proteins in Humans. bioRxiv. 2018 Jan;p. 243345.

76. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:16048.

77. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25.

78. Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, et al. An ontology for microbial phenotypes. BMC Microbiol. 2014;14:294.

79. Tierrafría VH, Mejía-Almonte C, Camacho-Zaragoza JM, Salgado H, Alquicira K, Gama-Castro S, et al. MCO: towards an ontology and unified vocabulary for a framework-based annotation of microbial growth conditions. Bioinformatics. 2018; bty689.

80. Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, et al. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. 2015;34:169–83.

81. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018; eaar4120.