CRISPR-Cas is associated with fewer antibiotic resistance genes in bacterial pathogens

Elizabeth Pursey[1]*, Tatiana Dimitriu[1], Fernanda L. Paganelli[2], Edze R. Westra[1] and Stineke van Houte[1]

[1]Environment and Sustainability Institute, Biosciences, University of Exeter, Penryn, Cornwall, United Kingdom

[2]Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands

*Correspondence to: ellie.pursey@gmail.com

Keywords: CRISPR-Cas, horizontal gene transfer, mobile genetic elements, antibiotic resistance, plasmids, integrative conjugative elements

Abstract

The acquisition of antibiotic resistance genes via horizontal gene transfer is a key driver of the rise in multidrug resistance amongst bacterial pathogens. Bacterial defence systems per definition restrict the influx of foreign genetic material, and may therefore limit the acquisition of antibiotic resistance. CRISPR-Cas adaptive immune systems are one of the most prevalent defences in bacteria, found in roughly half of bacterial genomes, but it has remained unclear if and how much they contribute to restricting the spread of antibiotic resistance. We analysed ~40,000 whole genomes comprising the full RefSeq dataset for 11 species of clinically important genera of human pathogens including *Enterococcus*, *Staphylococcus*, *Acinetobacter* and *Pseudomonas*. We modelled the association between CRISPR-Cas and indicators of horizontal gene transfer, and found that pathogens with a CRISPR-Cas system were less likely to carry antibiotic resistance genes than those lacking this defence system. Analysis of the mobile genetic elements targeted by CRISPR-Cas supports a model where this host defence system blocks important vectors of antibiotic resistance. These results suggest a potential "immunocompromised" state for multidrug-resistant strains that may be exploited in tailored interventions that rely on mobile genetic elements, such as phage or phagemids, to treat infections caused by bacterial pathogens.

Introduction

The spread of antibiotic resistance (ABR) genes between bacterial strains and species is of huge global importance; without access to working antibiotics, much of modern medicine is threatened, including surgery, cancer treatment and neonatal care [1]. Opportunistic pathogens in particular often have the ability to take up DNA readily from the environment, through various mechanisms of horizontal gene transfer (HGT) [2]. In some cases, strains are naturally competent and therefore able to uptake extracellular DNA from their environment, as occurs in some clinical isolates of *Acinetobacter baumannii* [3]. Mobile genetic elements (MGEs) also enter the cell through transduction (transfer of DNA by bacteriophages), or conjugation, in which mobile elements are physically transferred between cells through direct cell-cell contact. Conjugative plasmids encode all necessary self-transfer machinery,

while mobilizable plasmids require it to be encoded by another element, but both contribute hugely to the spread of ABR genes. For example, conjugative plasmids are enriched for ABR genes and are able to transfer these resistance determinants across species and genera [4]. In addition, they often harbour integrons, characterised by the *intI1* integrase gene, which enables them to capture ABR and other genes in cassettes that can be spread by the conjugative plasmid [5]. Integrative conjugative elements (ICEs) are widespread in bacteria, and more numerous than conjugative plasmids [6]. These MGEs can integrate into the host chromosome and transmit vertically through cell division, but can also excise themselves and be transferred horizontally through conjugation [7].

Host defences may block the acquisition of MGEs. One of the most prevalent defences is CRISPR-Cas (clustered regularly interspaced short palindromic repeat (CRISPR) loci and CRISPR-associated (*cas*) genes), an adaptive immune system found in in ~30-40% of bacteria [8,9], in which short spacer sequences, derived from incoming DNA, are incorporated at CRISPR loci in between repeat sequences. Functioning as a form of immunological “memory”, these spacer sequences guide their cognate Cas enzymes to cut and destroy any sequence matching this immune record. This enables the cell to defend itself from predation by bacteriophages, and block other sources of potentially costly incoming MGEs.

A number of studies have examined the extent to which CRISPR-Cas restricts HGT, using genomic approaches. In contrast to virus-mediated immunity, which is often circumvented through viral mutation, overcoming immunity to elements such as plasmids often involves loss or inactivation of the CRISPR-Cas system itself [10]. This means evidence for its potential role in blocking these elements may be found in the genomic and phylogenetic distributions of CRISPR-Cas systems in bacteria.

For example, some studies have examined whether CRISPR-Cas systems prevent HGT by looking at genome length. Theoretically, genomes possessing CRISPR-Cas may be shorter if they are blocking acquisition of foreign DNA elements. Evidence for this has been found in *Pseudomonas aeruginosa* [11,12], a species with a large core and accessory genome. More direct interactions between HGT and CRISPR-Cas can be tested using genomic comparisons. However, the outcomes of these studies

have been ambiguous. For example, a study of 1399 bacterial and archaeal genomes did not reveal a correlation between markers of recent HGT and spacer count (used as a proxy for CRISPR-Cas activity) [13]. However, a more recent analysis compared pairs of conspecific genomes with and without CRISPR-Cas as a method to control for relatedness, and found fewer plasmids when CRISPR-Cas was present within 29 genome pairs [14]. When comparing over 100,000 genomes from multiple species, another study revealed a complex situation when looking for correlations between particular ABR gene classes and CRISPR-Cas presence, in which positive and negative relationships were found dependent on species and ABR gene [15].

Research focused on within-species comparisons has often found more compelling evidence for CRISPR-Cas as a barrier to HGT and the spread of ABR genes. In *P. aeruginosa*, in paired within-ST (sequence type) genomes, fewer ICEs and prophages were detected when CRISPR-Cas was present [12], whilst another study found negative associations between the presence of type I-E and I-F systems and particular ABR genes [11]. Studies in Enterococci also support the role of CRISPR-Cas in blocking ABR acquisition, with fewer resistance genes detected in genomes with CRISPR-Cas in a set of 48 strains of *E. faecalis* and 8 strains of *E. faecium* [16]. Further genomic associations between CRISPR-Cas presence and a lack of ABR genes has been detected in *Klebsiella pneumoniae* [17,18], whilst there is evidence in *A. baumannii* that CRISPR-Cas presence is correlated with an absence of plasmids [19].

Here, we sought to expand on existing work using the complete RefSeq dataset for a group of bacterial human pathogens from 11 species (n=39,511), with diverse lifestyles, and CRISPR-Cas system types. We built upon the null hypothesis significance testing used in previous studies, which typically used *t*-tests (or their non-parametric alternatives) or correlation analyses. To do so, we applied linear and generalised linear models (LMs, GLMs) with model selection based on Akaike information criterion (AIC) values, which allows comparison of all candidate models and selection of the model with the best fit to the dataset. We also used Bayesian phylogenetic models to examine within-species relationships, to control for genetic distance between genomes. Both of these model types are advantageous in that

they allow predictions to be made about the dataset based on values of different variables.

We tested both between and within-species associations between ABR gene counts and CRISPR-Cas presence. First, we assessed whether the probability that a genome possesses a CRISPR-Cas system changes based on the presence of plasmids, ICEs and integrons, as well as looking at whether genomes with CRISPR-Cas were shorter. Subsequently, we built within-species models, controlling for genetic distance, to detect associations between specific CRISPR-Cas types and spacer counts, and the quantity of ABR genes accumulated in a genome. Finally, we assessed whether a CRISPR-Cas system possessing spacers that target MGEs that are vectors of ABR genes can effectively reduce the ABR count in the genome.

Methods

Genomes

RefSeq genomes for *P. aeruginosa, A. baumannii, Neisseria meningitidis, Staphylococcus epidermidis, Streptococcus pyogenes, Francisella tularensis, Mycobacterium tuberculosis, Neisseria gonorrhoeae*, *E. faecium* and *E. faecalis* were retrieved from NCBI between December 2020 and January 2021 in nucleotide fasta format using ncbi-genome-download v0.3.0 (https://github.com/kblin/ncbi-genome-download). The strains were selected to include a variety of CRISPR-Cas system types, pathogenic lifestyles (i.e. obligate pathogens such as *F. tularensis, M. tuberculosis* and *N. gonorrhoeae* as well as opportunistic pathogens like *P. aeruginosa*, *S. aureus and S. pyogenes*), and species of significance in the spread of ABR such as *A. baumannii* and *E. faecium*. The species used are summarised in Table 1.

Identification of CRISPR-Cas systems, ABR genes, plasmids, ICEs, and intI1

CRISPR-Cas systems were detected using CRISPRCasTyper [20], and orphan CRISPR arrays were not included in any spacer analysis. CRISPR-Cas systems with "Unknown" predictions were also removed from the dataset. Acquired ABR genes

and plasmid replicons were identified using abricate (https://github.com/tseemann/abricate) with the NCBI AMRFinderPlus [21] and PlasmidFinder [22] databases, both from 19/04/2019 and containing 5386 and 460 sequences, respectively. To detect ICEs and *intI1*, custom BLAST databases were created based on the ICEberg database (retrieved January 2021; Liu et al. 2019), and the NCBI gene entry for class 1 integron integrase *intI1* [NC_019069.1 (99852..100865, complement)], and used with abricate.

Identification of spacer targets and MGEs carrying ABR genes

Spacer sequences were searched against BLAST databases constructed from the aforementioned ABR, ICE and prophage databases as well as a curated plasmid database [24], and the complete RefSeq viral release (retrieved February 2021). Only hits with 100% identity were used for downstream analyses. The total number of spacers per genome was approximated by subtracting 1 from the number of repeats for each array and summing these values, and any spacers with no hits in any database were assigned as “unknown”.

MGEs carrying ABR genes were identified by running a blastn search with default parameters, querying all sequences in search databases used for ICE, plasmid and virus detection against a custom BLAST database built from the aforementioned NCBI AMRFinderPlus database. Subsequently, a percent identity cutoff of 90% was applied to hits. The presence of ABR genes on MGEs was classified in a binomial way, in which ≥ 1 ABR gene was counted as the presence of ABR on the element. In cases where multiple MGEs were targeted by a single spacer, the spacer was classified as targeting an MGE with ABR if at least one of these MGEs carried at least one ABR gene.

Data analysis and statistical modelling

The University of Exeter's Advanced Research Computing facilities were used to carry out this work. Data analyses were conducted in R v4.0.2 using the tidyverse suite of packages [25]. Unless otherwise specified, linear or generalized linear models were fitted using the *lm* or *glm* functions in base R. Linear models allow the

inclusion of multiple predictors in one model, termed fixed effects, to estimate the relationship between these predictors and a response variable. These models generate a formula that calculates the mean of a response variable based on the value of each fixed effect (predictor). Regression coefficients are fitted, which give the slope and the intercept (value of the response variable when the predictor = 0) for each fixed effect. For categorical predictors, a different intercept is fitted for each level of the predictor (e.g. each species in a multispecies model). To fit different slopes for levels of categorical predictor, interaction terms must be fitted between them and a continuous predictor. For example, an interaction between species and the count of ABR genes would allow varying slopes and intercepts for each individual species.

For these analyses, a maximal model was generated and all possible candidate models were compared using the AIC method using *dredge* from the MuMIn package [26]. AIC values assess the fit of a model by looking at the likelihood of a model given the data, penalising for increased number of parameters (as increased complexity of the model increases parameter uncertainty). The model with the lowest AIC value was selected, and no alternative models were within 2 AICs of the best fitting model.

Prediction data frames with 95% confidence intervals were generated using *ggpredict* from the package ggeffects [27], model dispersion was tested and scaled residuals were examined using DHARMa residual diagnostics [28], and the final predictions were visualised with ggplot2, ghibli [29] and cowplot (https://wilkelab.org/cowplot/index.html). Prediction plots were produced only for biologically realistic combinations of fixed effects and response, i.e. those where sufficient data was present in the data set.

Associations between genome length and CRISPR-Cas presence/absence were tested using a linear model with genome length as the response variable and CRISPR-Cas presence/absence and species as fixed effects, with CRISPR-Cas presence/absence $\times$ species as an interaction term.

The effect of various MGEs on the distribution of CRISPR-Cas was modelled using a binomial GLM with CRISPR-Cas presence/absence as the response variable and counts of ICEs, plasmid replicons, ABR genes and *intI1* copies, as well as species, as fixed effects. In addition, we fitted an interaction between species and every other fixed effect. Prediction data frames were created up to the maximum number of each MGE detected in each species (unless this value was an outlier, in which case a more biologically informative upper limit was set).

The relationship between spacers targeting MGEs with and without ABR and the count of ABR genes was modelled using a Bayesian Poisson GLM in MCMCglmm. The default weakly informative priors were used, and models were run for 20,000 iterations with a burn in period of 10,000 iterations and a thinning interval of 5. Two models were run, with the number of ABR genes as the response variable, species as a fixed effect, and either the counts of spacers targeting MGEs with ABR or targeting MGEs without ABR as an additional fixed effect. Prediction data frames were created up to the maximum number of spacers detected targeting MGEs with and without ABR.

Construction of trees and genetic distance-controlled single-species models

The package mashtree [30] with mash v2.0.0 [31] was used to create neighbour-joining trees for each species based on whole genomes. Genomes were sketched using mash with default parameters prior to using mashtree. Mash distance was calculated for all isolates relative to the first genome entry for the group, in order to remove outliers with a distance > 0.1, which are likely to have had their species misclassified. Trees were subsequently rooted to outgroups and ultrametric trees were produced with the ape package in R [32] using the chronos function with smoothing parameters (λ) of 1 and 10. The “correlated”, “discrete” and “relaxed” models were used for each value of λ, which vary in the extent to which they permit adjacent parts of the tree to evolve at different rates. The tree with the highest log-likelihood for each species was used. Due to the large number of genomes and low prevalence of CRISPR-Cas, we did not perform this analysis for *S. aureus*.

The MCMCglmm package [33] was used to run Bayesian GLMs incorporating trees. The default weakly informative priors were used, and all models were run for 30,000 iterations with a burn in period of 20,000 iterations and a thinning interval of 5. Poisson GLMs were run for each species, with count of ABR genes as the response variable and CRISPR-Cas type or spacer count (in which case genomes with no CRISPR-Cas were removed) as a fixed effect. In cases where more than one CRISPR-Cas type was present, CRISPR-Cas type × total spacers was fit as an interaction. The small sample size for *S. epidermidis* with CRISPR-Cas systems (n = 54), meant we chose not to model the association between spacer count and ABR genes for this species. Prediction data frames for spacer counts were created up to the maximum count of spacers detected for that species, with the exception of *S. pyogenes,* for which the top value was an outlier at 33 so the next highest value of 17 was taken.

Code availability and reproducibility

The snakemake pipeline used to calculate genome lengths and detect CRISPR-Cas systems, ABR genes, *intI1*, ICEs and plasmid replicons, as well as R scripts used for producing trees and statistical analysis for this work are available at https://github.com/elliekpursey/crispr-pathogens. Snakemake v5.18.1 [34] and Python v3.8.3 with Biopython v1.78 [35] were used for this work.

Results

Distribution of CRISPR-Cas systems, ABR and mobile elements across species

Genome analyses of ~40,000 genomes of *Pseudomonas aeruginosa, Acinetobacter baumannii, Neisseria meningitidis, Staphylococcus aureus, Staphylococcus epidermidis, Streptococcus pyogenes, Francisella tularensis, Mycobacterium tuberculosis, Neisseria gonorrhoeae*, *Enterococcus faecium* and *Enterococcus faecalis* revealed a large variety of CRISPR-Cas system types, ABR genes, ICEs and plasmids (Supp. Fig. 1). CRISPR-Cas systems were identified in every species except *N. gonorrhoeae*, which was therefore excluded from subsequent analyses.

Overall, type I-C, I-E, I-F, II-A, II-B, II-C, III-A, IV and V-A systems were represented across the dataset and ABR genes were detected in every species. However, *F. tularensis* was positive only for its native β-lactamase (FTU-1) [36], *M. tuberculosis* isolates were almost only ever positive for its intrinsic *blaA*, *aac(2')-Ic* and *erm(37)* genes, and *N. meningitidis* had a maximum of 1 ABR gene, so these species were not considered in downstream analyses. The proportion of genomes with CRISPR-Cas varied across species from around 0.75 for *M. tuberculosis* to around 0.005 for *S. aureus* (Supp. Fig. 2). The number of repeats per genome also varied considerably between system type and species (Supp. Fig. 3), reaching as high as ~200 in I-F systems in *A. baumannii*, and ~150 in III-A systems of *M. tuberculosis*. However, for all other types this number was typically below 50.

Plasmids were detected in >80% of genomes for species in the *Staphylococcus* and *Enterococcus* genera, and were also prevalent in *S. pyogenes* (~25%), *P. aeruginosa* (10%), and *A. baumannii* (~3%) (Supp. Fig.1). As genomes were assembled to varying levels within the dataset, it is assumed some plasmids will have been missed using this method. ICEs were also common in these species, except for *A. baumannii*, ranging from 93% of *E. faecalis* genomes having at least one ICE to 15% of *E. faecium* genomes. Finally, *intI1* was detected in ~20% of *P. aeruginosa* and *A. baumannii* genomes.

No clear association between genome size and CRISPR-Cas presence/absence

Consistent with previous work, *P. aeruginosa* genomes with CRISPR-Cas were on average smaller than those without the system. This was in fact the largest difference we saw, with predicted lengths of 6,579,416 ± 2885 bp and 6,692,831 ± 2601 bp for genomes with and without CRISPR-Cas, respectively (Supp. Fig. 4). However, the opposite was true in other species such as *A. baumannii* and *S. aureus*. For example, an *A. baumannii* genome with CRISPR-Cas was predicted to be 4,028,659 ± 5523 bp long, whilst one without had a predicted length of 3,976,319 ± 2181 bp, a difference of ~50,000 bp overall. Typically, differences between CRISPR-Cas positive and negative genomes were < 55,000 bp.

CRISPR-Cas presence is associated with fewer acquired ABR genes and mobile elements, but more ICEs

We modelled the association between CRISPR-Cas presence or absence and the counts of ABR genes, as well as their vectors. Predictions are presented for elements which are found in each species in Figure 1 and all model estimates are presented in Supplementary Table 4. Across species, every additional ABR gene led to a 0.08 reduction in the probability of having a CRISPR-Cas system (p = 9.2E-18). The direction and strength of this trend varied by species, with positive associations between probability of CRISPR-Cas presence and ABR gene count in *S. epidermidis* and *P. aeruginosa* (Fig. 1A). CRISPR-Cas presence appeared to have little association with plasmid replicon count across species (Fig. 1B), whilst for *intI1*, a negative association was present in *P. aeruginosa* but not *A. baumannii*. For genomes possessing ICEs, positive associations were detected between ICE count and ABR gene count in *P. aeruginosa* and *S. pyogenes*.

CRISPR-Cas presence is associated with fewer ABR genes across system types in individual species, controlling for genetic distance

To further explore the link between ABR genes and CRISPR-Cas, we made within-species models controlling for genetic distance, to look at the effect of CRISPR-Cas type and spacer count (Fig. 2). The overall trends detected are largely similar to those presented for across species comparisons in Fig. 1, in that there is generally a negative association between the presence of a CRISPR-Cas system and ABR counts. Across species, lacking a CRISPR-Cas system was associated with a higher predicted count of ABR genes (Fig. 2a). In some cases the size of this effect is striking; for example, an *E. faecium* genome lacking CRISPR-Cas is predicted to have 12 ABR genes compared to only 5 ABR genes for one possessing a type II-C system. However, in most cases this effect is more modest, with a reduction of 1 or 2 ABR genes when CRISPR-Cas is present. There are, however, two notable exceptions to this trend. Firstly, *Streptococcus pyogenes* type II-A systems were associated with a small but nonsignificant increase (0.44, 95% CI = 0.35 - 0.56) in the number of predicted ABR genes relative to genomes without CRISPR-Cas (0.33,

353 95% CI = 0.28 - 0.37). Secondly, both type I-C and IV systems are consistently
354 associated with an increased count of ABR genes.
355
356 Next, we modelled the number of ABR genes in response to the count of spacers
357 (for those genomes that possess CRISPR-Cas; Fig. 2b). Here we did not find strong
358 associations for most CRISPR-Cas system types. Exceptions were the type I-F
359 system of *A. baumannii,* which showed a dramatic reduction in ABR gene count with
360 increasing spacer count, and the type I-C system of *P. aeruginosa*, which
361 interestingly showed the opposite trend. We did not model this association for type
362 IV spacers due to the limited number of genomes (n = 20) in this category.
363
364 To expand on this dataset, we looked at the targets of all spacers by searching them
365 against virus, ICE and plasmid databases (Supp. Fig. 5). As expected, the majority
366 of the ~285,000 spacers detected (90%) did not have a match to any of these
367 elements. Overall, the majority of known spacer targets are viruses (n = 24,402).
368 However, we also detect many spacers targeting ICEs (n = 3,002) and plasmids (n =
369 1,416).
370
371 *Presence of spacers targeting ABR-carrying MGEs is negatively associated with*
372 *ABR genes*
373
374 As we detected many spacers matching potential vectors of ABR (n = 1,948), we
375 wanted to determine whether there was an association between the presence of
376 these spacers and an absence of ABR genes. We modelled the number of ABR
377 genes in response to the number of spacers targeting MGEs that were positive or
378 negative for ABR genes. The number of predicted ABR genes per genome was
379 consistently lower when spacers targeted MGEs carrying ABR than when they
380 targeted those that did not carry ABR, a trend that was universal across species (Fig.
381 3).
382
383 **Discussion**
384
385 We found little evidence for genome length as a marker of CRISPR-Cas blocking
386 HGT in this collection of bacterial pathogens. Genome reduction is a process that is

common in pathogens relative to their less-pathogenic relatives [37]. Bacteria may streamline their genomes in order to save energetic resources and remove genes that have become deleterious, or reductions in genome length may happen as a result of many other evolutionary processes that occur during the transition to pathogenicity. For example, in *P. aeruginosa*, strains adapted to the cystic fibrosis lung are frequently auxotrophic, as they are able to utilise amino acids produced by the host [38]. The extent of mobile DNA in genomes also reduces during the transition from facultative to obligate lifestyles, potentially due to reduced opportunities for their spread in more intracellular niches [39]. Therefore, this measure may be detecting potential adaptation to a pathogenic lifestyle or particular ecological environments, as well as a concurrent change in HGT frequencies.

However, we found compelling evidence that CRISPR-Cas can block the transfer of MGEs. We focused on the spread of ABR and its potential vectors due to their global importance, and identified a negative association between the count of ABR genes and the probability a genome will have a CRISPR-Cas system, which was repeatable across most CRISPR-Cas system types and species. Further supporting this hypothesis, we found genomes with spacers targeting vectors of ABR did indeed have fewer ABR genes. We did not detect strong trends for plasmid replicons, potentially because 1) genomes were assembled to varying levels and sequenced using different technologies within our dataset and many plasmids will have been missed and 2) the database we used was built based on plasmids found in *Enterobacteriaceae*, and has been shown to fail to predict plasmids in many species [40].

When looking at specific CRISPR-Cas types within species, we found that ABR genes were lower in frequency if CRISPR-Cas was present in most cases. Two exceptions were type I-C and IV systems, which were associated with more ABR genes. These system types may be driving the positive association we see between CRISPR-Cas presence and ABR genes in our multi-species model for *P. aeruginosa*. This result is also interesting in light of the fact that both of these system types are present on MGEs. Type I-C systems occur on conjugative elements in *P. aeruginosa* and have been found to correlate positively with the presence of particular ABR genes [11], whilst type IV CRISPR-Cas systems are usually found on

plasmids, where they can mediate inter-plasmid competition [41]. CRISPR-Cas system types that are more mobile may co-occur with ABR genes found on mobile elements more frequently. Interestingly, the species for which we detected a positive association between ICEs and CRISPR-Cas presence in our multi-species model (*P. aeruginosa* and *S. pyogenes*) were also the only species to possess type I-C systems. We also saw positive associations between the count of ABR genes and the number of type I-C spacers in *P. aeruginosa,* suggesting that this system may even allow the genome to accumulate more ABR genes, and could have a role in competition between MGEs rather than host genome defence. However, much remains to be learned about these widespread but relatively understudied type I-C systems to fully explain this trend [42].

We do not see strong negative associations between the total count of spacers and the presence of ABR genes across most species and CRISPR-Cas system types. As the majority of spacer targets that we could identify were viruses, which are rarely vectors of ABR [43], the large number of anti-phage spacers and comparative rarity of spacers targeting ABR vectors may obscure any trend we would otherwise see here. Alternatively, the lack of clear association may be explained by the tendency for arrays to include more inactive spacers as they increase in size, as the most recently acquired spacers are most active [44]. Therefore, genomes with more spacers and larger arrays may include more spacers with no activity in removing potential vectors of ABR. An exception to this trend was *A. baumannii*, which had a strong reduction in ABR gene count with increasing spacer count in its type I-F system. Our spacer target analysis found that, unusually, more spacers in this species targeted plasmids than viruses; therefore, there could be some specialisation towards blocking plasmid acquisition in this system, a trend which has also been suggested for the type I-F system in *E. coli* [45].

We looked for spacers that targeted vectors of ABR and found they were fairly common. An increase in the count of spacers targeting vectors of ABR had a striking effect on reducing the predicted count of ABR genes. Particularly compared with the weaker association between the count of spacers targeting vectors without ABR and ABR genes, this evidence strongly suggests CRISPR-Cas is an important barrier to vectors of ABR, thereby blocking the acquisition of ABR genes themselves. Many

ICEs, phages and plasmids remain to be characterised, and once our databases of these elements expand, this hypothesis will become more easily testable using genomic data.

This study supports the hypothesis that CRISPR-Cas system loss occurs where ABR acquisition is beneficial. In this case, the presence of CRISPR-Cas may be selected against in populations that are exposed to antibiotics, where the survival benefits of acquiring ABR genes outweigh the cost of phage predation or MGE maintenance. This has important implications for novel treatments, as multidrug-resistant bacteria may represent an immunocompromised population. In recent years, increasing attention has been given to antibiotic alternatives such as phage therapy [46,47] and CRISPR-Cas antimicrobials delivered using phage-like or plasmid elements [48,49]. Our results support the use of these interventions in cases where antibiotics do not work, as they are likely to be most effective in multidrug-resistant bacteria.

Our results also raise the question of why such trends have not been detected in previous work asking similar questions, particularly when looking across species. The strength of our work is primarily the use of more powerful and appropriate statistical tests, as well as controls for genetic distance. We selected the model with the best fit to the data, accounting and controlling for the effects of multiple predictors under one modelling framework as well as testing interactions between predictors. For example, we could test how the probability of a genome having a CRISPR-Cas system varies based on the number of ABR genes it has, whilst controlling for the number of ICEs, integrons and plasmid replicons in the genome overall. We could also control for interactions; for example, this allowed us to detect both positive and negative relationships between CRISPR-Cas presence and genome length for different species within the same model, avoiding some of the pitfalls of multiple testing. Such approaches are routinely used in evolution and ecology studies to account for the variability of datasets [50], and while suitable for these types of genomic comparisons, they are less commonly used in this context. Importantly, we also have access to a much larger dataset of genomes and improved computational tools since the publication of some previous work, which allows us greater statistical power to uncover associations between CRISPR-Cas and ABR genes.

Conclusion

We found that looking both across and within-species, CRISPR-Cas system presence was associated with a reduction in counts of ABR genes. In addition, where spacers targeted MGEs that carry ABR, a concurrent reduction in ABR genes was seen. These results have promising implications for the delivery of novel technologies to combat ABR such as phage therapy and CRISPR-Cas antimicrobials, which rely on bypassing bacterial immune systems to kill cells. In fact, they may be ideal for this purpose, if the most multidrug-resistant strains are also the most immunocompromised.

Figure legends

Figure 1 – Predicted probability of CRISPR-Cas presence (y-axis) from binomial GLMs for A) ABR gene count and numbers of the potential ABR vectors B) plasmids, C) *intI1* (integrons) and D) ICEs. Predictions are presented for elements which are found in each species. Error bars show 95% confidence intervals.

Figure 2 – Prediction plots from Bayesian Poisson GLMs of ABR gene counts for A) CRISPR-Cas type and B) count of spacers. Each column represents an individual species. Error bars show 95% credible intervals.

Figure 3 – Prediction plot overlaying two Bayesian Poisson GLMs of ABR gene counts according to counts of spacers (with known targets) targeting MGEs with and without ABR, faceted by species. Error bars show 95% credible intervals.

Table legend

Table 1 – Summary of species included

Acknowledgements

We would like to thank MD Sharma for his invaluable help with accessing computational resources needed for this work.

Funding

EP is supported by a PhD studentship equally funded by the European Research Council (ERC-STG-2016-714478 – EVOIMMECH) and the College of Life and Environmental Sciences, University of Exeter. SVH acknowledges support from the Biotechnology and Biological Sciences Research Council (BB/S017674/1 and BB/R010781/10). FLP acknowledges support from Utrecht Exposome Hub of Utrecht Life Sciences (www.uu.nl/exposome), funded by the Executive Board of Utrecht University.

Competing Interests

We have no competing interests.

Table

Species	Abbreviation used	RefSeq genomes (n)
Pseudomonas aeruginosa	PA	5360
Acinetobacter baumannii	AB	4864
Neisseria meningitidis	NM	1967
Staphylococcus epidermidis	SE	908
Staphylococcus aureus	SA	12148
Streptococcus pyogenes	SP	2106
Francisella tularensis	FT	675
Mycobacterium tuberculosis	MT	6688
Neisseria gonorrhoeae	NG	723
Enterococcus faecium	EFm	2247
Enterococcus faecalis	EFs	1825
	Total	39511

Bibliography

1. Laxminarayan R, Matsoso P, Pant S, Brower C, Røttingen J-A, Klugman K, Davies S. 2016 Access to effective antimicrobials: a worldwide challenge. *The Lancet* **387**, 168–175. (doi:10.1016/S0140-6736(15)00474-2)

2. Von Wintersdorff CJH, Penders J, Van Niekerk JM, Mills ND, Majumder S, Van Alphen LB, Savelkoul PHM, Wolffs PFG. 2016 Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7**, 1–10. (doi:10.3389/fmicb.2016.00173)

3. Traglia GM, Chua K, Centrón D, Tolmasky ME, Ramírez MS. 2014 Whole-Genome Sequence Analysis of the Naturally Competent *Acinetobacter baumannii* Clinical Isolate A118. *Genome Biol. Evol.* **6**, 2235–2239. (doi:10.1093/gbe/evu176)

4. Che Y, Yang Y, Xu X, Břinda K, Polz MF, Hanage WP, Zhang T. 2021 Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proc. Natl. Acad. Sci.* **118**, e2008731118. (doi:10.1073/pnas.2008731118)

5. Vrancianu CO, Popa LI, Bleotu C, Chifiriuc MC. 2020 Targeting Plasmids to Limit Acquisition and Transmission of Antimicrobial Resistance. *Front. Microbiol.* **11**, 761. (doi:10.3389/fmicb.2020.00761)

6. Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. 2011 The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genet.* **7**, e1002222. (doi:10.1371/journal.pgen.1002222)

7. Botelho J, Schulenburg H. 2021 The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. *Trends Microbiol.* **29**, 8–18. (doi:10.1016/j.tim.2020.05.011)

8. Hatoum-Aslan A, Marraffini LA. 2014 Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens. *Curr. Opin. Microbiol.* **17**, 82–90. (doi:10.1016/j.mib.2013.12.001)

9. Bernheim A, Bikard D, Touchon M, Rocha EPC. 2019 Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. *Nucleic Acids Res.* , gkz1091. (doi:10.1093/nar/gkz1091)

10. Jiang W, Maniv I, Arain F, Wang Y, Levin BR, Marraffini LA. 2013 Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. *PLoS Genet.* **9**, e1003844. (doi:10.1371/journal.pgen.1003844)

11. van Belkum A *et al.* 2015 Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. **6**, 13.

12. Wheatley RM, MacLean RC. 2020 CRISPR-Cas systems restrict horizontal gene transfer in Pseudomonas aeruginosa. *ISME J.* (doi:10.1038/s41396-020-00860-3)

582 13. Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV. 2015 No
583 evidence of inhibition of horizontal gene transfer by CRISPR–Cas on
584 evolutionary timescales. *ISME J.* **9**, 2021–2027. (doi:10.1038/ismej.2015.20)

585 14. O'Meara D, Nunney L. 2019 A phylogenetic test of the role of CRISPR-Cas in
586 limiting plasmid acquisition and prophage integration in bacteria. *Plasmid* **104**,
587 102418. (doi:10.1016/j.plasmid.2019.102418)

588 15. Shehreen S, Chyou T, Fineran PC, Brown CM. 2019 Genome-wide correlation
589 analysis suggests different roles of CRISPR-Cas systems in the acquisition of
590 antibiotic resistance genes in diverse species. *Philos. Trans. R. Soc. B Biol. Sci.*
591 **374**, 20180384. (doi:10.1098/rstb.2018.0384)

592 16. Palmer KL, Gilmore MS. 2010 Multidrug-Resistant Enterococci Lack CRISPR-
593 cas. *mBio* **1**, e00227-10. (doi:10.1128/mBio.00227-10)

594 17. Mackow NA, Shen J, Adnan M, Khan AS, Fries BC, Diago-Navarro E. 2019
595 CRISPR-Cas influences the acquisition of antibiotic resistance in *Klebsiella*
596 *pneumoniae*. *PLOS ONE* **14**, e0225131. (doi:10.1371/journal.pone.0225131)

597 18. Li H-Y, Kao C-Y, Lin W-H, Zheng P-X, Yan J-J, Wang M-C, Teng C-H, Tseng C-
598 C, Wu J-J. 2018 Characterization of CRISPR-Cas Systems in Clinical *Klebsiella*
599 *pneumoniae* Isolates Uncovers Its Potential Association With Antibiotic
600 Susceptibility. *Front. Microbiol.* **9**. (doi:10.3389/fmicb.2018.01595)

601 19. Mangas EL, Rubio A, Álvarez-Marín R, Labrador-Herrera G, Pachón J, Pachón-
602 Ibáñez ME, Divina F, Pérez-Pulido AJ. 2019 Pangenome of *Acinetobacter*
603 *baumannii* uncovers two groups of genomes, one of them with genes involved in
604 CRISPR/Cas defence systems associated with the absence of plasmids and
605 exclusive genes for biofilm formation. *Microb. Genomics* **5**.
606 (doi:10.1099/mgen.0.000309)

607 20. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ. 2020
608 CRISPRCasTyper: Automated Identification, Annotation, and Classification of
609 CRISPR-Cas Loci. *CRISPR J.* **3**, 462–469. (doi:10.1089/crispr.2020.0059)

610 21. Feldgarden M *et al.* 2019 Validating the AMRFinder Tool and Resistance Gene
611 Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations
612 in a Collection of Isolates. *Antimicrob. Agents Chemother.* **63**.
613 (doi:10.1128/AAC.00483-19)

614 22. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L,
615 Møller Aarestrup F, Hasman H. 2014 *In Silico* Detection and Typing of Plasmids
616 using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrob.*
617 *Agents Chemother.* **58**, 3895–3903. (doi:10.1128/AAC.02412-14)

618 23. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, Tai C, Deng Z, Ou H-Y. 2019 ICEberg 2.0:
619 an updated database of bacterial integrative and conjugative elements. *Nucleic*
620 *Acids Res.* **47**, D660–D665. (doi:10.1093/nar/gky1123)

621 24. Brooks L, Kaze M, Sistrom M. 2019 A Curated, Comprehensive Database of
622 Plasmid Sequences. *Microbiol. Resour. Announc.* **8**, e01325-18, e01325-18.
623 (doi:10.1128/MRA.01325-18)

624 25. Wickham H *et al.* 2019 Welcome to the Tidyverse. *J. Open Source Softw.* **4**,
625 1686. (doi:10.21105/joss.01686)

626 26. Barton K. 2009 MuMIn□: multi-model inference. *Httpr-Forg.-Proj.*

627 27. Lüdecke D. 2018 ggeffects: Tidy Data Frames of Marginal Effects from
628 Regression Models. *J. Open Source Softw.* **3**, 772. (doi:10.21105/joss.00772)

629 28. Hartig F. 2020 *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level /*
630 *Mixed) Regression Models.* See http://florianhartig.github.io/DHARMa/.

631 29. Henderson E. 2020 *ghibli: Studio Ghibli Colour Palettes.* See https://CRAN.R-
632 project.org/package=ghibli.

633 30. Katz LS, Griswold T, Morrison SS, Caravas JA, Zhang S, Bakker HC den, Deng
634 X, Carleton HA. 2019 Mashtree: a rapid comparison of whole genome sequence
635 files. *J. Open Source Softw.* **4**, 1762. (doi:10.21105/joss.01762)

636 31. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S,
637 Phillippy AM. 2016 Mash: fast genome and metagenome distance estimation
638 using MinHash. *Genome Biol.* **17**, 132. (doi:10.1186/s13059-016-0997-x)

639 32. Paradis E, Claude J, Strimmer K. 2004 APE: Analyses of Phylogenetics and
640 Evolution in R language. *Bioinformatics* **20**, 289–290.
641 (doi:10.1093/bioinformatics/btg412)

642 33. Hadfield JD. 2010 MCMC Methods for Multi-Response Generalized Linear Mixed
643 Models: The MCMCglmm *R* Package. *J. Stat. Softw.* **33**.
644 (doi:10.18637/jss.v033.i02)

645 34. Köster J, Rahmann S. 2012 Snakemake—a scalable bioinformatics workflow
646 engine. *Bioinformatics* **28**, 2520–2522. (doi:10.1093/bioinformatics/bts480)

647 35. Cock PJA *et al.* 2009 Biopython: freely available Python tools for computational
648 molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423.
649 (doi:10.1093/bioinformatics/btp163)

650 36. Antunes NT, Frase H, Toth M, Vakulenko SB. 2012 The Class A β-Lactamase
651 FTU-1 Is Native to Francisella tularensis. *Antimicrob. Agents Chemother.* **56**,
652 666–671. (doi:10.1128/AAC.05305-11)

653 37. Moran NA. 2002 Microbial Minimalism: Genome Reduction in Bacterial
654 Pathogens. *Cell* **108**, 583–586. (doi:10.1016/S0092-8674(02)00665-7)

655 38. Weinert LA, Welch JJ. 2017 Why Might Bacterial Pathogens Have Small
656 Genomes? *Trends Ecol. Evol.* **32**, 936–947. (doi:10.1016/j.tree.2017.09.006)

657 39. Newton ILG, Bordenstein SR. 2011 Correlations Between Bacterial Ecology and
658 Mobile DNA. *Curr. Microbiol.* **62**, 198–208. (doi:10.1007/s00284-010-9693-3)

659 40. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. 2017 On the
660 (im)possibility of reconstructing plasmids from whole-genome short-read
661 sequencing data. *Microb. Genomics* **3**. (doi:10.1099/mgen.0.000128)

662 41. Pinilla-Redondo R, Mayo-Muñoz D, Russel J, Garrett RA, Randau L, Sørensen
663 SJ, Shah SA. 2020 Type IV CRISPR–Cas systems are highly diverse and
664 involved in competition between plasmids. *Nucleic Acids Res.* **48**, 2000–2012.
665 (doi:10.1093/nar/gkz1197)

666 42. León LM, Park AE, Borges AL, Zhang JY, Bondy-Denomy J. 2021 Mobile
667 element warfare via CRISPR and anti-CRISPR in *Pseudomonas aeruginosa*.
668 *Nucleic Acids Res.* **49**, 2114–2125. (doi:10.1093/nar/gkab006)

669 43. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit M-A. 2017 Phages
670 rarely encode antibiotic resistance genes: a cautionary tale for virome analyses.
671 *ISME J.* **11**, 237–247. (doi:10.1038/ismej.2016.90)

672 44. McGinn J, Marraffini LA. 2016 CRISPR-Cas Systems Optimize Their Immune
673 Response by Specifying the Site of Spacer Integration. *Mol. Cell* **64**, 616–623.
674 (doi:10.1016/j.molcel.2016.08.038)

675 45. Aydin S, Personne Y, Newire E, Laverick R, Russell O, Roberts AP, Enne VI.
676 2017 Presence of Type I-F CRISPR/Cas systems is associated with antimicrobial
677 susceptibility in *Escherichia coli*. *J. Antimicrob. Chemother.* **72**, 2213–2218.
678 (doi:10.1093/jac/dkx137)

679 46. Lin DM, Koskella B, Lin HC. 2017 Phage therapy: An alternative to antibiotics in
680 the age of multi-drug resistance. *World J. Gastrointest. Pharmacol. Ther.* **8**, 162–
681 173. (doi:10.4292/wjgpt.v8.i3.162)

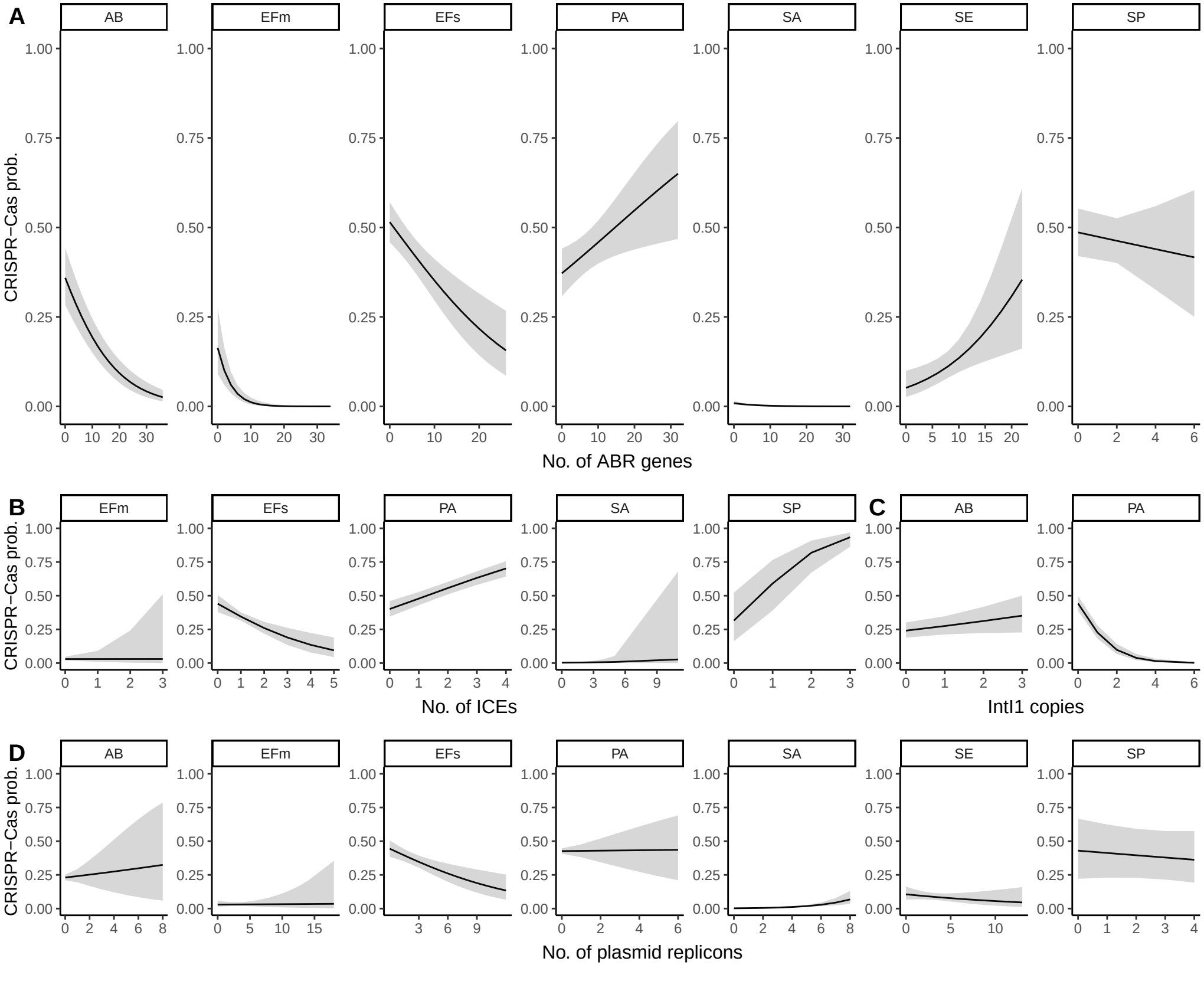
682 47. Pires DP, Costa AR, Pinto G, Meneses L, Azeredo J. 2020 Current challenges
683 and future opportunities of phage therapy. *FEMS Microbiol. Rev.* **44**, 684–700.
684 (doi:10.1093/femsre/fuaa017)

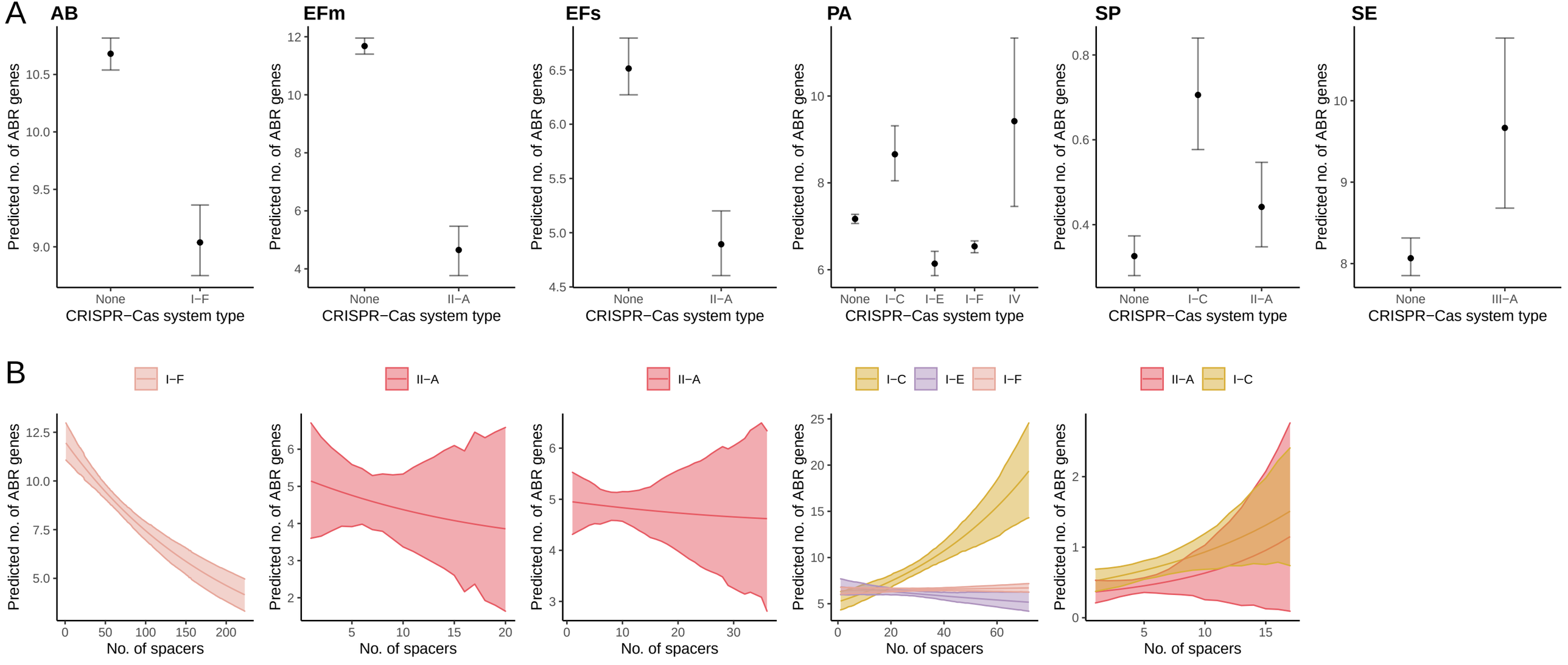
685 48. Bikard D, Barrangou R. 2017 Using CRISPR-Cas systems as antimicrobials.
686 *Curr. Opin. Microbiol.* **37**, 155–160. (doi:10.1016/j.mib.2017.08.005)

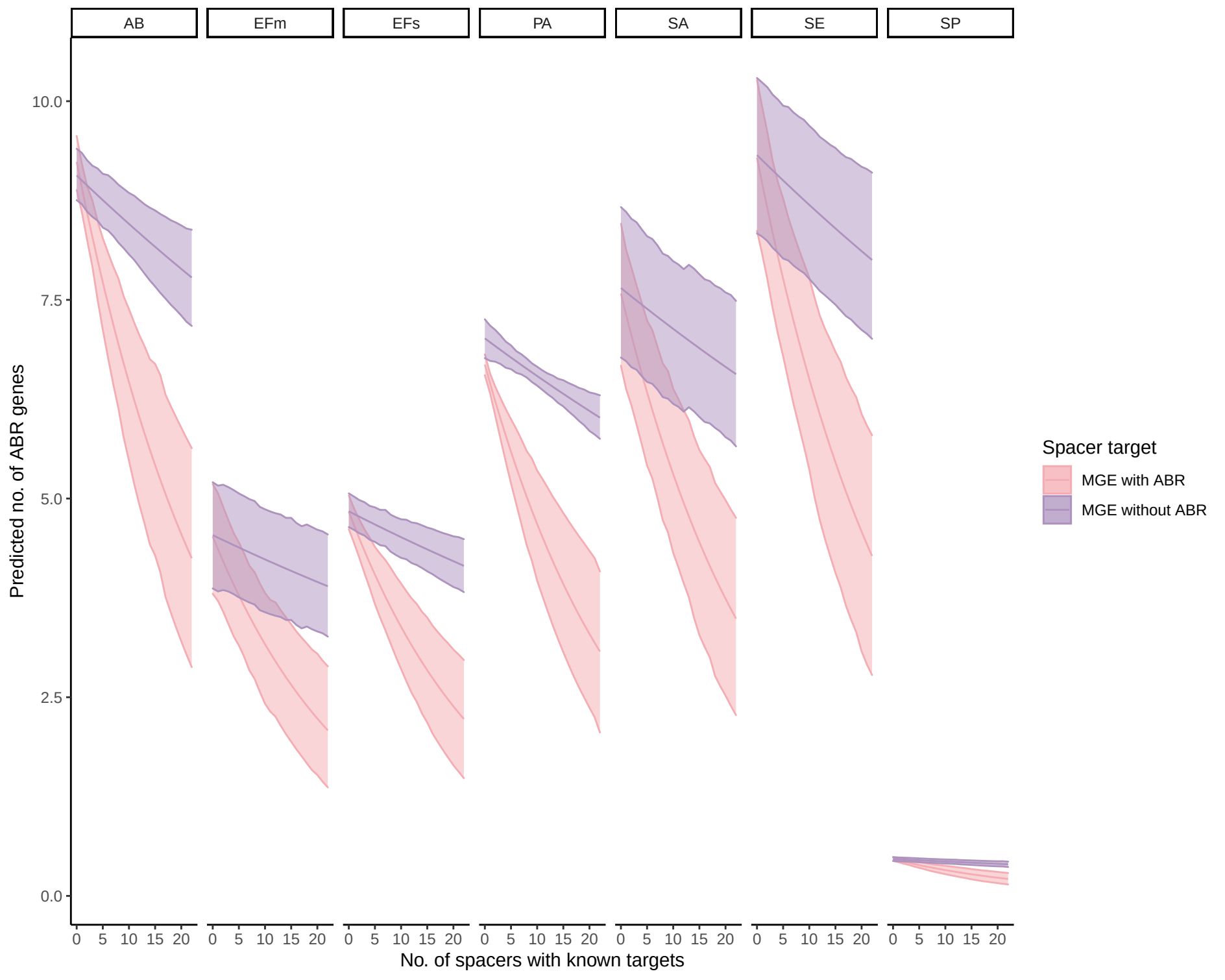
687 49. Pursey E, Sünderhauf D, Gaze W, Westra ER, van Houte S. 2018 CRISPR-Cas
688 antimicrobials□: Challenges and future prospects. *PLoS Pathog.* **14**, e1006990.

689 50. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White
690 J-SS. 2009 Generalized linear mixed models: a practical guide for ecology and
691 evolution. *Trends Ecol. Evol.* **24**, 127–135. (doi:10.1016/j.tree.2008.10.008)

692


A
AB
EFm
EFs
PA
SA
SE
SP
CRISPR−Cas prob.
No. of ABR genes
B
EFm
EFs
PA
SA
SP
CRISPR−Cas prob.
No. of ICEs
C
AB
PA
IntI1 copies
D
AB
EFm
EFs
PA
SA
SE
SP
CRISPR−Cas prob.
No. of plasmid replicons


A
AB
EFm
EFs
PA
SP
SE
Predicted no. of ABR genes
CRISPR−Cas system type
None
I−F
II−A
I−C
I−E
IV
III−A
B
No. of spacers


AB
EFm
EFs
PA
SA
SE
SP
Predicted no. of ABR genes
10.0
7.5
5.0
2.5
0.0
0
5
10
15
20
No. of spacers with known targets
Spacer target
MGE with ABR
MGE without ABR