A scaling law in CRISPR repertoire sizes arises from avoidance of autoimmunity

Hanrong Chen,[1, *] Andreas Mayer,[2, †] and Vijay Balasubramanian[1, 3]

[1]*David Rittenhouse Laboratory, Department of Physics and Astronomy, University of Pennsylvania, USA*
[2]*Lewis-Sigler Institute for Integrative Genomics, Princeton University, USA*
[3]*Theoretische Natuurkunde, Vrije Universiteit Brussel, Belgium*
(Dated: January 4, 2021)

Some bacteria and archaea possess an adaptive immune system that maintains a memory of past viral infections as DNA elements called spacers, stored in the CRISPR loci of their genomes. This memory is used to mount targeted responses against threats. However, cross-reactivity of CRISPR targeting mechanisms suggests that incorporation of foreign spacers can also lead to autoimmunity. We show that balancing antiviral defense against autoimmunity predicts a scaling law relating spacer length and CRISPR repertoire size. By analyzing a database of microbial CRISPR-Cas systems, we find that the predicted scaling law is realized empirically across prokaryotes, and arises through the proportionate use of different CRISPR types by species differing in the size of immune memory. In contrast, strains with nonfunctional CRISPR loci do not show this scaling. We also demonstrate that simple population-level selection mechanisms can generate the scaling, along with observed variations between strains of a given species.

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins form a prokaryotic defense against phage [1]. CRISPR loci are composed of DNA repeats alternating with variable DNA segments called spacers, acquired from phage and other foreign genetic material. In a process called interference, spacer RNA guides sequence-specific binding and cleavage of target DNA by Cas proteins. In this way, spacers acquired during phage attack confer acquired, heritable resistance against subsequent invasions.

CRISPR-Cas systems are remarkably diverse, characterized by functionally divergent Cas proteins and distinct mechanisms for each stage of immune defense [2]. Spacer acquisition is mediated by the conserved Cas1–Cas2 adaptation module, which sets spacer lengths within a narrow range varying by system [3, 4]. CRISPR arrays are also broadly distributed in size, ranging from less than 10 to hundreds of spacers, and the full repertoire of a host may comprise several CRISPR arrays [5]. Maintaining a broad spacer repertoire confers resistance against many phages and possible escape mutants [6]. However, there are constitutive costs associated with Cas protein expression [7], and diminishing returns of broad defense due to finite Cas protein copy numbers [8, 9]. In addition, CRISPR-Cas systems can prevent horizontal transfer of beneficial mobile genetic elements [10, 11].

CRISPR-Cas systems also cause autoimmunity, occurring when a spacer guides interference somewhere on the host genome, leading to cell death and strong mutational pressure in the CRISPR-cas locus and target region [12–15]. The patchy incidence of CRISPR-Cas systems in prokaryotes (roughly 40% of bacteria and 85% of archaea [2]), and the presence of diverse mechanisms for self–nonself discrimination [16], suggest that avoiding autoimmunity is a constraint in the evolution of CRISPR-Cas systems [2, 13–19].

Several mechanisms exist in divergent CRISPR-Cas types for suppressing autoimmunity arising from different forms of potential self-targeting [16]. In type I and II systems, interference requires presence of a protospacer-adjacent motif (PAM), a 2–5-nt-long sequence adjacent to target DNA but absent in CRISPR repeats, preventing interference within the CRISPR array [20, 21]. In type III systems, interference requires transcription of target DNA, which avoids targeting phages integrated into the host chromosome (prophage) [22]. Spacers acquired from the host genome are naturally self-targeting, but there are mechanisms to suppress such acquisition [23, 24]. For example, type I-E systems acquire spacers preferentially at double-stranded DNA breaks, which occur primarily at stalled replication forks of replicating phage DNA, and acquisition is confined by Chi sites which are enriched in bacterial genomes [23].

Here we propose that CRISPR evolution is also shaped by *heterologous autoimmunity*, which occurs if an acquired foreign spacer and a segment of the host genome are sufficiently similar. The likelihood of this effect depends on sequence statistics and the specificity of CRISPR targeting mechanisms. Heterologous autoimmunity is analogous to off-target effects that are an important concern in CRISPR-Cas genome editing [25, 26], but the possible effects on prokaryotic adaptive immunity have not been explored. We combine a probabilistic modeling approach with comparative analyses of CRISPR repertoires across prokaryotes to show that: (a) heterologous autoimmunity is a significant threat caused by CRISPR-Cas immune defense, (b) avoidance of autoimmunity leads to a scaling law in CRISPR repertoires, and (c) the scaling law can be achieved by population-level selection. Our work suggests that avoidance of heterologous autoimmunity is a key factor shaping CRISPR repertoires and the evolution of CRISPR-Cas systems.

---

* These two authors contributed equally to this work.; Present address: Computational & Systems Biology, Genome Institute of Singapore, Singapore
† These two authors contributed equally to this work.

I. RESULTS

A. Cross-reactivity leads to autoimmunity

We approach heterologous self-targeting as a sequence-matching problem [27–29], and derive estimates for the probability of a spacer being sufficiently similar to at least one site in the host genome. For a spacer of length l_s and PAM of length l_p (where it exists), an exact match at a given position requires $l \equiv l_s + l_p$ complementary nucleotides. In a host genome of length L, where $L \gg l$, there are $L - l + 1 \approx L$ starting positions for a match. At leading order, and ignoring nucleotide usage biases, we may treat matches as occurring independently with probability 4^{-l}. Thus, the probability of an exact match anywhere on the genome is (see Methods)

$$p_0 \equiv L 4^{-l}, \text{ where } l = l_s + l_p. \tag{1}$$

Considering order-of-magnitude parameter estimates for the E. coli type I-E system of $L = 5 \times 10^6$ nt, $l_s = 32$ nt, and $l_p = 3$ nt gives a negligible probability $p_0 \sim 10^{-15}$.

However, CRISPR interference tolerates several mismatches between spacer RNA and target DNA depending on position and identity [25, 26, 30, 31]. In general, mismatches in the PAM are not allowed, and mismatches in the PAM-distal region are tolerated to a greater extent than mismatches in the PAM-proximal region [21]. Up to ~ 5 mismatches are allowed in type II systems [25, 26], while in type I-E systems, errors are mostly tolerated at specific positions with a 6-nt periodicity [30, 31].

Partial spacer-target matching may also trigger primed spacer acquisition, which is the rapid acquisition of new spacers from regions surrounding target DNA [30, 32, 33]. In type I-E and I-F systems, primed acquisition tolerates many (up to 10) mismatches in the PAM and target region [30, 33]. Thus, a foreign spacer that does not cause direct interference may still trigger primed acquisition of self-spacers [33] and hence cause autoimmunity.

Given that the specificity of CRISPR interference and primed acquisition have been characterized for only a few systems, we consider two general classes of mismatch tolerance that include the above scenarios: (a) mismatches at k_{fix} fixed positions, and (b) mismatches at k_{var} variable positions anywhere else in the target region. These increase the per-spacer self-targeting probability by a combinatorial factor $\alpha(k_{\text{fix}}, k_{\text{var}}, l)$ (see Methods), so that

$$p_{\text{self}} = \alpha(k_{\text{fix}}, k_{\text{var}}, l) \, p_0 = \alpha(k_{\text{fix}}, k_{\text{var}}, l) \, L \, 4^{-l}. \tag{2}$$

A greater number of allowed mismatches greatly increases the likelihood of heterologous self-targeting (Fig. 1b). To gain intuition we can rewrite Eq. 2 as

$$p_{\text{self}} \equiv L \, 4^{-l_{\text{eff}}}, \text{ where} \tag{3}$$
$$l_{\text{eff}}(l, k_{\text{fix}}, k_{\text{var}}) \equiv l - \log_4 \alpha \approx l - k_{\text{fix}} - k_{\text{var}} \log_4 3(l - k_{\text{fix}}),$$

where l_{eff} is the effective spacer length after discounting for allowed mismatches (see Methods). This shows

that mismatches exponentially increase the probability of self-targeting, and variable-position mismatches particularly so. Considering the E. coli system as before, the matching probability increases to $p_{\text{self}} \sim 10^{-4}$ with $k_{\text{fix}} = 5$ nt and $k_{\text{var}} = 5$ nt (see Fig. 1b). Other CRISPR-Cas systems may similarly lie in parameter regimes with appreciable p_{self}, especially when including indirect self-targeting through primed acquisition [34]. Furthermore, the probability of self-targeting is likely higher than implied by our calculations as it can be increased by correlations in sequence statistics between host and phage genomes [28, 29]. Given our estimates, we thus hypothesize that heterologous autoimmunity may occur generally and be a significant cost of CRISPR-Cas immunity.

B. Spacer length scales with repertoire size

To test this hypothesis, we exploited the large natural variability in CRISPR systems across different microbial species. As the self-targeting probability depends exponentially on spacer length (Eqs. 2–3), we expect small differences in length to lead to large variations in the risk of autoimmunity. If CRISPR repertoire sizes are selected to balance broader immunity against the risk of autoimmunity, then qualitatively we expect that species with shorter spacers should have smaller repertoires, while species with longer spacers should have larger ones (Fig. 1c).

To make this prediction more quantitative, suppose prokaryotes tolerate a maximum probability P of self-targeting, and that CRISPR-Cas systems are selected to maximize protection against pathogens subject to this constraint. Repertoires with N spacers incur a self-targeting probability of $\sim N p_{\text{self}}$, and thus Eq. 2 implies

$$\ln N = l \ln 4 - \ln \alpha(k_{\text{fix}}, k_{\text{var}}, l) - \ln(L/P). \tag{4}$$

Linearizing the dependence of the combinatorial factor α around typical spacer lengths l_0 (see Methods) predicts a scaling relationship between spacer length and the logarithm of repertoire size

$$\ln N \sim l \left(\ln 4 - \frac{k_{\text{var}}}{l_0 - k_{\text{fix}}} \right) \sim 1.2 \, l, \tag{5}$$

where we arrived at the latter estimate by taking $k_{\text{var}} \sim k_{\text{fix}} \sim 5$ and $l_0 \sim 35$.

We analyzed a database of CRISPR-Cas systems identified in publicly available bacterial and archaeal genomes [5, 35] (see Methods). To sample widely from CRISPR-Cas systems while eliminating oversampling of certain species, we first selected strains carrying both CRISPR and cas loci, and then picked one strain at random from each species for further analysis (see Methods). We observed a multimodal distribution of spacer lengths acquired by these representative strains (Fig. 2a), consistent with different CRISPR-Cas types having narrow spacer length distributions (Fig. 3a). The distribution of
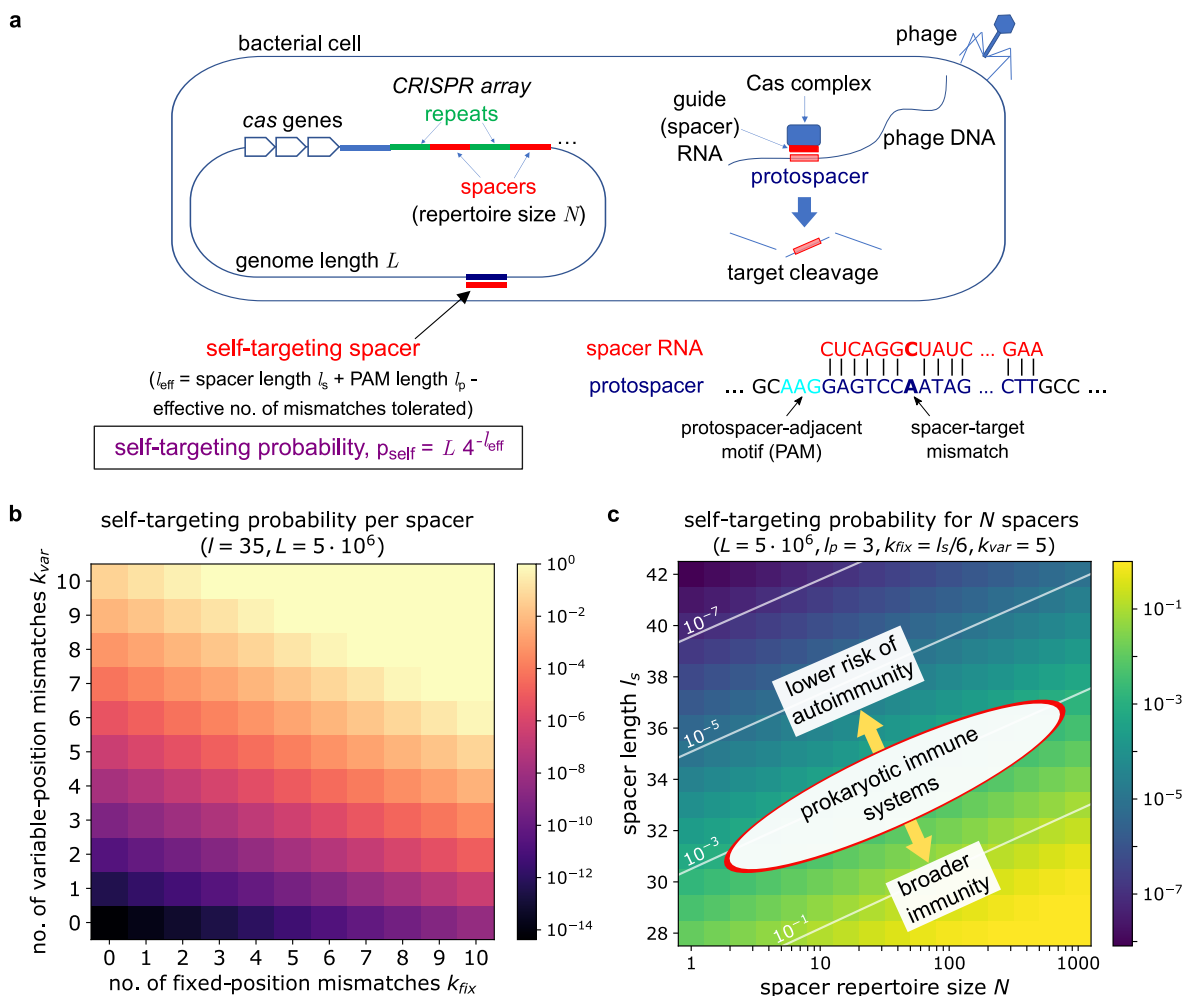
Figure 1. **CRISPR-Cas immune defense incurs a risk of heterologous autoimmunity.** **a**, Sketch of the main components of CRISPR-Cas immune defense. **b**, Per-spacer probability of heterologous self-targeting, p_{self}, as a function of the number of tolerated mismatches at fixed and variable positions along the spacer, k_{fix} and k_{var}, respectively. **c**, We hypothesize that the evolution of CRISPR-Cas systems is constrained by the risk of heterologous autoimmunity. As the self-targeting probability depends strongly on spacer length, this predicts a scaling of repertoire size with spacer length.

spacer repertoire sizes, defined as the sum of CRISPR array sizes in each genome, was broad, ranging from 1 to 812 spacers (Fig. 2b).

A linear regression between spacer length and log-repertoire size gave a slope of 1.1 ± 0.1 (Fig. 2c), in line with the predicted scaling (Eq. 5). A range of cross-reactivity parameters is broadly consistent with this scaling (Fig. S1), with a best-fit value of $k_{\text{var}} = 3.41 \pm 0.02$ obtained assuming $k_{\text{fix}} = l_s/6$ (consistent with a 6-nt periodicity in tolerated fixed-position mismatches as in type I-E systems) (see Fig. S1). The empirical law holds over two orders of magnitude in CRISPR repertoire size, but over this range the spacer length changes only modestly. These changes however lead to significant differences in the self-targeting probability, which is exponential in spacer length (Eq. 3).

Some prokaryotes may tolerate self-targeting spacers because they have defective cas genes [12] or contain

anti-CRISPRs [36]. To further test the link between autoimmune risk and spacer length, we investigated the incidence of missing cas genes across CRISPR-Cas systems. We expected that a higher autoimmune risk in species with shorter spacers would lead to a higher rate of cas gene loss. Thus, we analyzed how the fraction of strains with missing cas loci depends on spacer length, which shows the expected relationship (Fig. 2d). Once immunopathology from self-targeting is avoided by the loss of cas interference genes, the relation between spacer length and repertoire size should no longer be selected for. Indeed, we found no clear relation in strains with missing cas loci (Fig. 2e). Taken together, these observations strengthen the interpretation of the scaling law as arising from the modulation of autoimmunity risk by spacer length.
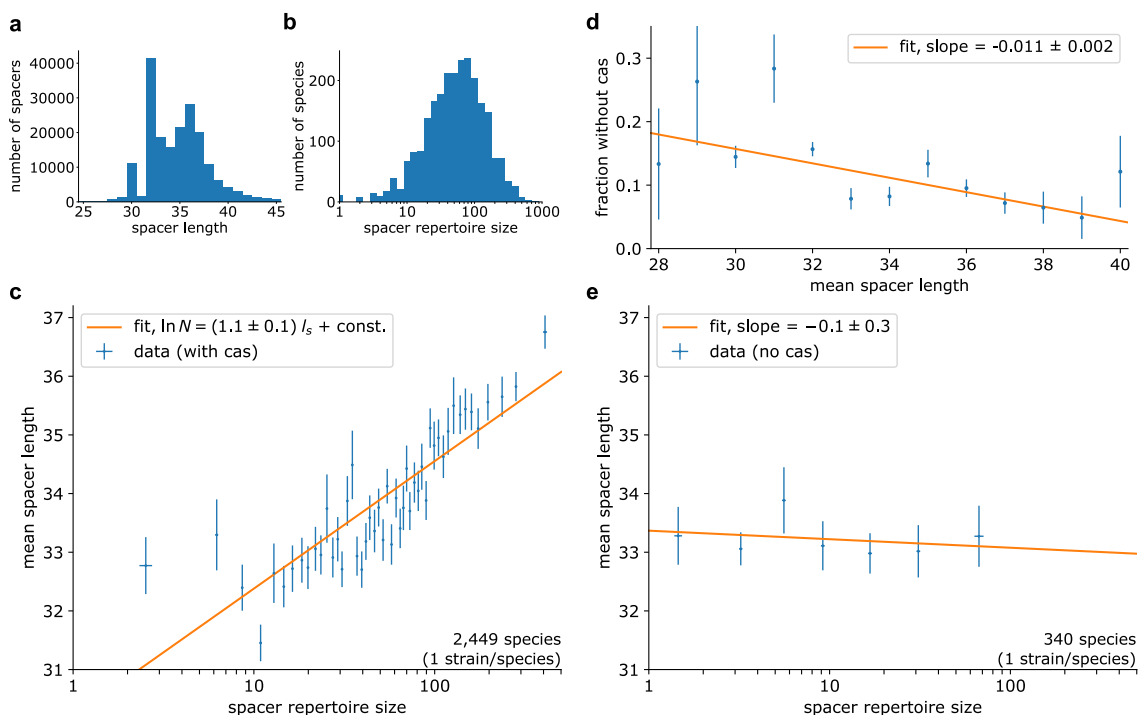
Figure 2. **A scaling law relates CRISPR repertoire size and spacer length. a, b** Distributions of spacer lengths (a) and repertoire sizes (b) across prokaryotes. For each of 2,449 species with CRISPR and cas loci we randomly picked a single strain (see Methods), and calculated its repertoire size as the sum of all CRISPR array sizes present in the genome. The length distribution of all spacers found in these filtered strains are shown in (a). Bins in (b) were formed by dividing each decade into 10 equal bins on a log scale. **c**, Scaling of repertoire size with spacer length. A linear fit of the mean spacer length against log repertoire size was performed on all 2,449 species, and is shown alongside the data, which is binned by repertoire size (50 strains/bin). The fitted slope is consistent with theory predictions (Eq. 5). **d**, Fraction of species with missing cas genes decreases with spacer length. **e**, Spacer length and repertoire size do not show a clear relation in strains with nonfunctional CRISPR loci. A linear fit was performed on 340 species with CRISPR arrays but no cas loci, and is shown alongside the data, which is binned by repertoire size (50 strains/bin). Error bars in panels c–e denote the standard error of the mean in each bin, which in (d) are calculated assuming a binomial probability distribution for the absence of cas at each spacer length.

C. Variable CRISPR-Cas type use underlies scaling

CRISPR-Cas systems are classified into different types and subtypes based on their evolutionary relationships and the use of different cas genes [2]. We wondered whether the aggregate scaling relationship between spacer length and repertoire size (Fig. 2) reflected differences at the level of CRISPR-Cas type usage. We thus grouped the species by subtype, when there is a single CRISPR-Cas system in the genome, or in a separate group when multiple subtypes are present.

For species carrying a single cas type, we aggregated all spacers found across species of each type to quantify the statistics of acquired spacer lengths. We observed differences in the spacer length distributions between types (Fig. 3a): (a) Type II-A and II-C systems have narrow distributions tightly clustered around 30 nt; (b) Type I-E and I-F systems also have narrow distributions, clustered around 32 nt, while other type I systems have spacers that are longer and more broadly distributed; (c) Type III systems have even longer and more broadly distributed spacers, with median lengths in the 36–39 nt range.

A broader distribution of acquired spacer lengths leads to a higher risk of autoimmunity than a narrow distribution with the same mean, since the self-targeting probability increases exponentially for shorter spacers. To account for an increase in autoimmune risk for broader distributions, we focused on the lower quartile of spacer lengths for each cas type as a proxy for autoimmune risk. Also, to account for the requirement of PAM recognition in type I and II (but not type III) systems, we added a PAM length of 3 nt to types I and II to obtain the overall length l. Strikingly, we observed that the predicted relationship between l and repertoire size also broadly holds between CRISPR-Cas types (Fig. 3b): Type II systems have the shortest spacers and the smallest repertoires, and among type I subtypes those with shorter spacers generally have smaller repertoires. Type III systems have smaller repertoires than type I systems despite somewhat longer spacers, but this is explained by the absence of PAMs and the broader spacer length distributions for
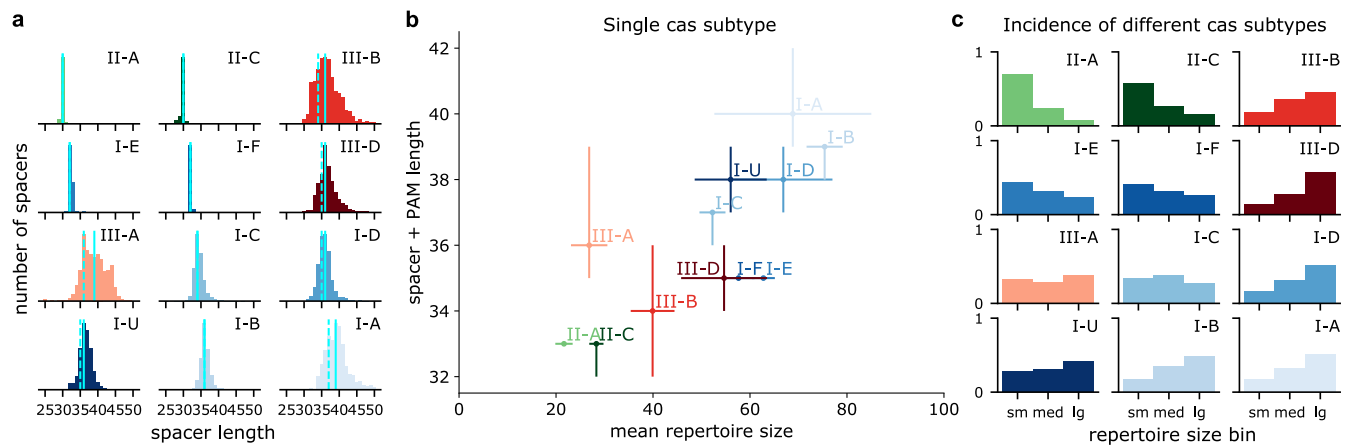
Figure 3. **Preferential CRISPR-Cas type use underlies the scaling law relating spacer length and repertoire size.** **a**, Length distributions of all spacers found in single-type strains aggregated by CRISPR-Cas type (for types with > 10 species in CRISPRCasdb [35]; see Methods). Also indicated are the median (solid vertical) and lower quartile (dotted vertical line) for each distribution. The subtypes are presented in order of lower quartile. **b**, A trend is observed between spacer + PAM length and repertoire size for different CRISPR-Cas types. For spacer lengths, the central dot is the lower quartile and the whisker runs between the lowest decile and the median. Repertoire sizes are indicated as the mean ± standard error. To indicate the requirement of PAM recognition, a length of 3 nt was added to all type I and II (but not type III) subtypes. **c**, Variable usage of cas subtypes among multiple-type strains. A total of 826 strains with multiple CRISPR-Cas systems, randomly picked from different species, were analyzed. They were divided into 3 groups of 275, 275 and 276 strains having small, medium, and large repertoire size, respectively. Each subfigure was normalized to 1, so that the bars indicate the relative incidence of a subtype in each repertoire size bin. The order of subtypes is the same as panel a.

the type III systems, both of which increase autoimmune risk.

We next tested whether this relation also carries over to species carrying multiple CRISPR-Cas systems, in the form of a differential use of cas types as a function of repertoire size. We divided species with multiple cas types into three equally sized groups by repertoire size, and determined the relative incidence of CRISPR subtypes within each group (Fig. 3c). We found that the use of types II, I-E, and I-F decreases with repertoire size in line with expectations, and an opposite pattern for two of the type III systems and the type I systems with the longest spacers. The relation between total repertoire size and spacer length in species with multiple cas subtypes was further reinforced by a direct analysis of the incidence of spacers of different lengths as a function of repertoire size, with a greater proportion of longer spacers present in larger repertoires (Fig. S2).

Taken together, we find that species carrying either single or multiple CRISPR-Cas systems differentially use CRISPR-Cas types having different spacer length distributions to form repertoires of different sizes. This differential use gives rise to the aggregate scaling observed in Fig. 2c, and is consistent with the hypothesis of minimizing the risk of heterologous autoimmunity.

D. Dynamical origin of the scaling law

Dynamical mechanisms can give rise to the scaling law that our theory predicts, and which is found in the empirical data. While spacer dynamics involves complex epidemiological feedbacks [8, 37–44], here we consider a simple effective model in which spacer acquisition and loss are described as a birth-death process, such that spacers are acquired at a rate b and lost at a per-spacer rate d (Fig. 4a, left panel). This gives rise to a Poisson distribution of repertoire sizes at steady state, with mean b/d (see SI). Our statistical theory requires that the mean of the distribution should shift with spacer length. There are two mechanisms by which selection could lead to such a dependence. First, the negative fitness effect of acquiring self-targeting spacers [13] purges lineages that undergo deleterious acquisition events. Indeed, CRISPR arrays are selected for the absence of self-targeting spacers [12]. Effectively, this reduces the net acquisition rate among surviving lineages. Second, over longer evolutionary timescales, different CRISPR-Cas systems may be selected to acquire spacers at different rates depending on their respective risks of autoimmunity. These differences in rates could arise from the maintenance of multiple copies of cas genes, or through regulation of cas expression [45]. Indeed, spacer repertoire size increases with the number of cas loci (Fig. S3), suggesting that larger gene copy numbers of cas1 and cas2, necessary for spacer acquisition, result in greater acquisition rates. Interestingly, strains having exactly one copy of both cas1
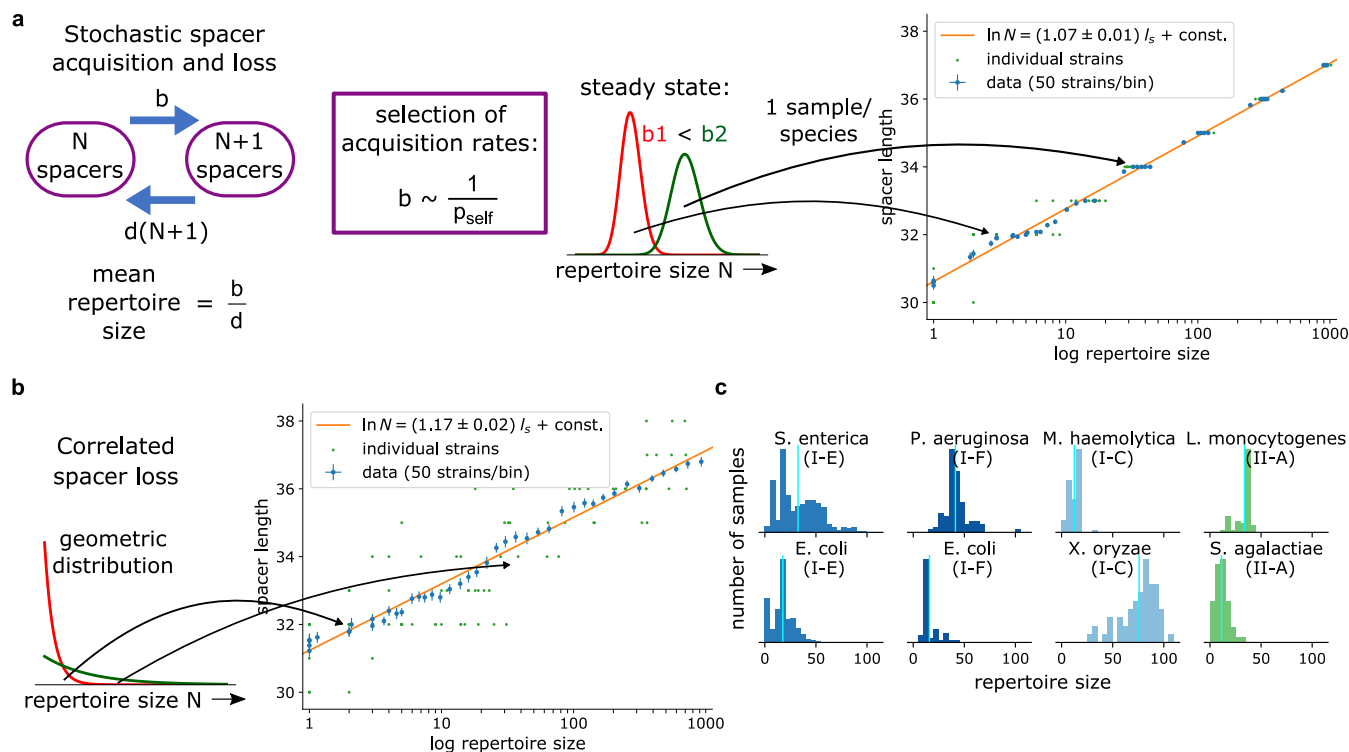
Figure 4. A spacer acquisition-loss model with population-level selection of acquisition rates produces scaling as well as substantial variability between strains of the same species. **a**, In our model, strains acquire spacers at a rate b and lose them with a per-spacer rate d, giving rise to a Poisson distribution at steady state with mean b/d (left panel). b is selected to minimize the risk of heterologous autoimmunity, such that species differing in p_{self} have different mean repertoire sizes (middle panel). We generate a synthetic dataset of strains by sampling from steady-state distributions with different spacer lengths and hence p_{self} (see Methods). The synthetic data displays scaling of the mean and variability on the single-strain level (right panel). The green points show 100 individually sampled strains, the blue points means after binning by repertoire size (50 species/bin, 2,450 species total), and the orange line is a fit to all 2,450 sampled points. **b**, Correlated spacer loss broadens the predicted distribution of repertoire sizes. We consider a model in which all spacers are lost simultaneously during a deletion event, which leads to a geometric steady-state distribution (see SI). Despite this additional variability, a synthetic sample generated as in panel a shows scaling of the means. **c**, The distributions of repertoire sizes of sequenced strains belonging to the same species are broad. 4 pairs of species with > 50 sequenced strains and with the indicated CRISPR-Cas type are displayed. Vertical lines denote the mean for each species.

and cas2 still obey a scaling relationship (Fig. S4), suggesting that regulation of these genes also contributes to minimizing autoimmune risk.

Let us suppose that one or both of these selection mechanisms lead to an effective spacer acquisition rate inversely proportional to the risk of self-targeting, $b \propto 1/p_{\text{self}}$. To replicate the empirical analysis, we created a synthetic dataset of the same size by sampling strains at random from steady-state distributions at different spacer lengths, which have different p_{self} (Fig. 4a, middle panel) (see Methods). Plotting spacer length against mean repertoire size in the same way as we did for the empirical data, we recover a scaling law as predicted by our theory (Fig. 4a, right panel).

In addition to providing a dynamical explanation for scaling of the means, this birth-death model produces substantial variability around the mean relationship (Fig. 4a, right panel, green dots). In fact, the Poisson variance of Fig. 4a, originating from a constant per-spacer loss rate, is likely an underestimate. Spacer loss occurs through double-stranded DNA breaks followed by homologous recombination at a different CRISPR repeat, a process which may delete chunks of an array in a single deletion event (see e.g. [46, 47]). Including such correlated spacer loss greatly increases the variance. For example, in a simple analytically solvable limiting case where entire arrays are lost at once, the distribution of repertoire sizes becomes geometric (see SI) and thus very broad (Fig. 4b, left panel). Given this substantial variability, we wondered whether a comparative analysis of many species could still recover a scaling in this model. We thus sampled strains from geometric steady-state distributions whose means obey a scaling law as in panel a, observing a much larger variability in individual strains (Fig. 4b, right panel, green dots). However, the scaling of the mean is recovered with a fit to the dataset and

when strains are binned by repertoire size (Fig. 4b, right panel).

Prompted by the broad variability predicted by correlated spacer loss, we analyzed the repertoire size distributions of species with many sequenced strains (see Methods). We indeed observed a broad distribution even among strains of the same species (Fig. 4c). We compared the repertoire size distributions of four pairs of highly-sampled species with the same CRISPR-Cas type, and found that the within-species variability comprises a substantial part of the overall variance. Additional variability between species, leading to different mean repertoire sizes for species with the same CRISPR-Cas type, might originate from different microbes inhabiting environments that differ in viral diversity and thus pressure to acquire broad immune defense. We tested the robustness of the comparative analysis to this additional source of variability, by sampling strains from steady-state distributions where we additionally sample the prefactor in $b \propto 1/p_{\text{self}}$ from a wide distribution (see Methods). This further increases the variability of individually sampled strains, but the means still show scaling (Fig. S5).

II. DISCUSSION

An adaptive immune system is dangerous equipment to have in an organism. There is always the risk that the immune receptors, intended as defenses against foreign invaders, will instead target the self. In CRISPR-Cas systems, biophysical mechanisms avoiding various forms of autoimmunity such as targeting of the CRISPR locus and self-spacer acquisition are known [16, 20, 21, 23], but here we propose that heterologous autoimmunity, where spacers acquired from foreign DNA seed self-targeting, is a significant threat to microbes carrying CRISPR-Cas. This threat is analogous to off-target effects in genome-editing applications [25, 26], and has been observed in an experimental CRISPR-Cas system [33], but its wider implications for the evolution of CRISPR-Cas systems are unexplored. We showed that avoidance of this form of autoimmunity while maximizing antiviral defense predicts a scaling law relating spacer length and CRISPR repertoire size. The scaling depends on the number and nature of sequence mismatches permitted during CRISPR interference and primed acquisition.

To test our prediction we used a comparative approach analyzing the natural variation in CRISPR-Cas systems across microbial species, and demonstrated that: (a) the predicted scaling law is realized, (b) the observed scaling constrains parameters for cross-reactive CRISPR targeting to lie in a range consistent with experimental studies, (c) the scaling arises in part from differential usage of different CRISPR-cas subtypes having different spacer length distributions, and (d) the scaling, and hence a balanced tradeoff between successful defense and autoimmunity, can be achieved by population-level selection mechanisms. In addition, we demonstrated a negative control:

CRISPR arrays in species that no longer have functional Cas proteins, and thus are not at risk of autoimmunity, do not show the predicted scaling relation. We propose two further tests of the link between spacer length and autoimmune risk: (1) If cross-reactivity leads to self-targeting, in addition to a depletion of self-targeting spacers in CRISPR arrays [12, 36], we predict a depletion of spacers several mismatches away from self-targets, and (2) Our theory predicts that CRISPR-Cas subtypes with longer spacers should acquire spacers more readily.

A similar tradeoff between sensitivity to pathogens and autoimmune risk shapes the evolution of vertebrate adaptive immune systems [27, 48]. In the light of our results it would be interesting to determine whether this tradeoff also leads to a relation between the size of the immune repertoire and specificity in vertebrates. Such a relation will likely be harder to ascertain for vertebrates as patterns of cross-reactivity between lymphocyte receptors and antigens are more complex. Interestingly, however, T cell receptor hypervariable regions in human are several nucleotides longer on average than those found in mice [49], which accompanies a substantial increase in repertoire size in human. If longer hypervariable regions translate to a greater specificity on average, one might view the increased human receptor length as an adaptation to a larger repertoire. The key to our current work was the ability to compare microbial immune strategies across a large panel of phylogenetically distant species. Further insight into how this tradeoff shapes vertebrate immune systems might thus be gained by building on recent efforts to survey adaptive immune diversity in a broader range of vertebrates [50, 51].

Many theoretical studies of adaptive immunity in both prokaryotes [8, 37–44] and vertebrates [52–55] consider detailed dynamical models of evolving immune repertoires. For prokaryotes, such dynamical models can be regarded as describing the role of CRISPR-Cas as a short-term memory for defense against a co-evolving phage [56]. Studying adaptive immunity in this way requires detailed knowledge of the parameters controlling the dynamics, many of which are not well-characterized experimentally. In this paper, we took an alternative approach of focusing on the statistical logic of adaptive immunity, where we regard the bacterial immune system as a functional mechanism for maintaining a long-term memory [56] of a diverse phage landscape [57], via probabilistic matching of genomic sequences. Previous work taking this perspective offered an explanation for why prokaryotic spacer repertoires lie in the range of a few dozen to a few hundred spacers [56]. As in our discussion of possible mechanisms for generating the observed scaling law, evolution should select dynamics that achieve the statistical organization that we predict, because this is what is useful for achieving a broad defense against phage while avoiding autoimmunity. A probability theory perspective of this kind has been applied to the logic of the adaptive immune repertoire of vertebrates [58–60], but to our knowledge we are presenting a novel approach to

the study of CRISPR-based autoimmunity.

III.   MATERIALS AND METHODS

a.   Derivation of self-targeting probability.   We estimate the probability of an alignment between a spacer + PAM sequence of length l and a host genome of length L. We assume that both sequences are random and uncorrelated, with nucleotide usage frequencies of $1/4$. In a length-L genome, where $L \gg l$, there are $L - l + 1 \approx L$ starting positions for an alignment. The matching probability at each position, p_m, depends on the number and nature of mismatches tolerated. In regimes where p_m is small, the matching probabilities at the different positions may be treated independently. Thus, the probability of having at least one alignment within the length-L genome is

$$p_{\text{self}} = 1 - (1 - p_m)^{L-l+1}$$
$$\approx Lp_m, \text{ since } p_m \ll 1, l \ll L. \qquad (6)$$

If no mismatches are tolerated, $p_m = 4^{-l}$ as in Eq. 1. At each site where a mismatch is allowed, four alternative nucleotide choices are possible. This gives a certain number α of unique complementary sequences matching to a given spacer, which we compute as a function of the number and nature of mismatches. If up to k_{fix} mismatches are tolerated at fixed positions in the alignment, $\alpha = 4^{k_{\text{fix}}}$. If instead up to k_{var} mismatches are tolerated anywhere in the complementary region, naively $\alpha \sim \binom{l}{k_{\text{var}}} 4^{k_{\text{var}}}$, where the binomial coefficient is the number of combinations of sites where mismatches are allowed. This is however an upper bound as matching sequences are overcounted, and the precise expression is

$$\alpha = \sum_{i=0}^{k_{\text{var}}} \binom{l}{i} 3^i, \qquad (7)$$

where each term in the sum is the number of unique complementary sequences having exactly i mismatches. The largest term dominates, giving $\alpha \approx \binom{l}{k_{\text{var}}} 3^{k_{\text{var}}}$. Thus, combining k_{fix} mismatches at fixed positions and up to k_{var} mismatches at any of the remaining $l - k_{\text{fix}}$ positions gives

$$\alpha(k_{\text{fix}}, k_{\text{var}}, l) \approx 4^{k_{\text{fix}}} \binom{l - k_{\text{fix}}}{k_{\text{var}}} 3^{k_{\text{var}}}. \qquad (8)$$

We can introduce an effective spacer length, l_{eff}, by $p_m \equiv 4^{-l_{\text{eff}}}$. To leading order the binomial expression in Eq. 8 is approximated by $(l - k_{\text{fix}})^{k_{\text{var}}}$. This gives $l_{\text{eff}} \approx l - k_{\text{fix}} - k_{\text{var}} \log_4 3(l - k_{\text{fix}})$ as in Eq. 3.

The probability that a repertoire of N spacers avoids self-targeting, $1 - P_{\text{self}}$, is one minus the probability that at least one spacer self-targets. This gives

$$P_{\text{self}} = 1 - (1 - p_{\text{self}})^N$$
$$\approx Np_{\text{self}}, \text{ since } Lp_m \ll 1. \qquad (9)$$

If CRISPR repertoires are selected to maximize repertoire size subject to the constraint $P_{\text{self}} \leq P$, we obtain Eq. 4. Taylor expanding $\ln N$ around $l = l_0$ gives Eq. 5 to lowest order in l.

b.   Comparative analyses.   For our comparative analyses we use CRISPRCasdb [5], which is a database of CRISPR and cas loci identified using CRISPRCasFinder [61] in public bacterial and archaeal whole-genome assemblies [35]. CRISPR arrays are assigned evidence levels 1–4, 4 being the highest confidence [61]. We restricted our analysis to level 4 CRISPR arrays only. Strains containing both annotated CRISPR and cas loci were used for the analyses in Figs. 2a–c, 3, and 4c. Strains containing annotated CRISPR but no cas loci were used for the analyses in Figs. 2d–e. In order to eliminate oversampling of certain species, we picked one strain at random from each species for further analysis (2,449 species with annotated CRISPR and cas loci, and 340 species with annotated CRISPR but no cas loci). To produce Fig. 3, the randomly chosen strains were grouped by annotated cas subtype, or into a separate group if they contain multiple subtypes. The 12 subtypes shown in Figs. 3a and c have >10 species represented in CRISPRCasdb. To produce Fig. 4c, 4 pairs of species, each with > 50 sequenced strains of the same CRISPR-Cas type, were chosen for analysis.

c.   Synthetic data generation and analysis.   A synthetic dataset producing a scaling law was generated in the following way: (1) A spacer of length l_s was drawn from the length distribution of Fig. 2a, and (2) a repertoire size distribution with mean A/p_{self} was created, from which one strain was sampled and added to the dataset. Parameter values of $L = 5 \cdot 10^6$, $l_p = 3$, $k_{\text{fix}} = l_s/6$, $k_{\text{var}} = 3$, and $A = 10^{-5.5}$ were used. The steady-state distributions are Poisson in Fig. 4a, and geometric with the same mean in Fig. 4b. In Fig. S5, A was sampled from a log-normal distribution with the same mean, and standard deviation chosen such that the coefficient of variation is 1.2.

[1] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath, Science **315**, 1709 (2007).

[2] K. S. Makarova, Y. I. Wolf, J. Iranzo, S. A. Shmakov, O. S. Alkhnbashi, S. J. J. Brouns, E. Charpentier, D. Cheng, D. H. Haft, P. Horvath, S. Moineau, F. J. M. Mojica, D. Scott, S. A. Shah, V. Siksnys, M. P. Terns, C. Venclovas, M. F. White, A. F. Yakunin, W. Yan, F. Zhang, R. A. Garrett, R. Backofen, J. van der Oost, R. Barrangou, and E. V. Koonin, Nat. Rev. Microbiol. **18**, 67 (2020).

[3] J. Wang, J. Li, H. Zhao, G. Sheng, M. Wang, M. Yin, and Y. Wang, Cell **163**, 840 (2015).

[4] J. K. Nunez, L. B. Harrington, P. J. Kranzusch, A. N. Engelman, and J. A. Doudna, Nature **527**, 535 (2015).

[5] C. Pourcel, M. Touchon, N. Villeriot, J. P. Vernadet, D. Couvin, C. Toffano-Nioche, and G. Vergnaud, Nucleic Acids Res. **48**, D535 (2020).

[6] S. van Houte, A. K. E. Ekroth, J. M. Broniewski, H. Chabas, B. Ashby, J. Bondy-Denomy, S. Gandon, M. Boots, S. Paterson, A. Buckling, and E. R. Westra, Nature **532**, 385 (2016).

[7] P. F. Vale, G. Lafforgue, F. Gatchitch, R. Gardan, S. Moineau, and S. Gandon, Proceedings of the Royal Society B: Biological Sciences **282**, 20151270 (2015).

[8] A. Martynov, K. Severinov, and I. Ispolatov, PLoS Comp. Bio. **13**, 1 (2017).

[9] S. Bradde, A. Nourmohammad, S. Goyal, and V. Balasubramanian, Proc. Natl. Acad. Sci. U.S.A. **117**, 5144 (2020).

[10] L. A. Marraffini and E. J. Sontheimer, Science **322**, 1843 (2008).

[11] W. Jiang, I. Maniv, F. Arain, Y. Wang, B. R. Levin, and L. A. Marraffini, PLoS Genetics **9**, e1003844 (2013).

[12] A. Stern, L. Keren, O. Wurtzel, G. Amitai, and R. Sorek, Trends Genet. **26**, 335 (2010).

[13] R. B. Vercoe, J. T. Chang, R. L. Dy, C. Taylor, T. Gristwood, J. S. Clulow, C. Richter, R. Przybilski, A. R. Pitman, and P. C. Fineran, PLoS Genet. **9**, e1003454 (2013).

[14] D. Paez-Espino, W. Morovic, C. L. Sun, B. C. Thomas, K. Ueda, B. Stahl, R. Barrangou, and J. F. Banfield, Nat. Commun. **4**, 1430 (2013).

[15] Y. Wei, R. M. Terns, and M. P. Terns, Genes Dev. **29**, 356 (2015).

[16] L. A. Marraffini, Nature **526**, 55 (2015).

[17] R. Edgar and U. Qimron, J. Bacteriol. **192**, 6291 (2010).

[18] G. W. Goldberg, E. A. McMillan, A. Varble, J. W. Modell, P. Samai, W. Jiang, and L. A. Marraffini, Nat. Commun. **9**, 61 (2018).

[19] C. Rollie, A. Chevallereau, B. N. J. Watson, T. Y. Chyou, O. Fradet, I. McLeod, P. C. Fineran, C. M. Brown, S. Gandon, and E. R. Westra, Nature **578**, 149 (2020).

[20] H. Deveau, R. Barrangou, J. E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, and S. Moineau, J. Bacteriol. **190**, 1390 (2008).

[21] E. Semenova, M. M. Jore, K. A. Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. Brouns, and K. Severinov, Proc. Natl. Acad. Sci. U.S.A. **108**, 10098 (2011).

[22] G. W. Goldberg, W. Jiang, D. Bikard, and L. A. Marraffini, Nature **514**, 633 (2014).

[23] A. Levy, M. G. Goren, I. Yosef, O. Auster, M. Manor, G. Amitai, R. Edgar, U. Qimron, and R. Sorek, Nature **520**, 505 (2015).

[24] J. W. Modell, W. Jiang, and L. A. Marraffini, Nature **544**, 101 (2017).

[25] Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung, and J. D. Sander, Nat. Biotechnol. **31**, 822 (2013).

[26] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao, and F. Zhang, Nat. Biotechnol. **31**, 827 (2013).

[27] J. K. Percus, O. E. Percus, and A. S. Perelson, Proc. Natl. Acad. Sci. U.S.A. **90**, 1691 (1993).

[28] S. Karlin and S. F. Altschul, Proc. Natl. Acad. Sci. U.S.A. **87**, 2264 (1990).

[29] A. Dembo, S. Karlin, and O. Zeitouni, The Annals of Probability **22**, 2022 (1994).

[30] P. C. Fineran, M. J. Gerritzen, M. Suárez-Diez, T. Künne, J. Boekhorst, S. A. van Hijum, R. H. Staals, and S. J. Brouns, Proc. Natl. Acad. Sci. U.S.A. **111**, E1629 (2014).

[31] C. Jung, J. A. Hawkins, S. K. Jones, Y. Xiao, J. R. Rybarski, K. E. Dillard, J. Hussmann, F. A. Saifuddin, C. A. Savran, A. D. Ellington, A. Ke, W. H. Press, and I. J. Finkelstein, Cell **170**, 35 (2017).

[32] K. A. Datsenko, K. Pougach, A. Tikhonov, B. L. Wanner, K. Severinov, and E. Semenova, Nat. Commun. **3**, 945 (2012).

[33] R. H. Staals, S. A. Jackson, A. Biswas, S. J. Brouns, C. M. Brown, and P. C. Fineran, Nat. Commun. **7**, 12853 (2016).

[34] T. J. Nicholson, S. A. Jackson, B. I. Croft, R. H. Staals, P. C. Fineran, and C. M. Brown, RNA Biology **16**, 566 (2019).

[35] Grissa, Ibtissem and Drevet, Christine and Couvin, David, CRISPRCasdb (2020), [Online; accessed 26-July-2020].

[36] K. E. Watters, C. Fellmann, H. B. Bai, S. M. Ren, and J. A. Doudna, Science **239**, 236 (2018).

[37] J. He and M. W. Deem, Physical Review Letters **105**, 128102 (2010).

[38] B. R. Levin, PLoS Genet. **6**, e1001171 (2010).

[39] L. M. Childs, N. L. Held, M. J. Young, R. J. Whitaker, and J. S. Weitz, Evolution **66**, 2015 (2012).

[40] B. R. Levin, S. Moineau, M. Bushman, and R. Barrangou, PLoS Genet. **9**, e1003312 (2013).

[41] J. Iranzo, A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, J. Bacteriol. **195**, 3834 (2013).

[42] A. D. Weinberger, C. L. Sun, M. M. Pluciński, V. J. Denef, B. C. Thomas, P. Horvath, R. Barrangou, M. S. Gilmore, W. M. Getz, and J. F. Banfield, PLoS Comp. Biol. **8**, e1002475 (2012).

[43] S. Bradde, M. Vucelja, T. Teşileanu, and V. Balasubramanian, PLoS Comp. Bio. **13**, 1 (2017), arXiv:1510.06082.

[44] P. Han and M. W. Deem, Journal of the Royal Society Interface **14** (2017).

[45] A. G. Patterson, M. S. Yevstigneyeva, and P. C. Fineran, Current Opinion in Microbiology **37**, 1 (2017).

[46] G. W. Tyson and J. F. Banfield, Environ. Microbiol. **10**, 200 (2008).

[47] P. Horvath, D. A. Romero, A. C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, and R. Barrangou, J. Bacteriol. **190**, 1401 (2008).

[48] C. J. E. Metcalf, A. T. Tate, and A. L. Graham, Nature Ecology & Evolution , 1 (2017).

[49] Z. Sethna, Y. Elhanati, C. S. Dudgeon, C. G. Callan, A. J. Levine, T. Mora, and A. M. Walczak, Proc. Natl. Acad. Sci. U.S.A. **114**, 201700241 (2017).

[50] R. Castro, S. Navelsaker, A. Krasnov, L. Du Pasquier, and P. Boudinot, Developmental and comparative immunology **75**, 28 (2017).

[51] R. Covacu, H. Philip, M. Jaronen, D. C. Douek, S. Efroni, F. J. Quintana, R. Covacu, H. Philip, M. Jaronen, J. Almeida, J. E. Kenison, and S. Darko, Cell Reports **14**, 2733 (2016).

[52] R. Antia, V. V. Ganusov, and R. Ahmed, Nature Reviews Immunology **5**, 101 (2005).

[53] G. Lythe, R. E. Callard, R. L. Hoare, and C. Molina-París, Journal of Theoretical Biology **389**, 214 (2016).

[54] J. Desponds, A. Mayer, T. Mora, and A. M. Walczak, arXiv preprint arXiv:1703.00226 (2017).

[55] M. Gaimann, M. Nguyen, J. Desponds, and A. Mayer, eLife **9**, e61639 (2020).

[56] S. Bradde, A. Nourmohammad, S. Goyal, and V. Balasubramanian, Proc. Natl. Acad. Sci. U.S.A. **117**, 5144 (2020).

[57] R. A. Edwards and F. Rohwer, Nature Reviews Microbiology **3**, 504 (2005).

[58] A. Mayer, V. Balasubramanian, T. Mora, and A. M. Walczak, Proc. Natl. Acad. Sci. U.S.A. **112**, 5950 (2015).

[59] A. Mayer, T. Mora, O. Rivoire, and A. M. Walczak, Proc. Natl. Acad. Sci. U.S.A. **113**, 8630 (2016).

[60] A. Mayer, V. Balasubramanian, A. M. Walczak, and T. Mora, Proc. Natl. Acad. Sci. U.S.A. **116**, 8815 (2019).

[61] D. Couvin, A. Bernheim, C. Toffano-Nioche, M. Touchon, J. Michalik, B. Neron, E. P. C. Rocha, G. Vergnaud, D. Gautheret, and C. Pourcel, Nucleic Acids Res. **46**, W246 (2018).

SUPPLEMENTARY INFORMATION

SI text on population dynamics

Consider a host population acquiring spacers of length l. Let the number of individuals in the population that have repertoire size n ($n \geq 0$) be X_n. Consider spacer acquisition to occur at a rate b:

$$X_n \xrightarrow{b} X_{n+1}. \tag{10}$$

Spacer acquisition is balanced by spacer loss leading to a well-defined steady-state distribution of repertoire size. Spacer loss occurs through double-stranded DNA breaks followed by homologous recombination at a subsequent repeat, which deletes chunks of the CRISPR array (see e.g. [46, 47]). The precise rate and mechanism by which this occurs is not well-understood. Here, we consider 3 solvable scenarios of this process:

$$1: X_n \xrightarrow{d} X_{n-1} \tag{11}$$

$$2: X_n \xrightarrow{dn} X_{n-1} \tag{12}$$

$$3: X_n \xrightarrow{d} X_0. \tag{13}$$

The first scenario represents spacer loss at the end(s) of the CRISPR array, hence independent of n. The second represents a constant per-spacer loss rate. For the third scenario, all spacers are lost in a deletion event, which is a solvable limit of several spacers being deleted at a time.

Scenario 1: $X_n \xrightarrow{d} X_{n-1}$. The probabilities P_n ($n \geq 0$) obey the following master equation:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = -(b+d)P_n + bP_{n-1} + dP_{n+1}, \ n \geq 1 \tag{14}$$

$$\frac{\mathrm{d}P_0}{\mathrm{d}t} = -bP_0 + dP_1. \tag{15}$$

The steady state fulfills the detailed balance condition,

$$dP_n = bP_{n-1}. \tag{16}$$

We can solve the recursion equation (Eq. 16) for the steady-state distribution,

$$P_n = (b/d)^n (1 - b/d), \tag{17}$$

which is geometric with parameter $1 - b/d$. Its mean is $b/(b-d)$, implying that a well-defined steady state is only possible if $d > b$.

Scenario 2: $X_n \xrightarrow{dn} X_{n-1}$. The master equation is:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = -(b+dn)P_n + bP_{n-1} + d(n+1)P_{n+1}, \ n \geq 1 \tag{18}$$

$$\frac{\mathrm{d}P_0}{\mathrm{d}t} = -bP_0 + dP_1. \tag{19}$$

At steady state again detailed balance holds

$$dnP_n = bP_{n-1}. \tag{20}$$

Eq. 20 implies that the steady-state distribution is Poisson with mean b/d:

$$P_n = \frac{1}{n!}(b/d)^n e^{-b/d}. \tag{21}$$

Scenario 3: $X_n \xrightarrow{d} X_0$. Here, the master equation is:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = -(b+d)P_n + bP_{n-1}, \ n \geq 1 \tag{22}$$

$$\frac{\mathrm{d}P_0}{\mathrm{d}t} = -bP_0 + d(1 - P_0). \tag{23}$$

Here there is no detailed balance, but probability flux is conserved,

$$(d+b)P_n = bP_{n-1}. \tag{24}$$

Eq. 24 implies that the steady-state distribution is geometric with parameter $d/(b+d)$:

$$P_n = \left[\frac{b}{b+d}\right]^n \frac{d}{b+d}. \tag{25}$$

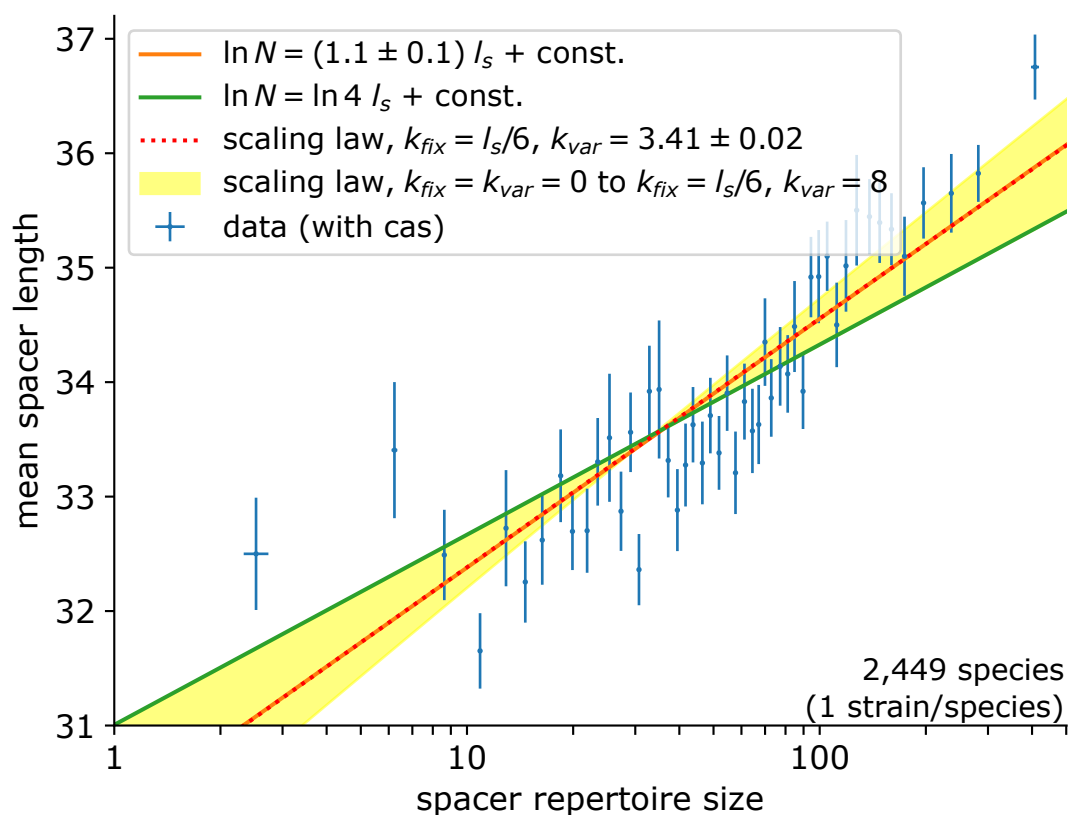The mean of this distribution is b/d.

SI figures

Figure S1. Cross-reactivity parameters obtained by a fit to the empirical data lie in a plausible range. The blue points are data from 2,449 species binned in increasing windows of repertoire size (50 species/bin), and the orange line is the linear fit to all species as in Fig. 2c. The green line is the naive $\ln 4$ scaling (Eq. 5). The fitted slope is consistent with a broad range of cross-reactivity parameters (yellow region). A best-fit to Eq. 4 was performed, in which l_p was fixed at 3, and k_{fix} was set to $l_s/6$, consistently with a 6-nt periodicity in mismatch tolerance in type I-E systems [30, 31] and approximately 5 allowed mismatches in type II systems in which most spacer lengths are ~ 30 nt [25, 26]. We found best-fit values of $k_{\mathrm{var}} = 3.41 \pm 0.02$ and $\log_{10}(L/P) = 11.47 \pm 0.02$, where the errors are 90% confidence intervals. The estimate of k_{var} is consistent with primed acquisition tolerating many mismatches, up to 10 in some systems [30, 33], and the estimate of L/P implies a maximum risk of self-targeting P in the range of 10^{-4} to 10^{-5}. We expect these cross-reactivity parameters to show significant variation around these means in individual species and systems (see Fig. 4).
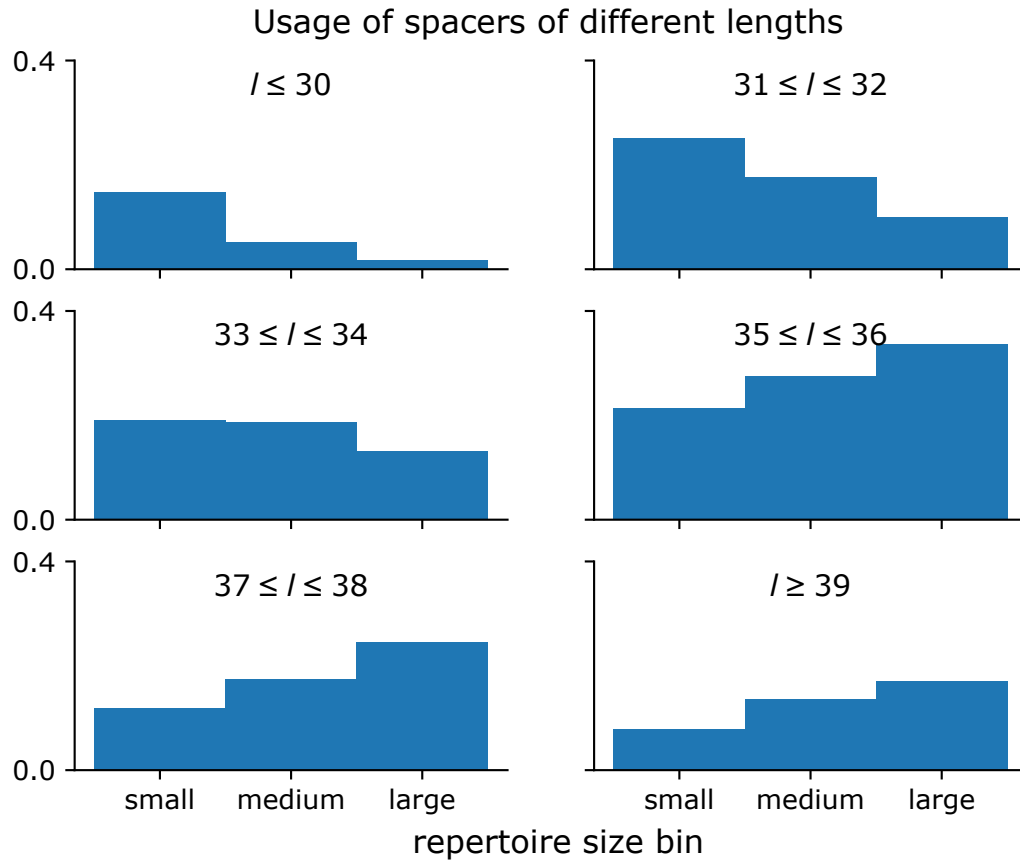
Figure S2. Variable usage of spacer lengths among multiple-type strains. 826 species with multiple CRISPR-Cas systems were divided into 3 groups of small, medium and large repertoire sizes containing 275, 275 and 276 species, respectively. Each repertoire size bin was normalized to 1, so that the bars indicate the fraction of spacers in each repertoire size bin with that length. The usage of spacers of length ≤ 32 nt decreases with repertoire size, while usage of spacers of length ≥ 35 nt increases with repertoire size among these strains.
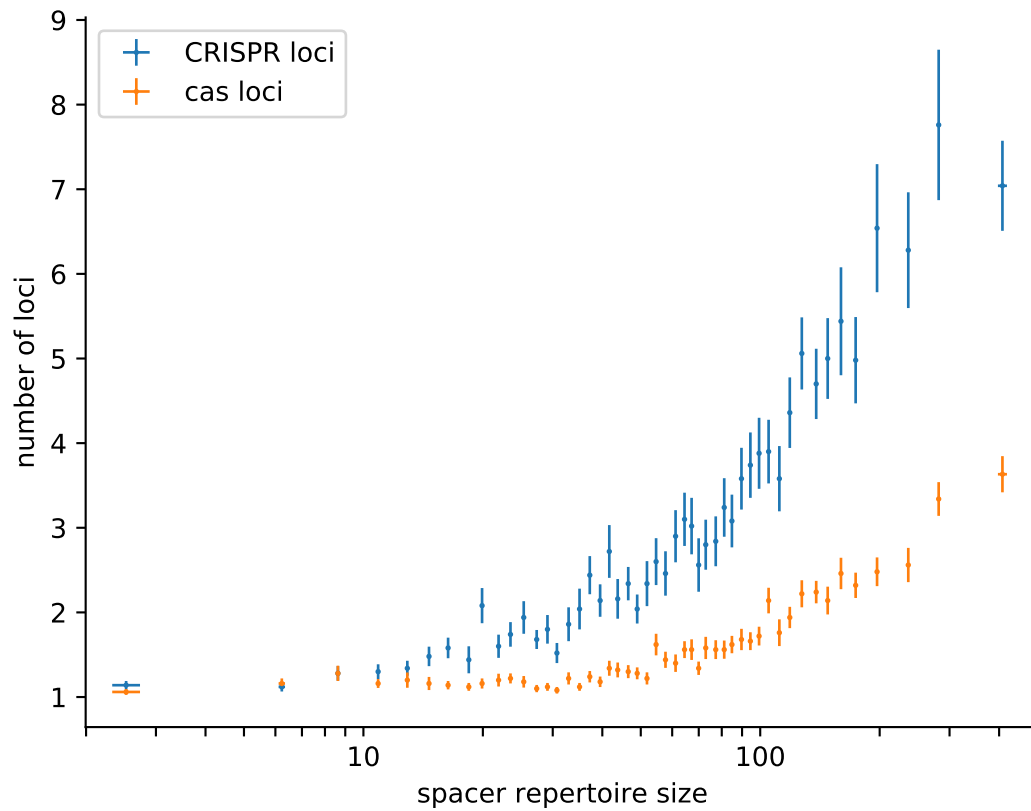
Figure S3. Spacer repertoire size is correlated with the number of CRISPR and cas loci. Data from 2,449 representative strains belonging to different species are binned by repertoire size (50 strains/bin). Error bars denote the standard error of the mean in each bin.
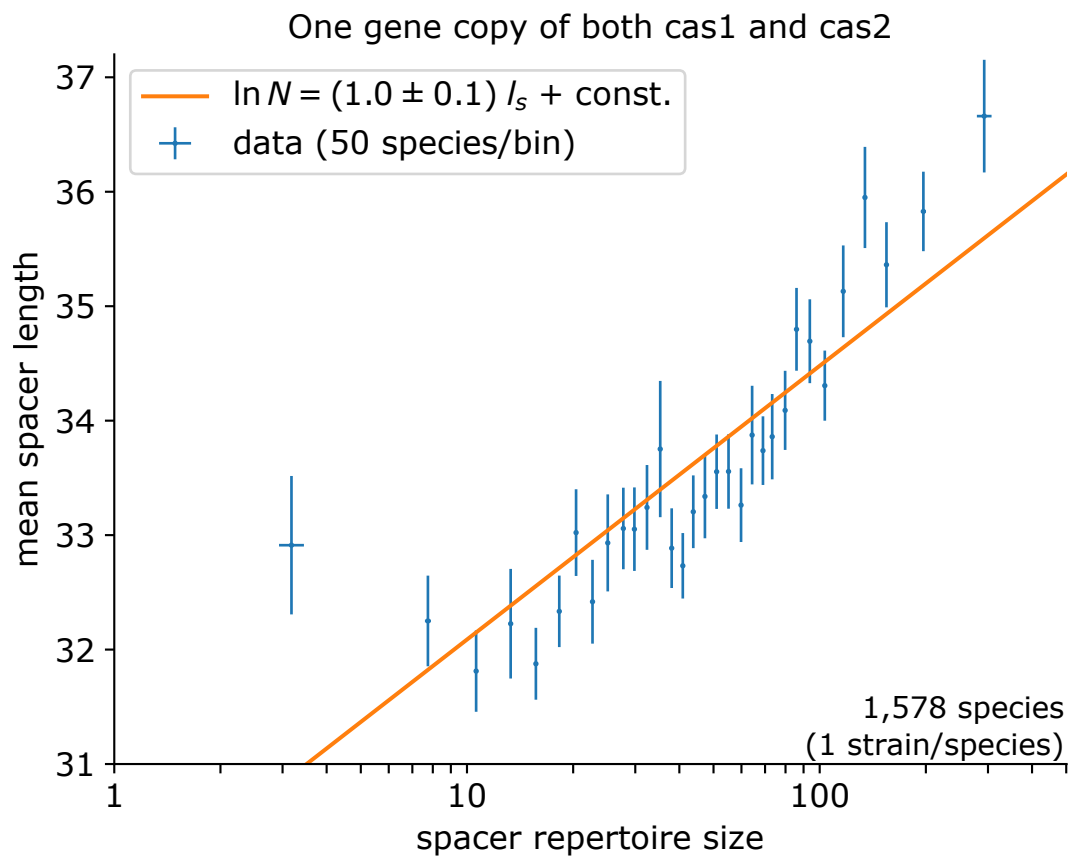
Figure S4. Repertoire size versus mean spacer length for strains restricted to one annotated gene copy of cas1 and cas2. 1,578 out of the 2,449 sampled species contain one gene copy of cas1 and cas2. The orange line is a linear fit to these species, shown alongside the data, which are binned by repertoire size (50 species/bin). Error bars denote the standard error of the mean in each bin.
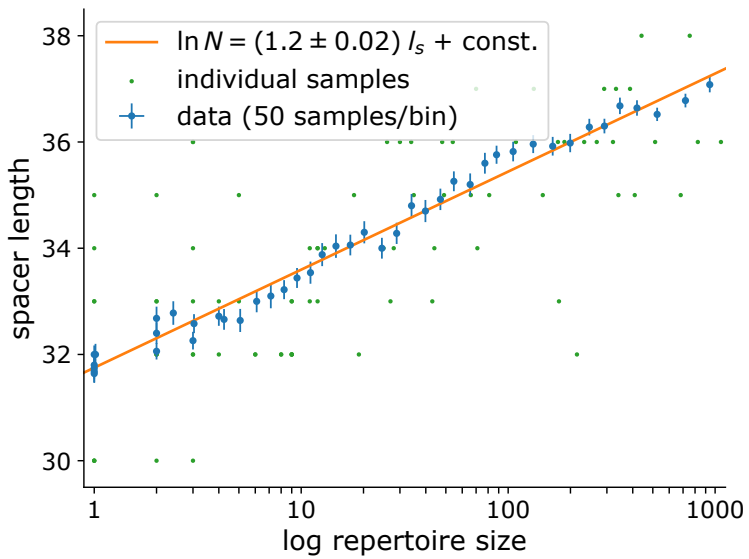
Figure S5. Species and system-specific stochasticity increases the variability of the sampled data, but a scaling law is recovered by binning by repertoire size. A sampling procedure on synthetically generated data is replicated as in Fig. 4. Individuals were drawn from steady-state distributions with mean proportional to $1/p_{\text{self}}$, but each time the prefactor A was also drawn from a wide (log-normal) distribution with the same mean as in Fig. 4a–b, and standard deviation chosen such that the coefficient of variation is 1.2. A large variability in the data results, but binning recovers a clear relation between mean repertoire size and spacer length. The green points show 100 individually sampled strains, the blue points means after binning by repertoire size (50 species/bin, 2,450 species total), and the orange line is a fit.