Finding CRISPR's Niche: Oxygen and Temperature Shape the Incidence of Adaptive Immunity

Jake L. Weissman[1], Rohan Laljani[1], William F. Fagan[1], and Philip L. F. Johnson[1,*]

[1]Department of Biology, University of Maryland College Park, MD, USA
[*]plfj@umd.edu

April, 2018

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 and environmental 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 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.

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, 3, 4, 5, 6, 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 ($\sim$ 90%), whereas less than half of mesophilic bacteria have CRISPR ($\sim$ 40%; [8, 9, 10, 11, 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 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 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, 17, 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, 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 intracellular immune systems in general.

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

Table 1: Top 10 variable loadings on the first three principal components of the microbial traits dataset. These three components explain 17%, 10%, and 7% of the total variance, respectively.

Results

Visualizing CRISPR Incidence in Trait Space

We visualized CRISPR incidence in microbial trait space using two unsupervised machine learning 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 CRISPR-encoding and CRISPR-lacking organisms in trait space, despite the fact that no explicit information about CRISPR was included when performing the decompositions.

First, principal components analysis (PCA) of the trait data reveals several well accepted patterns of microbial lifestyle choice and CRISPR incidence. The first principal component (19% variance explained) corresponds broadly to an axis running from host-associated to free-living microbes (Table 1), as observed by others [22, 23]. CRISPR-encoding and CRISPR-lacking microbes are not differentiated along this axis (S1 Fig). We see CRISPR-encoding and CRISPR-lacking organisms beginning to separate along the second (11% variance explained) and third (6% variance explained) principal components (Fig 1). The second component roughly represents a split between extremophilic, energy-stressed species 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 [24].

Second, we visualized the trait data using *t*-distributed stochastic neighbor embedding (t-SNE), which is a nonlinear method that can often pick up on more

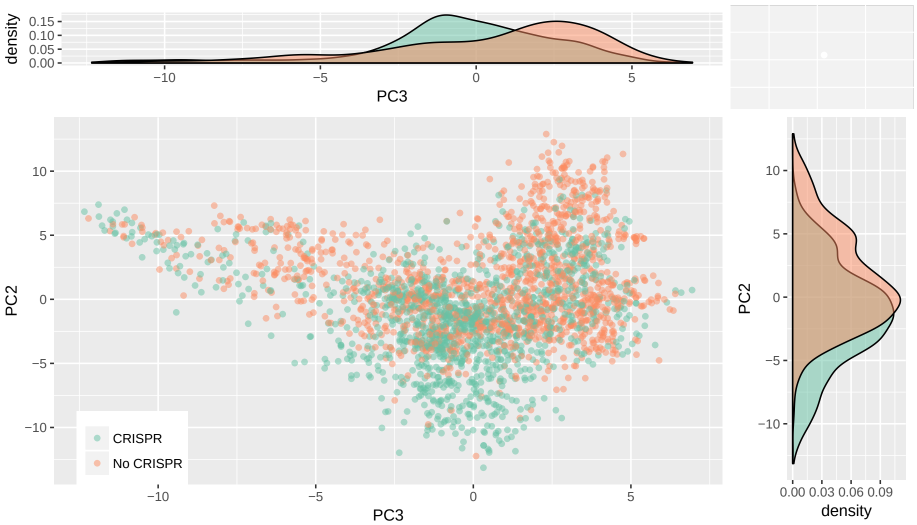


Figure 1: Organisms with CRISPR separate from those without in trait space. The second and third components from a PCA of the microbial traits dataset are shown. CRISPR incidence is indicated by color (green with, orange without), but was not included when constructing the PCA. Notice the separation of organisms with and without CRISPR along both components. Marginal densities along each component are shown to facilitate interpretation. See S1 Fig for the first component.

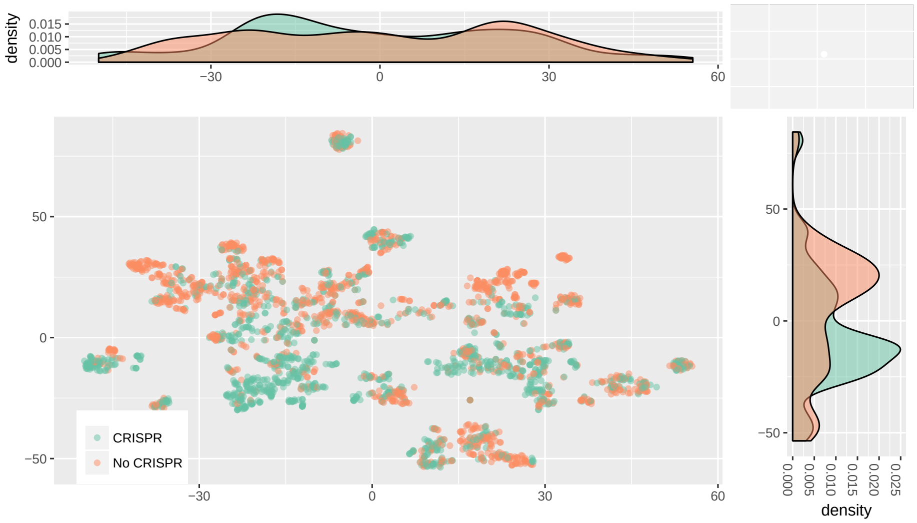


Figure 2: Organisms with CRISPR partially cluster in trait space away from those without. Two dimensional output of t-SNE dimension reduction on dataset. CRISPR incidence is indicated by color (green with, orange without), but was not included when performing dimension reduction. The axes of t-SNE plots have no clear interpretation due to the non-linearity of the transformation.

subtle relationships in a dataset (Fig 2; [25]). 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 (S2 Fig). Surprisingly, the most clearly aligned trait with CRISPR-incidence is having an obligately anaerobic metabolism (S3 Fig).

Predicting CRISPR Incidence

The unsupervised approaches (i.e. uninformed about the outcome variable, CRISPR) we employed above revealed clear patterns linking CRISPR incidence to microbial lifestyle. In order to further explore these patterns, and exploit them for their predictive ability, we applied several supervised prediction (i.e. trained with information about CRISPR incidence) methods to the data.

We tested each of our trained models of CRISPR incidence, using the Proteobacteria as our test set (left out during model training) to determine model accuracy. We emphasize here the choice of Proteobacteria, as they represent a phylogenetically-independent test set from our training set (see Methods). All models showed improved predictive ability over a null model only accounting for the relative frequency of each class in the dataset ($\kappa > 0$; Table 2), indicating that there is some ecological signal in CRISPR incidence. Unsurpris-

	Phylogenetic Correction			Performance		
Model Type	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

Table 2: Predictive ability of models of CRISPR incidence on the Proteobacteria test set. 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 [26]. Observe that the RF model appears to perform best at prediction in general.

ingly, given the difficulty of this task and the noise in the dataset, no model showed overwhelming predictive ability, though the RF model 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.

For the logistic regression models, taking phylogeny into consideration, both via blocked cross validation ($\kappa = 0.168$) and an explicit evolutionary model of trait evolution ($\kappa = 0.188$), improved predictive ability relative to the phylogenetically-uninformed logistic regression approach, though when combined these two corrections appeared to conflict with one another ($\kappa = 0.160$). Our cluster-based approach to phylogenetic correction (MINT) in the partial least squares model framework (sPLS-DA, see Methods) reduced overall predictive ability, but dramatically improved the true positive rate of the prediction (TPR = 0.538), at the cost of an increased false positive rate. The random forest (RF) and phylogenetically-informed RF ensemble models had nearly identical performance.

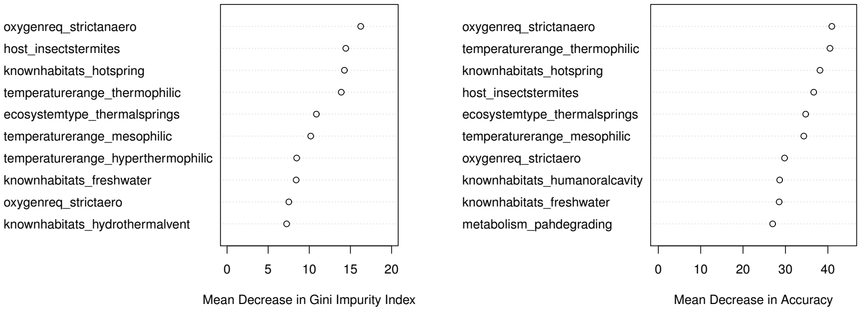


Figure 3: Importance of top ten predictors in the RF model, as measured by the mean decrease in the Gini impurity index or accuracy when that variable is excluded from the model. See S7 Fig for all predictor importances.

We note though, that the ensemble approach gave a much more reliable estimate of predictive ability on the training set (mean $\kappa = 0.258$ predicting on excluded clusters) than the internal estimate automatically generated by the global RF model (out-of-bag estimate, $\kappa = 0.441$). In general, with phylogenetically structured data the error estimates generated by an RF model will be misleading, and the blocked cross-validation approach we employ is one way to correct these estimates.

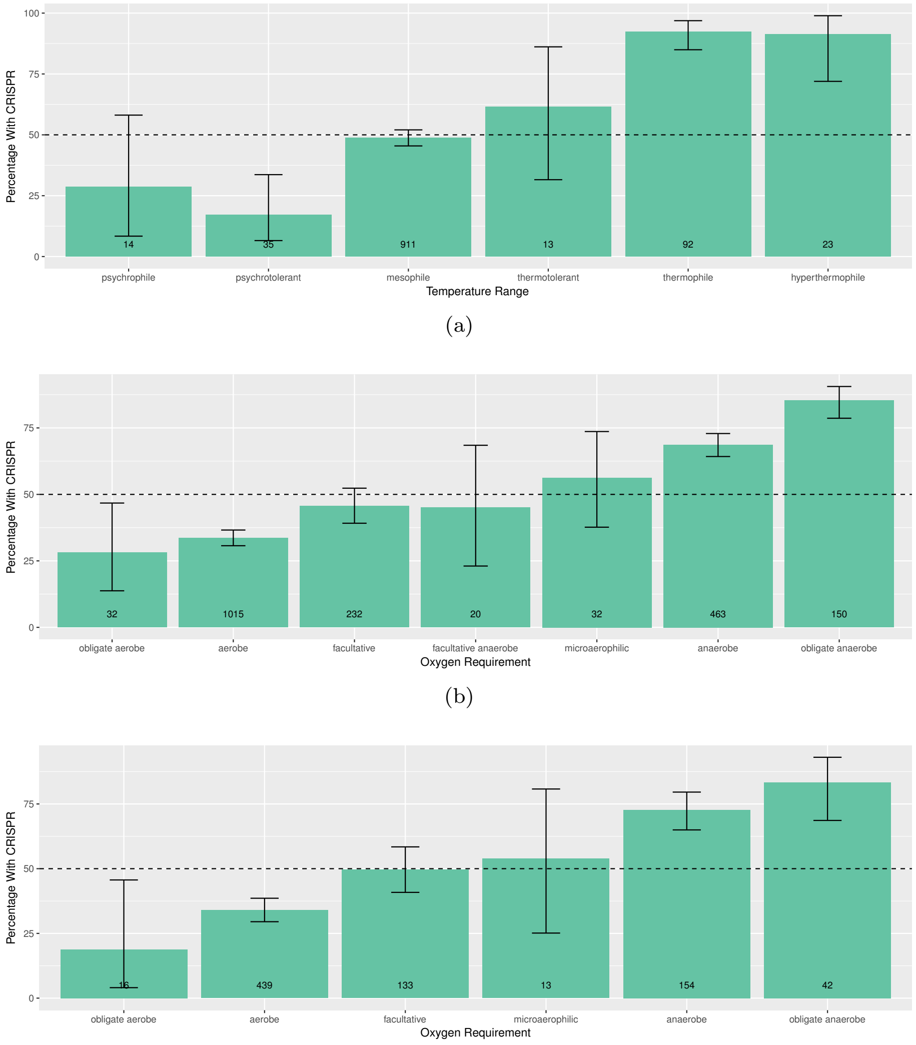
While each modeling framework revealed a distinct set of top predictors of CRISPR incidence, there was broad agreement overall (S1 Table, Fig 3, S4 Fig, and S5 Fig). Keywords indicating a thermophilic lifestyle (e.g. thermophilic, hot springs, hyperthermophilic, thermal springs) appeared across all models 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 logistic regression models that were not parametrically corrected for phylogeny and performed relatively poorly (S1 Table). 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 Fig, S5 Fig, and S6 Fig). 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 were largely in agreement with each other. 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 ($\kappa = 0.191$).

Using 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 psychrotrophic or psychrophilic to make any strong claims, these genomes seem to lack CRISPR most of the time, suggesting that CRISPR incidence decreases continuously as environmental temperatures decrease [10]. We were also able to confirm the 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 4c).

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 [27, 28, 29]. 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 our expectation.

Predicting Without Genomic Data

The ProTraits database, from which we take our trait data, combines various “sources” of text-based and genomic information to make trait predictions [30]. While the inclusion of genomic sources of information considerably improves the trait confidence scores, some of these sources explicitly consider gene presence/absence, and we worried it may lead to circularity in our arguments (e.g. if *cas* gene presence were used to predict a trait, which was then used to predict CRISPR incidence). Therefore we repeated our predictive analyses excluding the “phyletic profile” and “gene neighborhood” sources in ProTraits. We took the maximum confidence scores for having and lacking a trait respectively across all other sources in the database to produce a negative and positive trait score. We integrated these into a single score as described in Methods. We then built an RF model of CRISPR incidence, as this was the highest performing model on the complete dataset. This model had comparable predictive ability ($\kappa = 0.243$). We also found similar predictors to when the full dataset was used (S8 Fig). A notable change is that termite host and PAH degradation no longer appear as important predictors in the model.



(a)

(b)

(c)

Figure 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) The link between oxygen requirement and CRISPR incidence is apparent even when sub-setting to only mesophiles. Error bars are 95% binomial confidence intervals. Total number of genomes in each trait category shown at the bottom of each bar. Categories represented by fewer than 10 genomes were omitted.

Predicting CRISPR Type

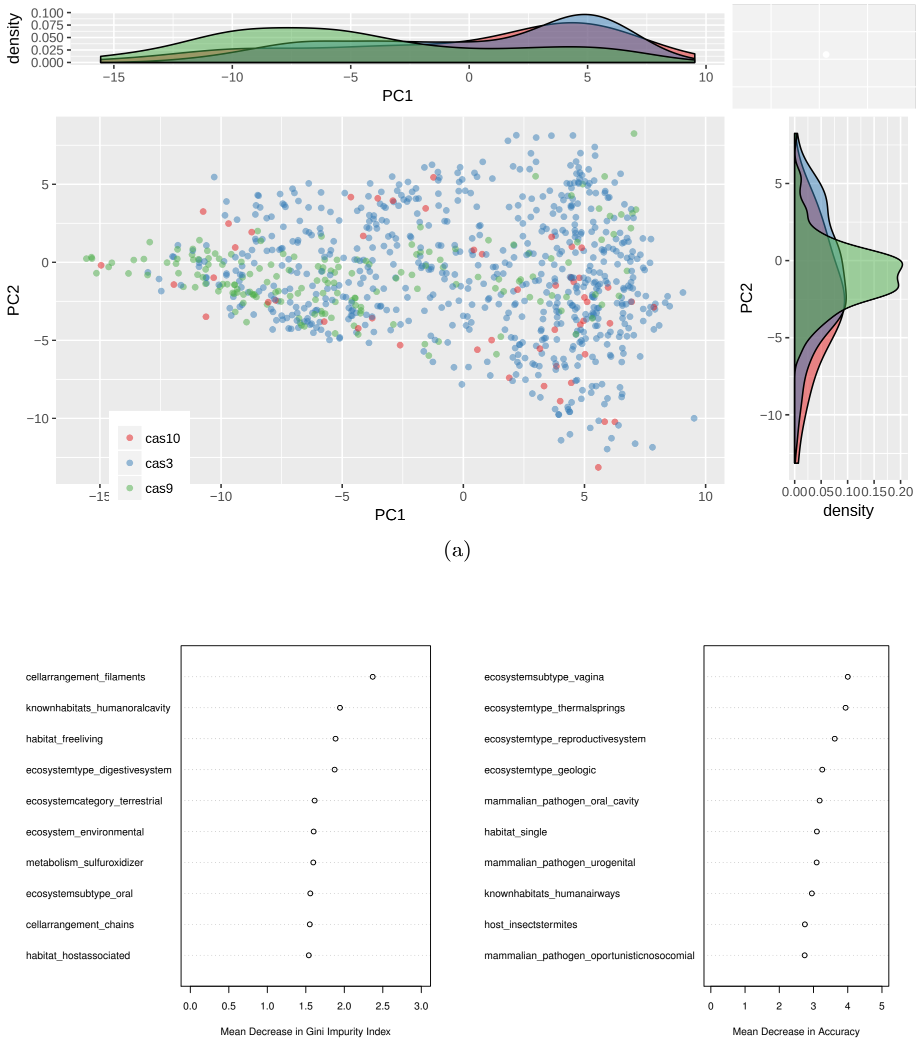
Each CRISPR system type is associated with a signature *cas* targeting gene unique to that type (*cas3*, *cas9*, and *cas10* for type I, II, and III systems respectively). There are many species in the dataset with *cas3* (605), but relatively few with *cas9* (160) and *cas10* (222), suggesting that the ecological correlates of CRISPR incidence that we identify above probably correspond primarily to type I systems. We mapped the incidence of each of these genes onto the PCA we constructed earlier (see S1 Fig 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

The Ku protein is essential to the NHEJ pathway some microbes possess [31, 32]. We searched for the gene encoding this protein and attempted to associate its presence with both microbial lifestyle and CRISPR incidence. Mapping Ku incidence onto our principal components found above we observed a pattern roughly the opposite of that of CRISPR incidence (S9 Fig). That is, Ku was favored in positive values on the second and negative values on the third component, roughly indicating a mesophilic, plant-associated, group-living lifestyle. Additionally, Ku was found in positive regions along the first component, indicating a free-living lifestyle, the opposite of type II CRISPR systems. We built an RF model of Ku incidence, in the same manner as we built one of CRISPR incidence above, and our top predictors appeared to show that the NHEJ pathway is favored in soil-dwelling, spore-forming, aerobic microbes, consistent with expectations of where NHEJ will be most important [33, 34] (S10 Fig). This model predicted Ku incidence well ($\kappa = 0.578$), indicating a clear association between microbial traits and the incidence of NHEJ.

Using our full set of RefSeq genomes, we found a weak negative association between CRISPR and Ku incidence overall (Pearson's correlation, $\rho = -0.012$; $\chi^2 = 15.015$, $p = 1.067 \times 10^{-4}$). Using metadata from NCBI, and restricting only to aerobes this negative association was much stronger ($\rho = -0.250$, $p =$



(a)

(b)

Figure 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 S1 Fig. 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. (2) A variable importance plot from an RF model of *cas9* incidence. Observe that keywords related to a host-associated lifestyle appear many times.

9.109×10^{-16}), while in anaerobes it was nonexistent ($\rho = -0.023$, $p = 0.704$). We found a similar pattern between *cas3* and Ku (aerobes, $p < 2 \times 10^{-16}$; anaerobes, $p = 0.377$), *cas9* and Ku (aerobes, $p = 2.416 \times 10^{-3}$; anaerobes, $p = 0.160$), and *cas10* and Ku (aerobes, $p < 3.16 \times 10^{-12}$; anaerobes, $p = 0.590$), suggesting that CRISPR and NHEJ are generally in conflict when oxygen is present. Nevertheless, anaerobes may have a higher incidence of CRISPR than aerobes overall, in addition to and independent of the effects of Ku incidence (S11 Fig).

Predicting RM Incidence

The majority of genomes in our dataset had at least one RM gene, with 97% of genomes encoding at least one RM-associated restriction enzyme. This agrees with previous results showing that the large majority of prokaryotes have RM systems [35]. We also confirmed the previously observed CRISPR-RM association, with CRISPR incidence being positively associated with the number of restriction enzymes on a genome (6.23 with versus 4.36 without CRISPR, $t = -9.038$, $p < 2.2 \times 10^{-16}$; $m = 0.0676$, $p = 7.212 \times 10^{-13}$, phylogenetic logistic regression; [35]) as well as whether or not a genome has any restriction enzymes ($\chi^2 = 35.065$, $p = 3.189 \times 10^{-9}$; $m = 1.96127$, $p = 1.853 \times 10^{-14}$, phylogenetic logistic regression).

We mapped the incidence of restriction enzymes onto the PCA decomposition of the trait data (Fig S12 Fig). Because very few genomes lacked a restriction enzyme (97), we hesitate to make any strong claims, but the restriction enzyme-lacking organisms seem to tend to be host associated (low values on PC1), thermophilic or anaerobic (low values on PC2), and solitary and motile (high values on PC3). With the exception of PC3, this is the opposite of the patterns we observed in CRISPR incidence. We also found that the number of restriction enzymes was negatively associated with an anaerobic lifestyle ($m = -4.53877$, $p = 2 \times 10^{-16}$, phylogenetic linear regression), and not significantly associated with a thermophilic lifestyle after considering the effects of multiple testing ($m = 1.51063$, $p = 0.03779$, phylogenetic linear regression).

We built 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 ($\kappa = 0.317$), and was able to accurately predict 77% of the enzyme-lacking genomes in the Proteobacteria without requiring a low true positive rate for enzyme incidence (0.898). The only variable that ranked highly in terms of importance that overlapped with our RF model of CRISPR incidence was association with a freshwater habitat (S13 Fig). Overall, the correlation between variable importance scores for the CRISPR and restriction enzyme RF models was low ($\rho = 0.169$ for mean decrease in Gini Impurity Index, $\rho = -0.0487$ for mean decrease in accuracy).

Discussion

We detected a clear association between ecological niche and CRISPR incidence among microbes. In line with previous analyses, temperature range appears to be a strong driver of CRISPR incidence [8, 9, 10]. We lend further support to these previous results by formally controlling for phylogeny using both parametric and non-parametric approaches. We also demonstrate that not only is temperature a predictor of CRISPR incidence, it is one of the most important predictors.

Surprisingly, we find that oxygen requirement appears to be just as important of a predictor 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 DNA repair. We found a clear link between the NHEJ DNA repair pathway and CRISPR incidence. Reactive oxygen species are produced during aerobic metabolism and can cause DNA damage [33], making NHEJ potentially particularly important in aerobes. Type II-A CRISPR systems have been shown to directly interfere with the action of the NHEJ DNA repair pathway in prokaryotes [21]. Thus, if CRISPR interferes with DNA repair, and such repair is more important in aerobes, we would expect CRISPR incidence to be inversely related to the presence of oxygen. While this negative epistatic interaction has only been experimentally observed between NHEJ and the Csn2 protein in type II-A systems, our results suggest that other Cas proteins may also suppress repair, since the interaction was found across system types and was oxygen-dependent in all cases. Alternatively, it is known that the process of CRISPR spacer acquisition prefers free DNA ends [36, 37], so that the cost of CRISPR due to autoimmunity may be heightened in situations where NHEJ is also necessary. This could cause a similar pattern between CRISPR and oxygen requirement, though it is unclear if this preference for breaks generally holds for all CRISPR systems nor if its effects on the rate of autoimmunity would be large. Additionally, if this autoimmunity-based hypothesis were true, we would expect aerobes to uniformly disfavor CRISPR regardless of Ku incidence. While oxygen requirement does have a weak effect on CRISPR incidence independent of Ku, the strong Ku-CRISPR interaction we observe in aerobes but not anaerobes cannot be explained by autoimmunity.

We found no strong link between the incidence or number of RM systems on a genome and a thermophilic or anaerobic lifestyle. In general, the ecological predictors of an RM immune strategy did not correspond to those of a CRISPR immune strategy. This suggests that the factors driving CRISPR incidence are CRISPR-specific, and not shared among intracellular immune strategies in general. This, in turn, partially supports previous work that shows in a theoretical context that CRISPR will be selected against in environments with dense and diverse viral communities, since such hypotheses are CRISPR-specific [11, 12]. In contrast to this conclusion, our results also suggest that host growth rate is not a strong predictor of CRISPR incidence, and that group-living microbes seem to favor CRISPR immunity, calling these prior viral diversity and density based explanations under 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 the local viral community. The disagreement between CRISPR and RM distribution could potentially be due to the high prevalence of RM systems overall, and the fact that these systems may serve other biological functions than immunity [38]. At this point we do not have sufficient empirical evidence to tease apart the mechanisms leading to the observed environmental associations, though others have suggested that thermophilic environments are not distinguished by especially high or low viral diversity [10].

We were also able to show that CRISPR types vary in in terms of the environmental niches 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 [39]. We note that type I and III systems do not appear to have a strong link to host-associated lifestyles.

Here we provide a broad view of how environmental factors shape the evolution of immune strategy. Using only publicly available data, we identified previously unobserved factors influencing the distribution of CRISPR immunity in microbes. More targeted approaches that examine shifts in immune strategy and viral communities along environmental gradients are sure to provide a more fine-grained understanding of how microbial populations adapt to their local pathogenic and abiotic environments. Finally, an increasing number of prokaryotic defense strategies are still being discovered (e.g. [40, 41]), each potentially filling a unique niche in strategy space.

Methods

Data

Trait Data

We downloaded the ProTraits microbial traits database [30] 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 transformed these confidence scores into a single score (p)

approximating the probability that a particular microbe has a particular trait so that a score of one would indicate complete confidence that a microbe has a particular trait, and a score of zero would indicate complete confidence that that microbe lacks that trait

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

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; [42]). There are a number 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. In both cases, we removed these traits from the dataset for increased interpretability (post-imputation).

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. 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 [43]. Additionally, data on the number of CRISPR arrays found among all available RefSeq genomes from a species were taken from Weissman et al. ([44]).

We downloaded the REBASE Gold database of experimentally verified RM proteins and performed blastx searches of our genomes against this database [45, 46]. 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}/number of genomes; Pfam PF02735; [47, 31, 32]). 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 [48]. 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.

Phylogeny

We used Phylosift to locate and align a large set of marker genes (738) found broadly across microbes, generally as a single copy [49, 50]. 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; [51]) to build a phylogeny (S14 Fig).

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) and t-distributed stochastic neighbor embedding (t-SNE, perplexity = 50, 5000 iterations; [25]).

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. 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 [52]. 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 (S15 Fig). The remaining phyla were used to train our models.

We built both linear and nonlinear predictive models. First we performed logistic regression to predict CRISPR incidence among species, using forward subset selection to choose traits to include in the model. We used the minimum mean squared error of prediction under 5-fold cross-validation as our criterion for forward selection, and the minimum BIC as the criterion for choosing model size. Similar to choosing a test set, it is important to take care when dividing the data for cross validation. We performed cross validation both with randomly drawn folds and with blocked folds, where the data were divided into phylogenetically-cohesive chunks [52]. 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); [53, 54]). We note that this method of blocked cross-validation is a non-parametric form of phylogenetic correction, since by testing fit on largely independent sections of the tree we prevent fitting to the underlying phylogenetic structure of the training set. We repeated this analysis using phylogenetic logistic regression to more formally correct for phylogeny (R package phylolm; [26, 55]). While the non-parametric blocking approach is less powerful than the parametric approach used in phylogenetic regression, it has a clear advantage in that it does not require us to specify an underlying evolutionary model.

The trait data exhibit strong multicolinearity (R package mctest; [56, 57]), and so we sought out methods that deal well with this type of data, specifically partial least squares (PLS) regression. We used sparse partial least squared regression discriminant analysis (sPLS-DA) to simultaneously perform feature selection and classification (tune.splsda() and splsda() functions in R package mixOmics; [58, 59]). An extension of sPLS-DA, multivariate integrative (MINT) sPLS-DA, takes into account clustering in the data, where clusters may vary systematically from one another (tune() and mint.splsda() functions in R package mixOmics; [59, 60]). We used MINT sPLS-DA alongside the phylogenetically blocked folds we defined earlier to control for phylogeny. 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 [61], we are comfortable making this assumption.

While regression has the clear advantages of interpretability and computational efficiency, 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 [62]. We built an RF classifier on our training data from 5000 trees (otherwise default settings in R package randomForest; [63]). To prevent fitting to phylogeny, we also took an ensemble approach. Using the phylogenetically blocked folds defined above we fit five 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; [64]). We predicted using our ensemble of forests by choosing the predicted outcome with the greatest total weight.

Acknowledgments

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.

References

[1] Munson-McGee, J. H. *et al.* A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. *The ISME Journal* page 1 (2018).

[2] Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).

[3] Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution* **60**, 174–182 (2005).

[4] Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709–1712 (2007).

[5] Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nature Reviews. Microbiology* **8**, 317–327 (2010).

[6] Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research* **41**, 4360–4377 (2013).

[7] Houte, S. v., Buckling, A. & Westra, E. R. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiology and Molecular Biology Reviews* **80**, 745–763 (2016).

[8] Mojica Francisco J. M., Díez-Villaseñor Cesar, Soria Elena & Juez Guadalupe. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology* **36**, 244–246 (2002).

[9] Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. 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. *Biology Direct* **1**, 7 (2006).

[10] Anderson, R. E., Brazelton, W. J. & Baross, J. A. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiology Ecology* **77**, 120–133 (2011).

[11] Weinberger, A. D., Wolf, Y. I., Lobkovsky, A. E., Gilmore, M. S. & Koonin, E. V. Viral Diversity Threshold for Adaptive Immunity in Prokaryotes. *mBio* **3**, e00456–12 (2012).

[12] Iranzo, J., Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Evolutionary Dynamics of the Prokaryotic Adaptive Immunity System CRISPR-Cas in an Explicit Ecological Context. *Journal of Bacteriology* **195**, 3834–3844 (2013).

[13] Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature Communications* **7**, 10613 (2016).

[14] Makarova, K. S., Wolf, Y. I. & Koonin, E. V. The basic building blocks and evolution of CRISPR–Cas systems. *Biochemical Society Transactions* **41**, 1392–1400 (2013).

[15] Westra, E. R. *et al.* Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology* **25**, 1043–1049 (2015).

[16] Chung, Y. J., Krueger, C., Metzgar, D. & Saier, M. H. Size Comparisons among Integral Membrane Transport Protein Homologues in Bacteria, Archaea, and Eucarya. *Journal of Bacteriology* **183**, 1012–1021 (2001).

[17] Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research* **33**, 3390–3400 (2005).

[18] Ledford, H. Five big mysteries about CRISPR's origins. *Nature News* **541**, 280 (2017).

[19] Bikard, D., Hatoum-Aslan, A., Mucida, D. & Marraffini, L. A. Crispr interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell host & microbe* **12**, 177–186 (2012).

[20] Jiang, W. *et al.* Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. *PLOS Genetics* **9**, e1003844 (2013).

[21] Bernheim, A. *et al.* Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. *Nature Communications* **8**, 2094 (2017).

[22] Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* **6**, 776–788 (2008).

[23] Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

[24] Patterson, A. G. *et al.* Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas Systems. *Molecular Cell* **64**, 1102–1108 (2016).

[25] Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

[26] Ives, A. R. & Garland, T. Phylogenetic Logistic Regression for Binary Dependent Variables. *Systematic Biology* **59**, 9–26 (2010).

[27] Condon, C., Liveris, D., Squires, C., Schwartz, I. & Squires, C. L. rRNA operon multiplicity in Escherichia coli and the physiological implications of rrn inactivation. *Journal of Bacteriology* **177**, 4152–4156 (1995).

[28] Vieira-Silva, S. & Rocha, E. P. C. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLOS Genetics* **6**, e1000808 (2010).

[29] Roller, B. R. K., Stoddard, S. F. & Schmidt, T. M. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nature Microbiology* **1**, 16160 (2016).

[30] Brbić, M. *et al.* The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Research* **44**, 10074–10090 (2016).

[31] Doherty Aidan J., Jackson Stephen P. & Weller Geoffrey R. Identification of bacterial homologues of the Ku DNA repair proteins. *FEBS Letters* **500**, 186–188 (2001).

[32] Aravind, L. & Koonin, E. V. 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 Research* **11**, 1365–1374 (2001).

[33] Karanjawala, Z. E., Murphy, N., Hinton, D. R., Hsieh, C.-L. & Lieber, M. R. Oxygen Metabolism Causes Chromosome Breaks and Is Associated with the Neuronal Apoptosis Observed in DNA Double-Strand Break Repair Mutants. *Current Biology* **12**, 397–402 (2002).

[34] Pitcher, R. S., Brissett, N. C. & Doherty, A. J. Nonhomologous end-joining in bacteria: a microbial perspective. *Annual Review of Microbiology* **61**, 259–282 (2007).

[35] Oliveira, P. H., Touchon, M. & Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Research* **42**, 10618–10631 (2014).

[36] Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015).

[37] Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* **544**, 101–104 (2017).

[38] Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews* **77**, 53–72 (2013).

[39] Charlesworth, C. T. *et al.* Identification of Pre-Existing Adaptive Immunity to Cas9 Proteins in Humans. *bioRxiv* page 243345 (2018).

[40] Goldfarb, T. *et al.* BREX is a novel phage resistance system widespread in microbial genomes. *The EMBO Journal* **34**, 169–183 (2015).

[41] Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* page eaar4120 (2018).

[42] Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).

[43] Biswas, A., Staals, R. H., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).

[44] Weissman, J. L., Fagan, W. F. & Johnson, P. L. F. Is having more than one CRISPR array adaptive? *bioRxiv* page 148544 (2017).

[45] Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).

[46] Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research* **38**, D234–D236 (2010).

[47] Eddy, S. R. Profile hidden Markov models. *Bioinformatics (Oxford, England)* **14**, 755–763 (1998).

[48] Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research* **44**, 6614–6624 (2016).

[49] Lang, J. M., Darling, A. E. & Eisen, J. A. Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices. *PLOS ONE* **8**, e62510 (2013).

[50] Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).

[51] Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

[52] Roberts David R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

[53] Reynolds, A. P., Richards, G., Iglesia, B. d. l. & Rayward-Smith, V. J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475–504 (2006).

[54] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. *cluster: Cluster Analysis Basics and Extensions* (2018). R package version 2.0.7-1.

[55] Ho, L. s. T. & Ané, C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* **63**, 397–408 (2014).

[56] Farrar, D. E. & Glauber, R. R. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics* **49**, 92–107 (1967).

[57] Imdadullah, M., Aslam, M. & Altaf, S. mctest: An R Package for Detection of Collinearity among Regressors. *The R Journal* **8** (2016).

[58] Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).

[59] Rohart, F., Gautier, B., Singh, A. & Cao, K.-A. L. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology* **13**, e1005752 (2017).

[60] Rohart, F., Eslami, A., Matigian, N., Bougeard, S. & Lê Cao, K.-A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* **18**, 128 (2017).

[61] Puigbò, P., Makarova, K. S., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Reconstruction of the evolution of microbial defense systems. *BMC Evolutionary Biology* **17**, 94 (2017).

[62] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

[63] Liaw, A., Wiener, M. & others. Classification and regression by randomForest. *R news* **2** (2002).

[64] Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20**, 37–46 (1960).