How adaptive immunity constrains the composition and fate of large bacterial populations - supporting information

Madeleine Bonsma-Fisher[1], Dominique Soutière[1], and Sidhartha Goyal[1,2]

[1]Department of Physics, University of Toronto, 60 St George St, Toronto, ON M5S 1A7
[2]Institute of Biomaterials & Biomedical Engineering, University of Toronto, 164 College Street, Toronto, ON M5S 3G9

1 Data analysis

We used data from [1] which is publicly available in the NCBI Sequence Read Archive under the accession SRA062737. It includes four data files (SRR630110, SRR630111, SRR630412, and SRR630413) which we used for our analysis. We extracted the data corresponding to the MOI2 deep sequencing experiment and separated it into time points by checking each read for matches to the primers identified in the supplementary information of [1]. Any reads with a mismatch between the annotation of the forward and reverse primers were discarded. Any remaining unsorted reads were excluded from the following analysis.

1.1 Identifying and sorting spacers

We extracted and catalogued spacers from the published raw read data of [1]. Since only the expanding CRISPR end was sequenced, each read represents the longest possible sequence from wild type to leader end and so further assembly was not required (SI Figure 1 and SI Figure 2).

Because of this very specialized data structure, detecting CRISPR spacers and inferring their order was conceptually straightforward. A spacer was defined as any sequence flanked by two repeats. Since each read was bordered by wild type sequence and leader end sequence, all repeat sequences were complete and not truncated. SI Figure 2 shows a typical read in more detail. Note that in this orientation, the spacer numbered "1" is found at the end of the read. To collect spacers, we (1) detected repeat sequences, reversing the read if the repeats were reversed, (2) inferred spacers as sequences between repeats, and (3) categorized spacers by comparing to previously detected spacers.

Repeat sequence variation was present due to sequencing errors or naturally occurring SNPs. We used a regular expression to match variations on the number of Ts in a 5-T region of the repeat - the forward repeat was matched with "GTTT*GTACTCTCAAGATTTAAGTAACTGTACAAC" and the reverse repeat was matched with "GTTGTACAGTTACTTAAATCTTGAGAGTACAAA*C". These expressions match an identical string with three or more Ts or As in the region of the asterisk. This is a reasonable allowance to make since the 454 sequencing platform used to sequence this data is known to have high insertion and deletion rates in homopolymer regions [2].

To detect the most possible spacers, we developed methods to deal with repeat sequence variation beyond simple insertions and deletions in the homopolymer region. We inferred the presence of an undetected repeat by measuring the length of sequence before the first detected repeat, after the last detected repeat, and between two repeats. If any of these lengths exceeded its threshold (determined based on the known primer lengths and average spacer length, respectively), a more careful search

www.pnas.org/cgi/doi/10.1073/pnas.1802887115

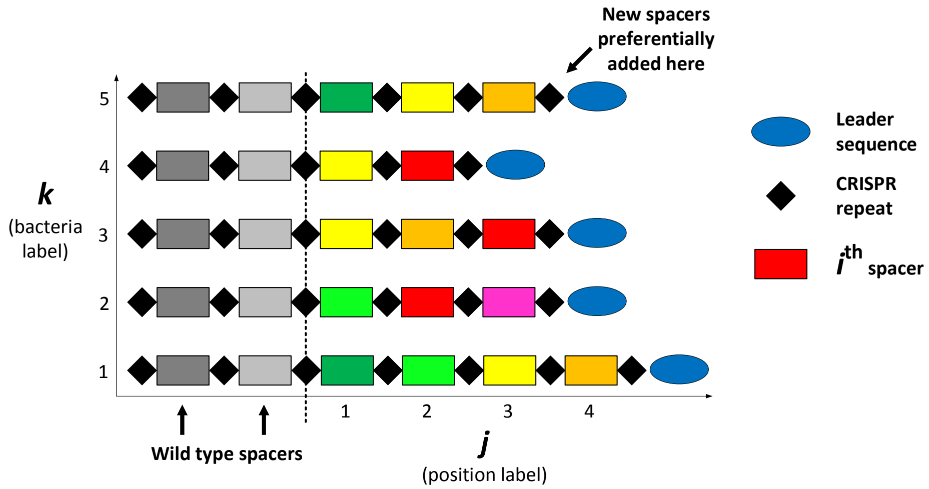

Figure 1: Schematic of the portion of the S. *thermophilus* CRISPR locus sequenced in [1]. We identified spacers with a type i, a locus position j, and bacteria number k. Coloured rectangles to the right of the dashed line represent spacers sequenced as the locus expands. Wild type spacers are shown in greyscale.

*ATCAGACACGGAAAGTAAGGATTGACAAGG*ACAGTTATTGATTTTATAATCACTATGTGGGTATAA
AACGTCAAGATTTTATTTGAG**GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACCCAACACTC**
AAACGTTGCAAACGCAAGCTTGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACATCACTTAC
GAGGTTGACGGTTTTGTAGATGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC*TGTTTGACA*
GCAAATCAAGCGTGTCTGATCTGAGCGGGCTGGCAAGGCGCATAGN

CRISPR repeat, *forward primer*, *reverse primer*, **spacer**

Figure 2: Example read covering the expanding CRISPR locus. The forward primer which overlaps with the leader sequence is shown in blue italics. The reverse primer which overlaps the first wild type spacer is shown in green italics. CRISPR repeats are shown in bold red and spacers in bold black.

for repeats was performed using the *pairwise2* module in Biopython which performs a local pairwise alignment between the ideal repeat sequence and the read in question.

1.1.1 Pairwise alignment settings

If the alignment with the true repeat (36 nucleotides long) was less than 31 nucleotides long, the alignment was discarded. The scoring system was as follows: match score of 1, mismatch score of -1, gap open score of -0.8 for the target sequence, gap open score of -0.7 for the repeat, and gap extend penalty of -1 for each sequence. The gap open scores were chosen to be different for the repeat and read so that the algorithm could identify how many gaps were opened and in which each sequence, in order to properly identify the start and end of each spacer.

If no good match was found in a region between two repeats, the remaining "long" spacer was discarded and a placeholder was inserted to preserve position information. Using this method, the number of detected repeats increased from 550931 to 622067, a 12.9% increase.

Repeats detected in this second search sometimes contained gaps with respect to the read or vice versa. In these cases, conventional labelling of nucleotide position prevented accurate detection of the start and end of adjacent spacers. We detected how many gaps were present and whether they occurred in the repeat or the read and then adjusted the indices of adjacent spacers accordingly. The scoring scheme was carefully chosen so that the number and placement of gaps could be inferred from the score.

1.1.2 Spacer type assignment

We compared newly detected spacers to a growing list of previously detected spacers to assign it a type. If it matched an existing spacer exactly, it was assigned that type. Otherwise, a global pairwise alignment was performed between the new spacer and all existing spacers. If a match was found for which the score subtracted from the spacer length was within a chosen cutoff, the new spacer was assigned that type. This definition of cutoff is equivalent to the number of allowed SNPs between spacers under the scoring scheme used. If no match was found in either case, the new spacer was assigned a new type.

To choose an appropriate tolerance for spacer alikeness, we tested this spacer sorting algorithm on a small sample of data (190 reads) as the cutoff was increased from 0 to 9. SI Figure 3 shows the number of unique spacer types detected as the cutoff is changed. It can be seen that there is a clear plateau between cutoff values of 1 and 8, which indicates that the system is insensitive to the cutoff if it falls in this range. We chose a cutoff of 2 for the analysis.

In this way, we created a master dataset for each time point that contained each detected spacer, a number indicating the source read, the spacer position in the read, and the assigned spacer type. The definition of spacer type was consistent across time points, or in other words the same comparison list was carried through all time points.

1.2 Analysis

We extracted CRISPR spacers from the raw reads at each time point by finding sequences flanked by an S. *thermophilus* CRISPR repeat (SI Figure 1). Newly detected spacers were added to an existing group if they were within an edit distance of 2 of another spacer in that group. Data was organized into an array s_{ijk} (equation 1).

$$s_{ijk}(t) = \begin{cases} 1 & \text{if spacer type } i \text{ is at position } j \text{ in bacterium } k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We tracked individual spacer types, or "clones", $n_B^i(t)$, by summing over all bacteria and all locus positions: $n_B^i(t) = \sum_{j,k} s_{ijk}(t)$.

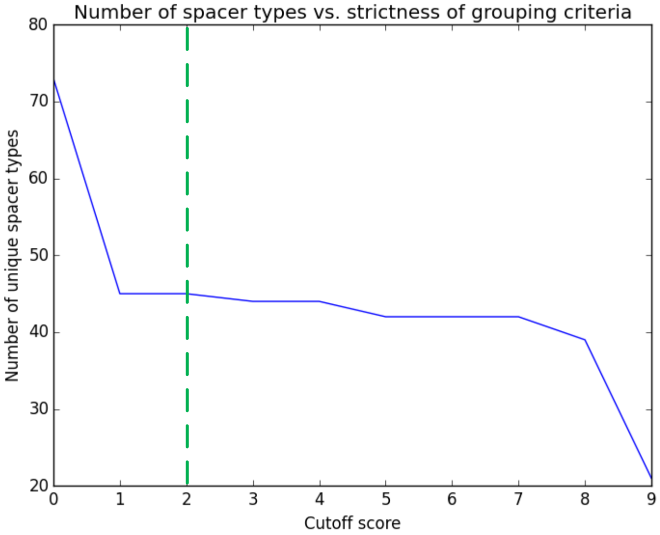

Figure 3: Number of unique spacer types vs. cutoff for 190 reads from time point 11. The green vertical dashed line indicates the selected cutoff.

Most bacteria acquired only a single spacer; over half of bacteria from days 4-14 which had acquired 1 or more spacers only acquired a single spacer (SI Figure 4).

2 Model description

Table 1: Model parameters

Parameter	Description	Value
$\frac{1}{gC_0}$	Bacterial doubling time	41.7 min
C_0	Inflow nutrient concentration in units of bacterial cell density	
α	Phage adsorption rate	2×10^{-10} min^{-1}
B	Phage burst size	170
F	Chemostat flow rate	
p_V	Probability of phage success for bacteria without spacers	
e	Spacer effectiveness	
r	Rate of spacer loss	
η	Probability of spacer acquisition	

Parameter values are as above unless otherwise indicated. Representative values estimated for Streptococcus thermophilus bacteria in lab conditions.

We model bacteria and phages interacting in a chemostat. The populations we track are nutrient concentration C, phages n_V, and bacteria n_b which can either have no spacer (n_b^0) or a spacer of type i (n_b^i). Nutrients flow in at concentration C_0 with rate F, and all species flow out with rate F. The total number of bacteria with a spacer is n_b^s and the total number of bacteria is n_B. The phage in

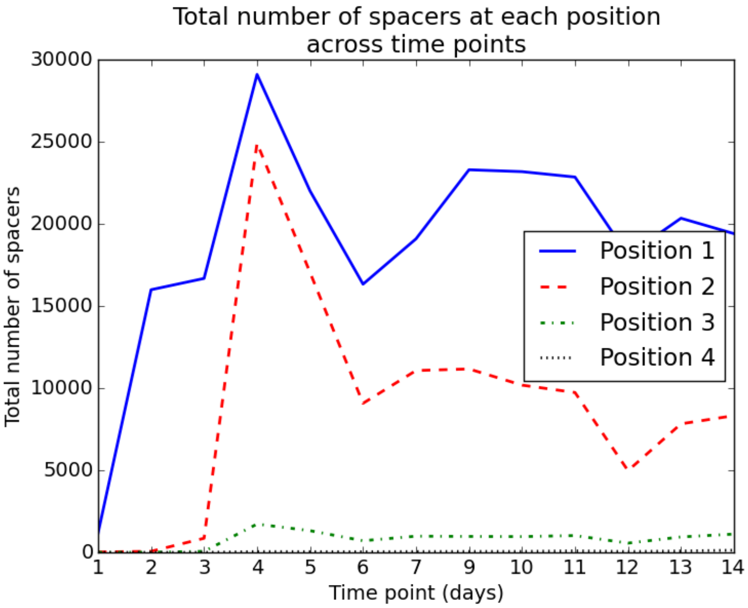

Figure 4: Total number of spacers at each time point in the experiment. Position 1 represents the oldest spacer (closest to the wild type spacers). Over half of all bacteria that acquired spacers, even at the end of the experiment, only acquired a single spacer.

the solution are all clonal and have m distinct protospacers. Bacteria grow at rate gC. With rate α, a phage interacts with a bacterium. With probability p_V, the phage will kill bacteria without spacers and produce a burst of new phages with size B, while for bacteria with spacers that probability is reduced to $p_v^s = (1-e)p_V$ ($0 \leq e \leq 1$). Bacteria without spacers that survive an attack have a chance to acquire a spacer with probability η. Bacteria with a spacer lose their spacer at rate r. Parameter descriptions and default values are shown in SI Table 1.

2.0.1 Reactions

Table 2 lists all the interactions present in our model between individual bacteria (b), phages (V) and nutrients (C).

2.1 Master equation

The reactions in Table 2 can be formulated as a master equation describing the probability of observing n_b^0 bacteria without spacers, the set n_b^i bacteria with spacers of type i, n_V phages, and a nutrient

Table 2: Model reactions

$b^{0,i} + C \xrightarrow{g} 2b^{0,i}$	bacterium divides
$b^{0,i} \xrightarrow{F} \emptyset$	bacterium flows out
$V \xrightarrow{F} \emptyset$	phage flows out
$\emptyset \xrightarrow{FC_0} C$	nutrients flow in
$C \xrightarrow{F} \emptyset$	nutrients flow out
$b^0 + V \xrightarrow{\alpha p_V} BV$	interaction, phage wins
$b^0 + V \xrightarrow{\alpha(1-p_V)(1-\eta)} b^0$	interaction, bacterium survives
$b^0 + V \xrightarrow{\alpha(1-p_V)\eta/m} b^i$	interaction, bacterium survives and acquires a spacer
$b^i + V \xrightarrow{\alpha p_v^s} BV$	interaction, phage wins
$b^i + V \xrightarrow{\alpha(1-p_v^s)} b^i$	interaction, bacterium survives
$b^i \xrightarrow{r} b^0$	bacterium loses spacer

concentration of C at time t (equation 2).

$$\begin{aligned}
\frac{dP(n_b^0, \{n_b^i\}, n_V, C, t)}{dt} &= g(C+1)(n_b^0 - 1)P(n_b^0 - 1, \{n_b^i\}, n_V, C+1, t) \\
&+ \sum_{j=1}^{m} g(C+1)(n_b^j - 1)P(n_b^0, \{n_b^{i\neq j}\}, n_b^j - 1, n_V, C+1, t) \\
&+ F(n_b^0 + 1)P(n_b^0 + 1, \{n_b^i\}, n_V, C, t) \\
&+ \sum_{j=1}^{m} F(n_b^j + 1)P(n_b^0, \{n_b^{i\neq j}\}, n_b^j + 1, n_V, C, t) \\
&+ F(n_V + 1)P(n_b^0, \{n_b^i\}, n_V + 1, C, t) \\
&+ F(C+1)P(n_b^0, \{n_b^i\}, n_V, C+1, t) \\
&+ FC_0 P(n_b^0, \{n_b^i\}, n_V, C-1, t) \\
&+ \alpha p_V (n_b^0 + 1)(n_V - B + 1)P(n_b^0 + 1, \{n_b^i\}, n_V - B + 1, C, t) \\
&+ \alpha(1 - p_V)(1 - \eta)n_b^0(n_V + 1)P(n_b^0, \{n_b^i\}, n_V + 1, C, t) \\
&+ \sum_{j=1}^{m} \frac{\alpha(1 - p_V)\eta}{m}(n_b^0 + 1)(n_V + 1)P(n_b^0 + 1, \{n_b^{i\neq j}\}, n_b^j - 1, n_V + 1, C, t) \\
&+ \sum_{j=1}^{m} \alpha p_v^s (n_b^j + 1)(n_V - B + 1)P(n_b^0, \{n_b^{i\neq j}\}, n_b^j + 1, n_V - B + 1, C, t) \\
&+ \sum_{j=1}^{m} \alpha(1 - p_v^s)n_b^j(n_V + 1)P(n_b^0, \{n_b^{i\neq j}\}, n_b^j, n_V + 1, C, t) \\
&+ \sum_{j=1}^{m} r(n_b^j + 1)P(n_b^0 - 1, \{n_b^{i\neq j}\}, n_b^j + 1, n_V, C, t) \\
&- \Bigg(F(n_b^0 + \sum_{j=1}^{m} n_b^j + n_V + C + C_0) + gC(n_b^0 + \sum_{j=1}^{m} n_b^j) \\
&+ \alpha n_V(n_b^0 + \sum_{j=1}^{m} n_b^j) + r\sum_{j=1}^{m} n_b^j \Bigg) P(n_b^0, \{n_b^i\}, n_V, C, t)
\end{aligned} \tag{2}$$

The 1st term is included only for $n_b^0 > 1$, the 2nd term if $n_b^j > 1$, the 7th term for $C \geq 1$, 8th term if $n_V > B-1$, the 10th term for $n_b^j \geq 1$, the 11th term for $n_V > B-1$ and the 13th term for $n_b^0 \geq 1$.

2.2 Mean-field dynamics

We can also write equations for the averages of the microscopic quantities (equations 3 to 6).

2.2.1 Microscopic equations

$$\frac{d\left\langle n_b^0\right\rangle}{dt} = -F\left\langle n_b^0\right\rangle + g\left\langle C n_b^0\right\rangle - \alpha p_V\left\langle n_b^0 n_V\right\rangle - \alpha(1-p_V)\eta\left\langle n_b^0 n_V\right\rangle + \sum_{j=1}^{m} r\left\langle n_b^j\right\rangle \tag{3}$$

$$\frac{d\left\langle n_b^j\right\rangle}{dt} = -F\left\langle n_b^j\right\rangle + g\left\langle C n_b^j\right\rangle - \alpha p_v^s\left\langle n_b^j n_V\right\rangle - r\left\langle n_b^j\right\rangle + \frac{\alpha(1-p_V)\eta}{m}\left\langle n_b^0 n_V\right\rangle \tag{4}$$

$$\begin{aligned}\frac{d\left\langle n_V\right\rangle}{dt} =& -F\left\langle n_V\right\rangle + \alpha p_V(B-1)\left\langle n_b^0 n_V\right\rangle - \alpha(1-p_V)\left\langle n_b^0 n_V\right\rangle + \\ & \sum_{j=1}^{m}\alpha p_v^s(B-1)\left\langle n_b^j n_V\right\rangle - \sum_{j=1}^{m}\alpha(1-p_v^s)\left\langle n_b^j n_V\right\rangle\end{aligned} \tag{5}$$

$$\frac{d\left\langle C\right\rangle}{dt} = F(\left\langle C\right\rangle - C_0) - g\left\langle C\left(n_b^0 + \sum_{j=1}^{m} n_b^j\right)\right\rangle \tag{6}$$

We approximate the correlations $\langle XY\rangle \approx \langle X\rangle\langle Y\rangle$.

$$\frac{d\left\langle n_b^0\right\rangle}{dt} = -F\left\langle n_b^0\right\rangle + g\left\langle C\right\rangle\left\langle n_b^0\right\rangle - \alpha p_V\left\langle n_b^0\right\rangle\left\langle n_V\right\rangle - \alpha(1-p_V)\eta\left\langle n_b^0\right\rangle\left\langle n_V\right\rangle + \sum_{j=1}^{m} r\left\langle n_b^j\right\rangle \tag{7}$$

$$\frac{d\left\langle n_b^j\right\rangle}{dt} = -F\left\langle n_b^j\right\rangle + g\left\langle C\right\rangle\left\langle n_b^j\right\rangle - \alpha p_v^s\left\langle n_b^j\right\rangle\left\langle n_V\right\rangle - r\left\langle n_b^j\right\rangle + \frac{\alpha(1-p_V)\eta}{m}\left\langle n_b^0\right\rangle\left\langle n_V\right\rangle \tag{8}$$

$$\begin{aligned}\frac{d\left\langle n_V\right\rangle}{dt} =& -F\left\langle n_V\right\rangle + \alpha p_V(B-1)\left\langle n_b^0\right\rangle\left\langle n_V\right\rangle - \alpha(1-p_V)\left\langle n_b^0\right\rangle\left\langle n_V\right\rangle + \\ & \sum_{j=1}^{m}\alpha p_v^s(B-1)\left\langle n_b^j\right\rangle\left\langle n_V\right\rangle - \sum_{j=1}^{m}\alpha(1-p_v^s)\left\langle n_b^j\right\rangle\left\langle n_V\right\rangle\end{aligned} \tag{9}$$

$$\frac{d\left\langle C\right\rangle}{dt} = F(\left\langle C\right\rangle - C_0) - g\left\langle C\right\rangle\left(\left\langle n_b^0\right\rangle + \sum_{j=1}^{m}\left\langle n_b^j\right\rangle\right) \tag{10}$$

Then, we replace means by deterministic variables n_b^0, n_b^j, n_V, and C.

$$\frac{dn_b^0}{dt} = -Fn_b^0 + gCn_b^0 - \alpha p_V n_b^0 n_V - \alpha(1-p_V)\eta n_b^0 n_V + \sum_{j=1}^{m} r n_b^j \tag{11}$$

$$\frac{dn_b^j}{dt} = -Fn_b^j + gCn_b^j - \alpha p_v^s n_b^j n_V - r n_b^j + \frac{\alpha(1-p_V)\eta}{m} n_b^0 n_V \tag{12}$$

$$\begin{aligned}\frac{dn_V}{dt} &= -Fn_V + \alpha p_V(B-1)n_b^0 n_V - \alpha(1-p_V)n_b^0 n_V \\ &+ \sum_{j=1}^{m} \alpha p_v^s (B-1) n_b^j n_V - \sum_{j=1}^{m} \alpha(1-p_v^s) n_b^j n_V\end{aligned} \tag{13}$$

$$\frac{dC}{dt} = F(C - C_0) - gC\left(n_b^0 + \sum_{j=1}^{m} n_b^j\right) \tag{14}$$

2.2.2 Macroscopic equations

We can define new variables, $n_b^s = \sum_{j=1}^{m} n_b^j$, $n_B = n_b^0 + n_b^s$, $\nu = n_B^s/n_B$ $(1 - \nu = n_B^0/n_B)$, and $p_V^s = (1-e)p_V$.

$$\frac{dn_b^0}{dt} = -Fn_b^0 + gCn_b^0 - \alpha p_V n_b^0 n_V - \alpha(1-p_V)\eta n_b^0 n_V + rn_b^s \tag{15}$$

$$\frac{dn_b^s}{dt} = -Fn_b^s + gCn_b^s - \alpha(1-e)p_V n_b^s n_V - rn_b^s + \alpha(1-p_V)\eta n_b^0 n_V \tag{16}$$

$$\frac{dn_V}{dt} = -Fn_V - \alpha n_B n_V + \alpha p_V(1-e\nu)Bn_B n_V \tag{17}$$

$$\frac{dC}{dt} = F(C - C_0) - gCn_B \tag{18}$$

$$\frac{dn_B}{dt} = -Fn_B + gCn_B - \alpha p_V(1-e\nu)n_B n_V \tag{19}$$

2.3 Description of simulations

Simulations were written in C++ and performed on a Lenovo ideapad Y700 and on SciNet. We primarily used the tau leaping method [3] and compared with Gillespie simulations for some cases. Both methods showed good agreement for the mean-field behaviour of bacteria and phages (SI Figure 5) and produced the same qualitative behaviour for individual spacer types (SI Figure 6).

2.4 Parameter choices

Burst size for phage that target S. thermophilus is between 140-200 [4]. The rate of adsorption for phage is of the order of 10^{-8} min^{-1} ml [5]. Using a volume of $V = 50ml$, our total adsorption rate is $\alpha = 2 \times 10^{-10}$ min^{-1} per bacteria and phage.

[6] measured the maximum growth rate of S. thermophilus in milk at 42°C to be 2.4×10^{-2} min^{-1}. This corresponds to gC_0 in our model.

The other parameters were picked in order to get a stable fixed point where phage and bacteria coexist, with population sizes relevant to experiments such as [1].

2.5 Simulation results

Our simulations were performed with a maximum of $m = 500$ spacer types that can be acquired by bacteria. This upper limit on the number of spacer types limits the total diversity of spacer types that can be observed and only impacts the spacer abundance distribution at large η. The qualitative simulation results, namely a continuous turnover of individual spacers and the presence of a non-trivial steady-state spacer abundance distribution, are insensitive to the choice of η provided not all m spacer

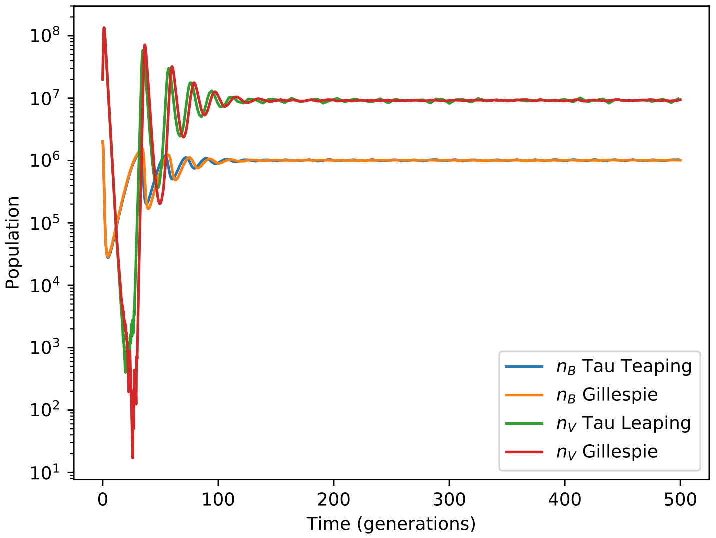

Figure 5: Total bacteria (n_B) and total phage (n_V) as a function of time for a Gillespie simulation and a tau leaping simulation. The two simulation techniques produce very similar results.

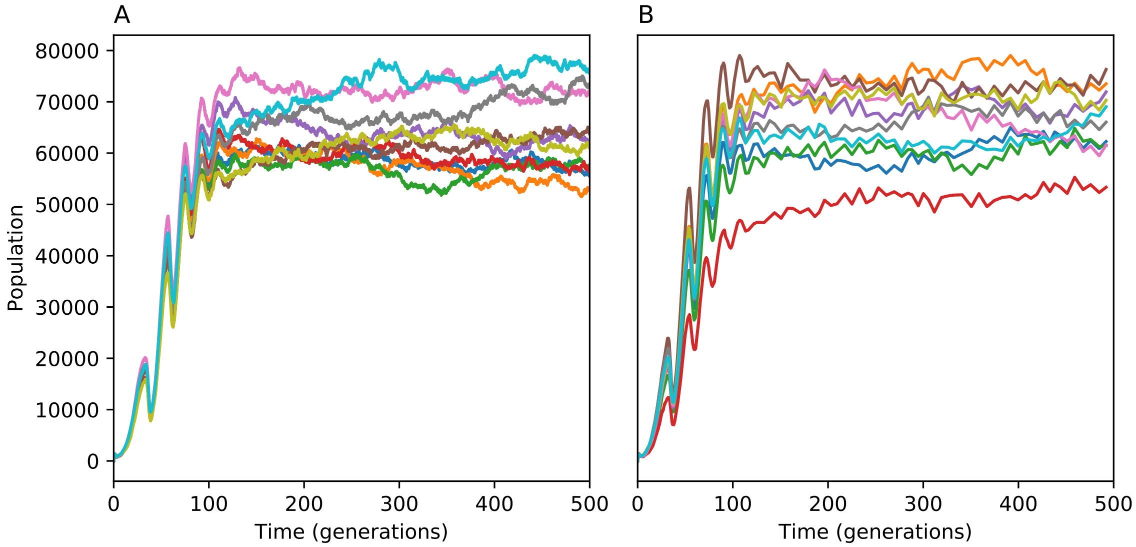

Figure 6: Comparison of individual spacer type trajectories using tau leaping and Gillespie simulation techniques. (A) 10 spacer type trajectories vs time using Gillespie simulation methods. (B) 10 spacer type trajectories vs time using tau leaping simulation methods.

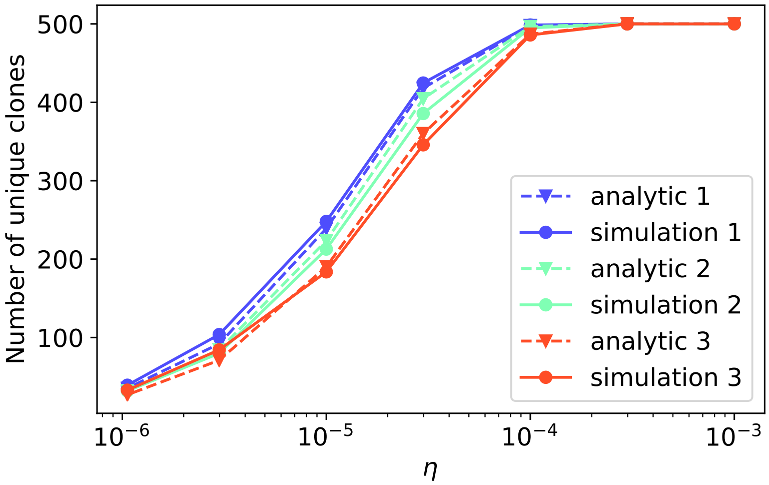

Figure 7: The average number of unique spacer types present at steady-state in simulations (circles and solid lines) increases with increasing η. The simulation results are well-matched by the analytic prediction from equation 29 (triangles and dashed lines). The parameters η and e were chosen for each simulation so that all points on each colored curve correspond to a constant total bacterial population size of $0.15C_0$ (blue points), $0.1C_0$ (green points) and $0.05C_0$ (red points).

types are acquired. This puts an upper bound on η of $\approx 10^{-4}$ in our simulation, but simulations with higher η can be performed with large values of m. SI Figure 7 shows the average total number of unique spacer types at steady state as a function of η for simulation data, compared to the analytic prediction given by equation 29: the number of unique spacer types predicted at steady state is $\sum_k b_k$.

We initialized each simulation with no bacteria with spacers, or in other words the rank-abundance distribution is uniform at 0 abundance at the start of all simulations. The steady-state distribution evolves from a very different shape at early times. SI Figure 8 shows the spacer rank-abundance distribution for various time points of a simulation run.

2.6 Origin of rank-abundance curve

The spacer rank-abundance distribution resulting from our simulations can be analytically derived from the following master equation, which describes b_k, the number of spacer types, or clones, of size k. The size of a clone can increase through bacterial division with rate gC (first term) or decrease through flow (F), spacer loss (r), and phage predation ($\alpha n_V p_v(1-e)$). The third term in equation 20 describes spacer acquisition: since in our simulations the total number of protospacers is fixed at $m = 500$, a newly acquired spacer will be added to an existing clone of size k with probability η/m, where η is the probability of acquiring any spacer in an interaction in which the phage does not succeed (which happens with probability $1 - p_V$).

$$\begin{aligned}\partial_t b_k = gC[(k-1)b_{k-1} - kb_k] + (F + r + \alpha n_V p_V(1-e))[(k+1)b_{k+1} - kb_k] \\ + \alpha n_b^0 n_V (1-p_V)\frac{\eta}{m}[b_{k+1} - b_k]\end{aligned} \tag{20}$$

The variables n_V, n_B^0, and C evolve according to their mean-field equations (15, 17, 18). The total number of bacteria with spacers $n_B^s = \sum_k kb_k$; $\partial_t \sum_k kb_k$ is equivalent to equation 16.

At steady-state, all the population variables are constant, and equation 20 can be solved using a generating function and the method of characteristics.

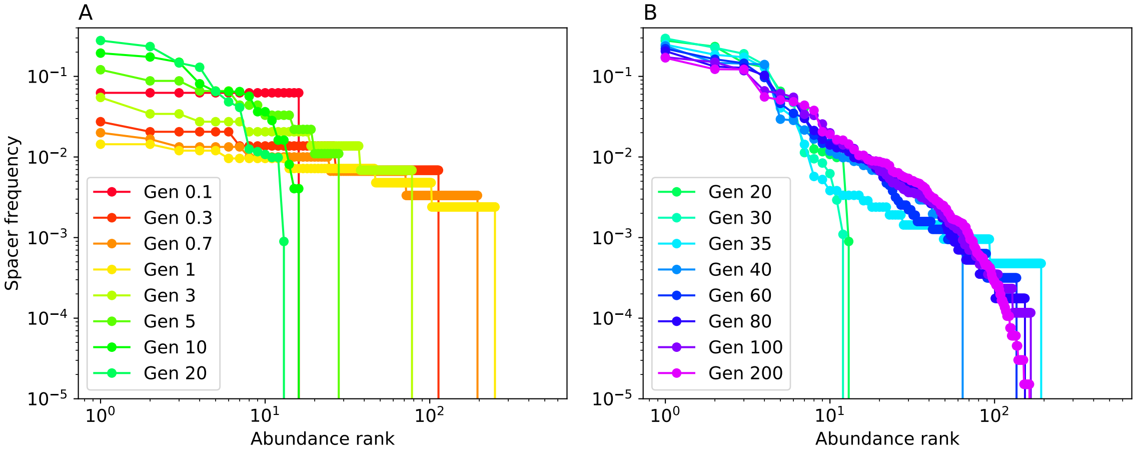

Figure 8: Spacer rank-abundance distributions for a simulation with $\eta = 10^{-5}$ and $e = 0.387$. Time is rescaled into units of bacterial generations. (A) The distribution at early times begins as a flat distribution with a few bacteria having a single spacer (0.1 generations). As time progresses, more bacteria acquire spacers and some spacer types grow to larger sizes, making the distribution steeper and broader. (B) At longer times, the distribution reaches its steady-state shape at about generation 200.

2.6.1 Generating function solution

The generating function for the probability distribution $b_k(t)$ is $G(z,t) = \sum_k z^k b_k(t)$.

Let $\beta = gC$, $\mu = F + r + \alpha n_V p_V (1-e)$, and $D = \alpha\eta n_B^0 n_V (1-p_V)$. Multiplying equation 20 by $\sum_k z^k$ and noting that $\partial_z G(z,t) = \sum_k k z^{k-1} b_k(t)$, we get the following differential equation:

$$\partial_t G(z,t) = \partial_z G(z,t)\left(z^2\beta - z(\beta+\mu) + \mu\right) + G(z,t)\frac{D}{m}(z-1) \tag{21}$$

Equation 21 can be solved with the method of characteristics [7]. We parametrize the function $G(z,t)$ with a new variable s. Applying the chain rule:

$$\partial_s G(z(s), t(s)) = \frac{\partial G}{\partial z}\frac{\partial z}{\partial s} + \frac{\partial G}{\partial t}\frac{\partial t}{\partial s} \tag{22}$$

And by comparison with equation 21, the characteristic equations are

$$\frac{\partial t}{\partial s} = 1 \tag{23}$$

$$\frac{\partial z}{\partial s} = (1-z)(\beta z - \mu) \tag{24}$$

$$\frac{\partial G}{\partial s} = G\frac{D}{m}(z-1) \tag{25}$$

From equation 23 we see $t = s + c_1$, so we can choose $t_0 = c_1 = 0$ and replace s with t going forward.

Solving the characteristic equation for z by integrating both sides gives equation 26.

$$\frac{1-z}{\mu - \beta z}\mathrm{e}^{(\beta-\mu)t} = c_2 \tag{26}$$

At $t = 0$, z will pass through some point z_0, so we have the initial condition $z(0) = z_0$. With z_0 in equation 26 at $t = 0$, we get equation 27, where c_2 is given by equation 26.

$$z_0 = \frac{c_2\mu - 1}{c_2\beta - 1} \tag{27}$$

The variation of G along the $z - t$ curve is

$$\frac{\partial G}{\partial z} = -\frac{\frac{GD}{m}(z-1)}{z^2\beta - z(\beta+\mu) + \mu} = -\frac{GD}{m(\beta z - \mu)}$$

Integrating both sides, we get

$$G(z) = \Omega(c_2)(\beta z - \mu)^{-\frac{D}{\beta m}}$$

The constant Ω is a function of the characteristic z-t curve (equation 26). To find the particular form of $\Omega(c_2)$, we apply the initial condition $G(z, 0) = zN_0$, meaning that we start with N_0 clones of size 1 at time $t = 0$.

$$G(z,0) = zN_0 = \Omega\left(\frac{1-z}{\mu - \beta z}\right)(\beta z - \mu)^{-\frac{D}{\beta m}}$$

Let $\xi = \frac{1-z}{\mu-\beta z}$, therefore $z = \frac{\xi\mu-1}{\xi\beta-1}$.

$$\Omega(\xi)(\beta\left(\frac{\xi\mu-1}{\xi\beta-1}\right) - \mu)^{-\frac{D}{\beta m}} = \left(\frac{\xi\mu-1}{\xi\beta-1}\right)N_0$$

Solving for $\Omega(\xi)$:

$$\Omega(\xi) = \left(\frac{\xi\mu-1}{\xi\beta-1}\right)N_0(\beta\left(\frac{\xi\mu-1}{\xi\beta-1}\right) - \mu)^{\frac{D}{\beta m}}$$

The full solution for $G(z, t)$ can be written by replacing the constant $\Omega(c_2)$ with the expression for $\Omega(\xi)$ and replacing ξ with $\xi\epsilon$, where $\epsilon = \mathrm{e}^{(\beta-\mu)t}$ is the time-dependent part of the $z - t$ curve.

$$G(z,t) = N_0(\beta z - \mu)^{-\frac{D}{\beta m}}\left(\frac{\xi\epsilon\mu-1}{\xi\epsilon\beta-1}\right)(\beta\left(\frac{\xi\epsilon\mu-1}{\xi\epsilon\beta-1}\right) - \mu)^{\frac{D}{\beta m}}$$

Finally, replacing ξ with $\frac{1-z}{\mu-\beta z}$, we get

$$G(z,t) = N_0(\beta z - \mu)^{-\frac{D}{\beta m}}\left(\frac{(1-z)\epsilon\mu + \beta z - \mu}{(1-z)\epsilon\beta + \beta z - \mu}\right)(\beta\left(\frac{(1-z)\epsilon\mu + \beta z - \mu}{(1-z)\epsilon\beta + \beta z - \mu}\right) - \mu)^{\frac{D}{\beta m}}$$

$G(1, t) = N_0$, meaning that the total population remains conserved, consistent with our assumption that all the population variables are at steady-state.

The limit as $t \to \infty$ of $G(z, t)$ is

$$G(z) = N_0\left(\frac{\beta z - \mu}{\beta - \mu}\right)^{-\frac{D}{\beta m}}$$

We can construct b_k by taking successive derivatives of $G(z)$: $b_k = \frac{1}{k!}\frac{\partial G}{\partial z}|_{z=0}$

$$b_k = \frac{N_0\prod_{i=1}^{k}[D/m + (i-1)\beta](\frac{\mu-\beta}{\mu})^{D/(\beta m)}}{k!\mu^k} \tag{28}$$

$$b_0 = N_0\left(\frac{\mu-\beta}{\mu}\right)^{D/(\beta m)}$$

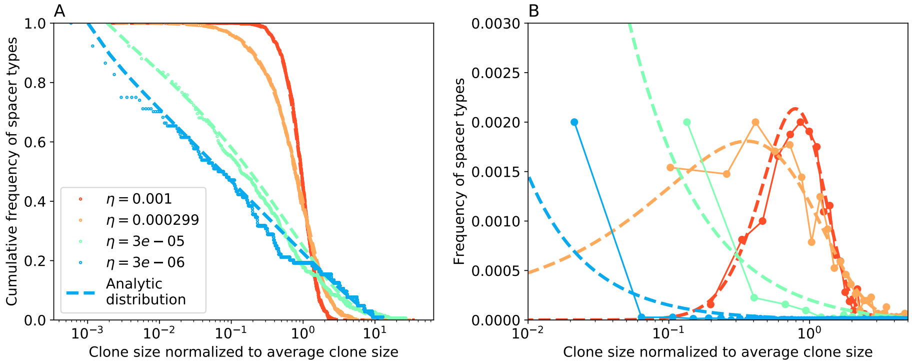

Figure 9: Equation 29 (dashed lines) compared with spacer clone size distributions from simulations at steady-state (dots). (A) Distributions in cumulative form. (B) Distributions shown as a histograms (dots). The analytic steady-state distribution matches well with the simulation results, except in the cases where the acquisition rate η is very low and the choice of bin size has a large effect. Here $N_0 = m$, the total number of unique clones.

We can re-write this expression using Stirling's approximation for $k!$ to facilitate evaluation at large k.

$$b_k = \frac{N_0}{\sqrt{2\pi k}}\exp\left[\frac{D}{\beta m}\ln\left(\frac{\mu-\beta}{\mu}\right) + \sum_{i=1}^{k}\ln\left(\frac{e}{k\mu}(\frac{D}{m} + (i-1)\beta)\right)\right] \tag{29}$$

Equation 29 is an analytic expression describing the steady-state spacer abundance distribution that results from our simulations. SI Figure 9 compares the analytic distribution to the steady-state spacer clone size distribution from our simulations at several values of the spacer acquisition probability η, with e chosen such that the total number of bacteria is the same for all cases. The corresponding rank-abundance distribution can be obtained from the cumulative distribution (SI Figure 9A) by flipping the axes and rescaling the frequency axis.

2.6.2 Rank-abundance distribution in ecology

The steady-state clone size distribution given in equation 29 can be approximated for large clone size k and large m to give a gamma distribution and logseries distribution respectively, both of which have a long history as descriptions of species abundance in ecology [8].

In the following expressions we replace N_0 with m, since at steady-state the total number of clones remains fixed at m.

We find the following expression for b_k in the limit of large clone size (large k) by taking a series expansion as $k \to \infty$ and keeping the first term.

$$b_k \approx \frac{m\left(1-\frac{\beta}{\mu}\right)^{\frac{D}{\beta m}}}{\Gamma\left(\frac{D}{\beta m}\right)} \mathrm{e}^{-\ln(\mu/\beta)k}\left(\frac{1}{k}\right)^{1-\frac{D}{\beta m}} \tag{30}$$

This is a gamma distribution with shape parameter $\frac{D}{\beta m}$ and rate parameter $\ln(\mu/\beta)$. Note that $\left(1-\frac{\beta}{\mu}\right)^{\frac{D}{\beta m}} \approx \ln(\mu/\beta)^{\frac{D}{\beta m}}$, consistent with the canonical form of the gamma distribution. The additional factor of m in equation 30 arises because we treat b_k as the *number* of clones of size k; to normalize b_k we would divide by m, the total possible number of unique clones.

The gamma distribution has been used to describe species abundance in a number of ecological situations [9, 10, 11, 12]. For example, Dennis and Patil [9] arrive at a gamma distribution as “the approximate stationary distribution for the abundance of a population fluctuating around a stable equilibrium,” and Plotkin and Muller-Landau [12] use a gamma distribution to fit species abundance distributions on a tropical island.

For practical purposes the gamma distibution given by equation 30 is a good approximation to the true distribution for all the parameter values we considered in our simulation.

When the total number of unique spacer types m is large, our model is effectively an infinite alleles model in which each newly acquired spacer is assumed to be completely unique. In the limit of large m, we find the following expression for b_k.

$$b_k \approx \frac{D}{\beta}\frac{1}{k}\left(\frac{\beta}{\mu}\right)^k \tag{31}$$

Up to a constant, this is a log-series distribution, made famous by Fisher et al. [13] and appearing many times since [14].

3 Mean-field steady-state solutions

3.1 $e = 0$ model (no adaptive immunity)

Equations 15 to 18 describe the full model. If spacer effectiveness $e = 0$, the model reduces to three dimensions: bacteria n_B, phages n_V, and nutrients C. Equations 32 to 34 describe this simpler model.

$$\frac{dn_V}{dt} = -\alpha n_B n_V + \alpha B p_V n_B n_V - F n_V \tag{32}$$

$$\frac{dn_B}{dt} = gCn_B - \alpha p_V n_V n_B - F n_B \tag{33}$$

$$\frac{dC}{dt} = FC_0 - gCn_B - FC \tag{34}$$

Solving equations 32 to 34 at steady state gives the following fixed points.

3.1.1 Trivial fixed point

There is a trivial fixed point where bacteria and phages are both zero.

$$n_B^* = 0$$

$$n_V^* = 0$$

$$C^* = C_0$$

The eigenvalues of the Jacobian at this fixed point are $1 - f, -f$, and $-f$, where $f = F/(gC_0)$. This means that this fixed point is stable for $f > 1$. $f > 1$ is a reasonable stability condition: this is the case where the flow rate is too high for bacteria to persist.

3.1.2 Phages unable to persist

$$n_B^* = C_0(1-f)$$

$$n_V^* = 0$$

$$C^* = C_0 f$$

$0 < f < 1$ is required for physical existence of this fixed point.

The eigenvalues of the Jacobian at this fixed point are $f-1$, $-f$, and $-\frac{(f-1)p(Bp_V-1)+fp_V}{p_V}$, where $p = p_V\alpha/g$. The first two are negative under the requirement for existence. The third is negative for $Bp_V < \frac{gf}{(1-f)\alpha} + 1$. If this stability condition is satisfied, phages cannot persist in the population — they will be driven to extinction.

3.1.3 All populations finite and stable

If all variables are non-zero, the fixed point is

$$\frac{n_B^*}{C_0} = \frac{fp_V}{p(-1+Bp_V)}$$

$$\frac{n_V^*}{C_0} = \frac{(1-f)p(Bp_V-1)-fp_V}{p(p(Bp_V-1)+p_V)}$$

$$\frac{C^*}{C_0} = \frac{p(Bp_V-1)}{p(Bp_V-1)+p_V}$$

The condition for existence is

$$Bp_V > \frac{gf}{(1-f)\alpha} + 1$$

The eigenvalues are

$$-f$$

$$-\frac{\sqrt{f}\sqrt{4(f-1)p^2(Bp_V-1)^2+4fpp_V(Bp_V-1)+fp_V^2}+fp_V}{2p(Bp_V-1)}$$

$$\frac{\sqrt{f}\sqrt{4(f-1)p^2(Bp_V-1)^2+4fpp_V(Bp_V-1)+fp_V^2}-fp_V}{2p(Bp_V-1)}$$

The first is always negative. The second is negative for

$$\frac{4p^2(Bp_V-1)^2}{(2p(Bp_V-1)+p_V)^2} \leq f < 1$$

The third is negative for

$$\frac{4p^2(Bp_V-1)^2}{(2p(Bp_V-1)+p_V)^2} \leq f < \frac{p(Bp_V-1)}{p(Bp_V-1)+p_V} = \frac{C^*}{C_0}$$

The upper limit on f is the same as the existence condition (requiring all be solutions > 0).

3.2 Nonlinear bacterial growth rate

Instead of the growth rate for n_B being gC, we check what happens when the growth rate is a Hill function of the form $\frac{gkC}{C+k}$, where k is the nutrient concentration at which bacterial growth rate is at half maximum. If $k >> C$, the linear approximation used in our results is valid and $\frac{gkC}{C+k} \approx gC$.

Solving for the non-trivial steady-state variables in the case when bacteria have no CRISPR spacers, we find that n_B^* is unchanged:

$$n_B^* = \frac{F}{\alpha(Bp_V - 1)} \tag{35}$$

C^* and n_V^*, however, now depend on k:

$$C^* = \frac{1}{2}\left(C_0 - k - \frac{gkn_B^*}{F}\right) + \frac{1}{2}\sqrt{\left(C_0 - k - \frac{gkn_B^*}{F}\right)^2 + 4C_0k} \tag{36}$$

$$n_V^* = \frac{gkC^*}{C^* + k} - F \tag{37}$$

This solution for C reduces to the linear growth rate solution (section 3.1.3) when k is large. This can be seen by expanding the square root in C^* and keeping terms up to order $\frac{1}{k^3}$:

$$\frac{C^*}{C_0} \approx 1 - \frac{1}{p(B - 1/p_V)} + \frac{1}{p^2(B - 1/p_V)^2} \approx \frac{p(B - 1/p_V)}{p(B - 1/p_V) + 1}$$

The stability condition for bacteria and phage coexistence now depends on k. k must be greater than the following parameter combination in order for phages to persist.

$$k > \frac{F(F + \alpha(C_0 - BC_0p_V))}{Fg + \alpha(F - C_0g)(Bp_V - 1)} = C_0\frac{f(fg + \alpha(1 - Bp_V))}{fg + \alpha(1 - f)(1 - Bp_V)}$$

For $f = F/(gC_0) = 0.1, B = 170, p_V = 0.02, g = 2.4 \times 10^{-11}, C_0 = 10^9$, and $\alpha = 2 \times 10^{-10}$, k/C_0 must be greater than ≈ 0.11. SI Figure 10 compares the full nonlinear growth solutions (equations 35 to 37) to the solutions for linear growth (equations 3.1.3). Provided k is large enough that phages can persist, the picture is not qualitatively different, and in the low-nutrient limit ($k >> C$), the two solutions are very nearly the same.

3.3 $e \neq 0$ model (adaptive immunity)

If $e > 0$, then the system is fully four-dimensional and all four variables are coupled.

3.3.1 Trivial fixed points

The two partially trivial fixed points are the same as in the case when $e = 0$, since if $n_V = 0$, then $\nu = n_B^s/n_B = 0$ at steady state. The stability and existence conditions are also the same; effectively ν becomes uncoupled and the system is reduced to three dimensions if $n_V = 0$.

3.3.2 Non-trivial fixed point

For convenience we define rescaled population sizes $x = n_B/C_0$, $y = n_V/C_0$, and $z = C/C_0$. Solving in the case where all dynamical variables are non-trivial, we get

$$z^* = \frac{p(Bp_V(e\nu^* - 1) + 1)}{p(Bp_V(e\nu^* - 1) + 1) - p_V} \tag{38}$$

$$x^* = \frac{fp_V}{p}\frac{1}{Bp_V(1 - e\nu^*) - 1} \tag{39}$$

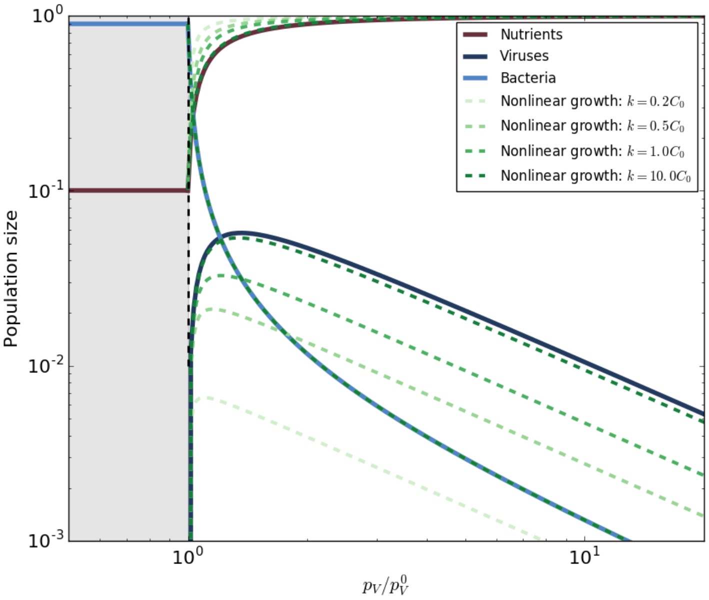

Figure 10: Solid lines: solutions to equations 32 to 34 with linear growth for n_B. Green dashed lines: solutions to equations 35 to 37 for different values of k.

$$y^* = \frac{(f-1)p(Bp_V(e\nu^*-1)+1)-fp_V}{p(e\nu^*-1)(p(Bp_V(e\nu^*-1)+1)-p_V)} \tag{40}$$

And an implicit cubic equation for ν, where $R = r/(gC_0)$.:

$$\begin{aligned} 0 = (1-\nu)\left[-p_V\nu e - \eta(1-p_V)\right]\left[(1-f)p(p_V B(1-e\nu)-1)-fp_V\right] \\ +R\nu p_V(1-e\nu)(Bpp_V(1-e\nu)-p+p_V) \end{aligned} \tag{41}$$

This cubic equation is analytically solvable, but the full solutions in terms of all parameters are cumbersome.

Only one of the three solutions of equation 41 is physical in the parameter range we use (real-valued and properly bounded):

$$\begin{aligned} \nu^* = &-\frac{\left(1+i\sqrt{3}\right)\sqrt[3]{\sqrt{\left(-27a^2d+9abc-2b^3\right)^2+4\left(3ac-b^2\right)^3}-27a^2d+9abc-2b^3}}{6\sqrt[3]{2}a} \\ &+\frac{\left(1-i\sqrt{3}\right)\left(3ac-b^2\right)}{3\;2^{2/3}a\sqrt[3]{\sqrt{\left(-27a^2d+9abc-2b^3\right)^2+4\left(3ac-b^2\right)^3}-27a^2d+9abc-2b^3}}-\frac{b}{3a} \end{aligned} \tag{42}$$

where the coefficients are

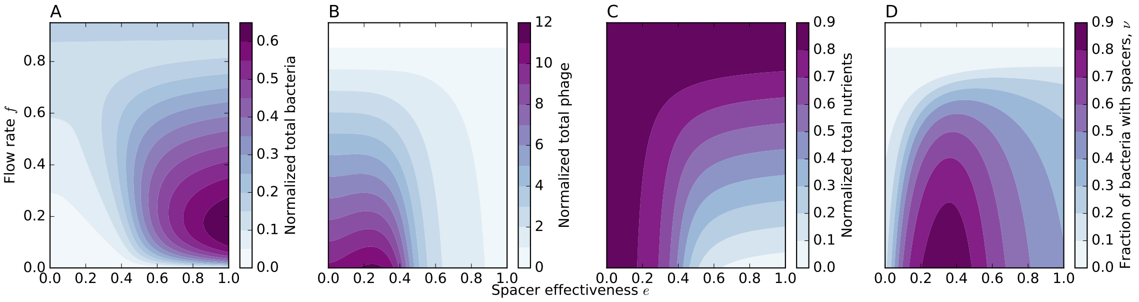

Figure 11: x, y, z, and ν (A-D respectively) vs. f and e with $R = 0.04$, $\eta = 0.0001$, $B = 170$, $p_V = 0.01$, $\alpha = 2 \times 10^{-10}$, and $gC_0 = 0.024$.

$$a = Be^2 fpp_V^2(f + R - 1) \tag{43}$$

$$\begin{aligned} b = &-efp_V(p(f(B(p_V(e + \eta + 1) - \eta) - 1) \\ &+ B(\eta - p_V(e + \eta - 2R + 1)) - R + 1) + p_V(f + R)) \end{aligned} \tag{44}$$

$$\begin{aligned} c = &fp\left[Bp_V^2(e(f - 1)(\eta + 1) + (f - 1)\eta + R)\right. \\ &\left.- (e - 1)(f - 1)p_V(B\eta + 1) - (2B + 2)(f - 1)\eta p_V + (f - 1)\eta - p_V(f + R - 1)\right] \\ &+ fp_V(efp_V - f\eta + p_V(f\eta + R)) \end{aligned} \tag{45}$$

$$d = -f\eta(p_V - 1)((f - 1)p(Bp_V - 1) + fp_V) \tag{46}$$

Total bacteria, phage, nutrients, and the fraction of bacteria with spacers are plotted for a range of parameters in SI Figure 11 and SI Figure 14.

This fixed point is stable for a wide range of parameters, which we explored numerically. SI Figure 13 shows the number of negative eigenvalues vs. parameters; where all four eigenvalues have a negative real part, this fixed point is stable.

We observed that the minimum success probability $p_V^0 = \frac{1}{B}\left(\frac{gf}{(1-f)\alpha} + 1\right)$ required for phages to invade a bacterial culture is independent of e, which parametrizes adaptive immunity. To understand this, note that the fraction of bacteria with spacers ($\nu = n_B^s/n_B$) $= 0$ whenever n_V, the number of phages is 0, since spacers are continually lost with a small rate r but cannot be acquired if there are no phages. As a result, p_V^0 is independent of e since there are no bacteria with spacers at the point of phage extinction. SI Figure 12 shows ν and phages with and without adaptive immunity, illustrating that both ν and n_V go to zero at $p_V = p_V^0$.

3.3.3 Large α limit

For large α ($\alpha >> \alpha_0$, where $\alpha_0 = \frac{gf}{(1-f)(Bp_V-1)}$), we can find an approximate value of e, e^*, at which ν and n_V peak (equation 47). This solution is plotted as a yellow dashed line in SI Figure 14.

$$e^* = \frac{1}{Bp_V} + \frac{R(1 - Bp_V)}{Bp_V(f - 1)(1 - B(\eta + p_V) + B\eta p_V)} \tag{47}$$

4 Spacer dynamics

In our analysis of data from [1], we found that spacer abundance distributions were stable in time after three days and were broad, spanning four orders of magnitude. This distribution $\rho(v)$ is created

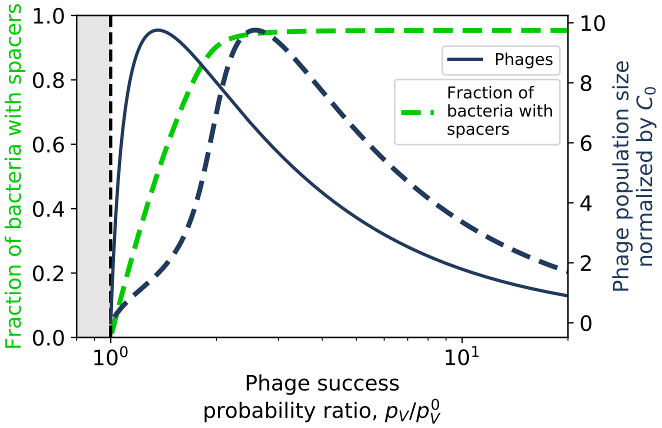

Figure 12: Phage and the fraction of bacteria with spacers at steady state as a function of the probability of phage success (p_V) for a model without CRISPR ($e = 0$, solid line) and for a model where bacteria have CRISPR systems and are able to acquire spacers ($e = 0.5$, dashed lines). The phage population size is normalized by the inflow nutrient concentration C_0 (y-axis labels on the right). Below $p_V = p_V^0$, phages cannot persist and the fraction of bacteria with spacers is 0.

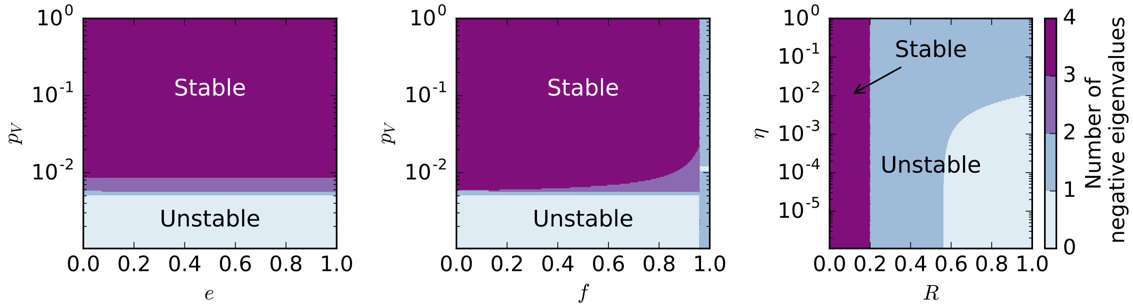

Figure 13: The number of eigenvalues with a negative real part for various parameter combinations (p_V and e, p_V and f, and R and η). The unstable regions in the first two plots reflect parameter combinations for which phages cannot persist. In the third plot, Equation 42 becomes unstable for large R, but one of the other roots takes its place as a stable and physical solution in this regime (confirmed numerically).

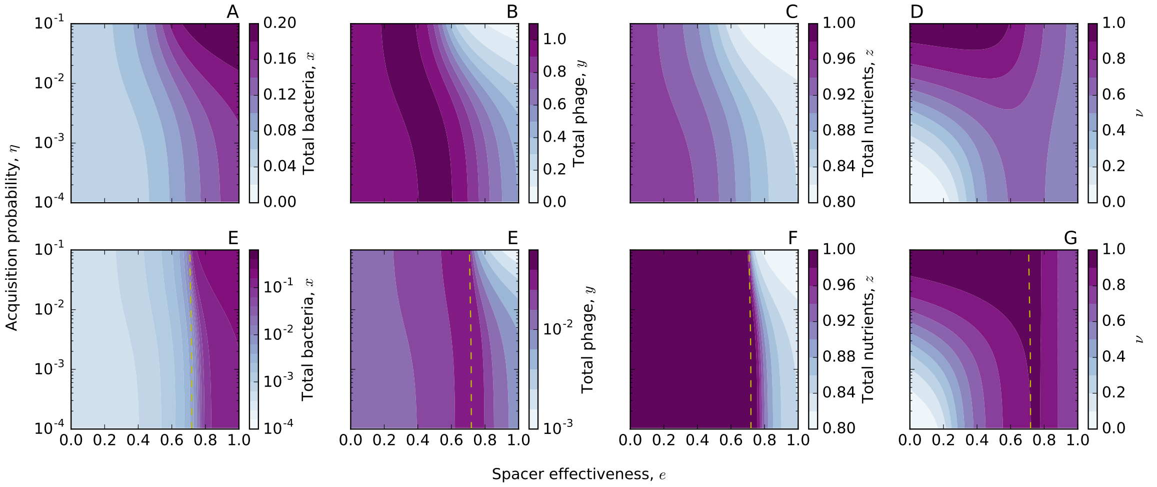

Figure 14: x, y, z, and ν vs. η and e for $\alpha \gtrsim \frac{gf}{(1-f)(Bp_V-1)}$ (top row) and $\alpha >> \frac{gf}{(1-f)(Bp_V-1)}$ (bottom row). The yellow dashed line is the approximate value of e at which both ν (G) and y (F) are maximized (equation 47) which agrees well with the full solution for large α.

by summing all spacer types of a particular abundance: $\rho(v) = \sum_i \delta(n_B^i - v)$. The normalized cumulative distribution, $\sum_v^\infty \rho(v)/\sum_0^\infty \rho(v)$, is plotted in SI Figure 15. The corresponding rank-abundance distribution is plotted in Figure 3C in the main text.

Individual spacer types experience continual turnover, both in our simulations and in experimental data from [1]. In the experimental data, both high-abundance and low-abundance spacers can change in abundance by an order of magnitude or more between time points, while in our simulations we find that the large abundance spacers are approximately stable once the system has reached a population-level steady state (SI Figure 16).

The observed turnover in large spacer types in the experimental data may reflect additional stochasticity not accounted for in our model, changes in fitness for individual spacer types over time, or the fact that the sequenced spacers are strongly undersampled. There are $\approx 10^8$ to 10^9 bacteria at the end of each day in the experiment, and there are $\approx 3 \times 10^4$ spacers recovered from sequencing each day. The data is undersampled by a factor of $\approx 10^4$, and apparent turnover may result from this.

SI Figure 17 compares the original simulation data with data undersampled by a factor of 10^2, 10^3, or 10^4. The mean fractional abundance over time for a particular type appears mostly unaffected by the undersampling, but there is indeed more variability when the degree of undersampling is higher. At an undersampling factor of 10^4, spacer counts are in the ones and tens, much lower than than counts of $\approx 10^3$ or 10^4 in the experimental data. Variability in the experimental data is over more orders of magnitude between time points than in the undersampled simulated data.

This undersampling of simulated data only considered that fewer organisms are sequenced than are present in the population and does not take into account that the experiment was performed with 100:1 serial dilutions and so each time point was seeded with a random subsample from the previous time point.

4.1 Time to extinction

To further investigate ongoing turnover in individual spacer types in our simulations and the experimental data, we calculated the mean time to extinction as a function of spacer abundance. Only data at steady state was used, beginning at Day 4 in the experimental data and generation 200 in the simulated data. For each spacer type that went extinct during the simulation or experiment, the time

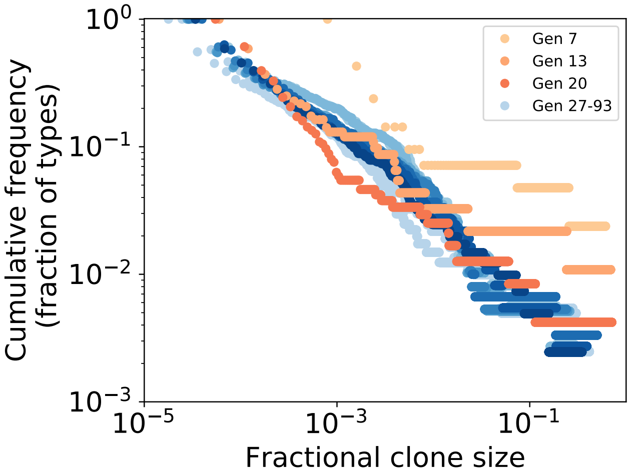

Figure 15: Cumulative frequency of spacer types (clones) as a function of normalized clone size. Darker blue indicates earlier times.

remaining to extinction was recorded as a function of its abundance at each time point after steady state, and the average and standard deviation over all types were calculated at each abundance. Figure 18 shows the standard deviation envelope for simulated data overlaid with experimental data, indicating that for both simulations and experiment spacers continue to experience turnover at steady-state and that the simulated time to extinction closely matches the experimental observations. Note that the longest observed time to extinction can never exceed the length of the simulation or experiment, meaning that shorter measurement windows will result in shorter average times to extinction. To illustrate this effect we calculate two time to extinction distributions for a short simulation of similar length to the experiment (generation 300 to 400) and a longer simulation (generation 300 to 500). Figure 19 shows that the mean time to extinction is finite even for high abundance spacers in the simulated data.

5 Regulation of CRISPR-Cas

5.1 Extent of bistability

We add regulation of CRISPR-Cas to our model by making spacer effectiveness e a function of bacterial cell density, assuming Cas expression to also be a sigmoidal function of cell density. Many bacterial behaviours controlled by quorum sensing are threshold-dependent: cells must switch between discrete states such as motile and non-motile, biofilm and free-living, virulent and non-virulent. In many quorum sensing systems, production of the autoinducer molecule is under positive feedback and increases nonlinearly with increasing cell density, and so many of the resulting changes in gene expression are switch-like [15]. For this reason we assume that spacer effectiveness depends strongly on cell density.

However, we observe bistability for a wide range of parameters and note that spacer effectiveness does not necessarily need to be a sharp function of x, where $x = n_B/C_0$. SI Figure 20 illustrates the additional dependence of e on x — wherever $e(x)$ intersects the original solution, there is a fixed point. SI Figure 20B shows that even a linear $e(x)$ can intersect the original solution in three places for certain parameters, in this case for certain values of f. Any curve that intersects one of the solid lines in three places will result in bistability.

Changing the precise location of the transition from low to high spacer effectiveness does not change

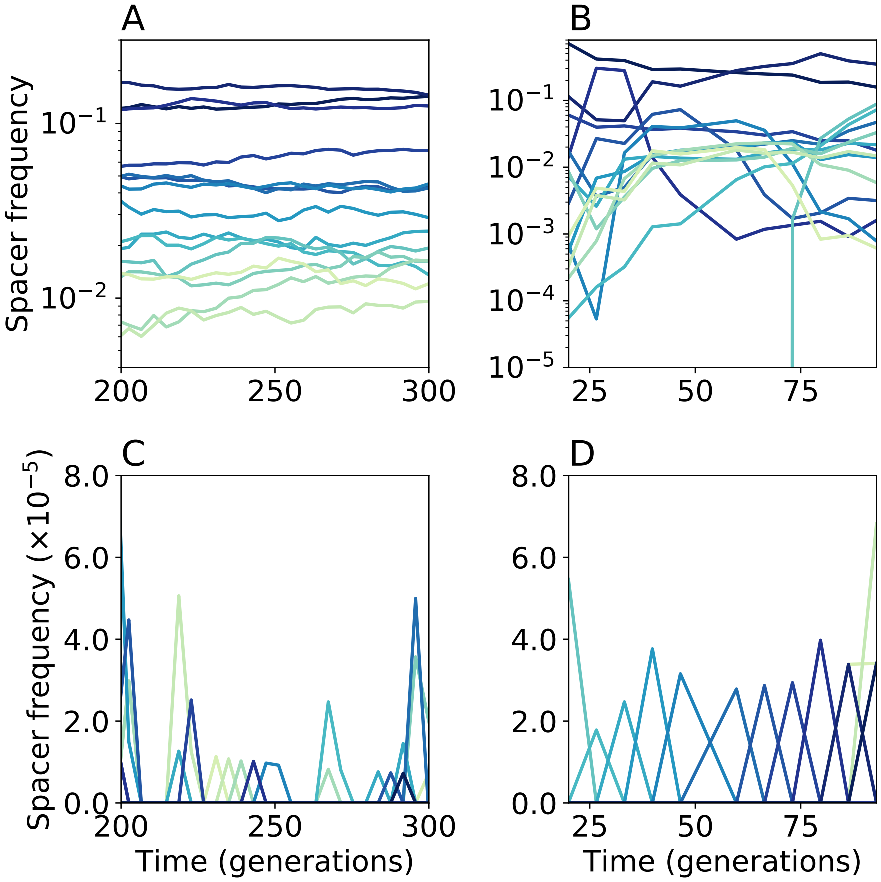

Figure 16: Spacer type trajectories vs. time for experimental data from [1] (B and D) and for data from our simulations (A and D). Colours indicate different spacer types. (A and B) show the largest 15 spacer types vs. time and (C and D) show the lowest 15 unique spacer type trajectories vs. time. Both large and small abundance spacers experience turnover in the experimental data, while in simulations the large abundance spacers are approximately stable once the system has reached population-level steady state.

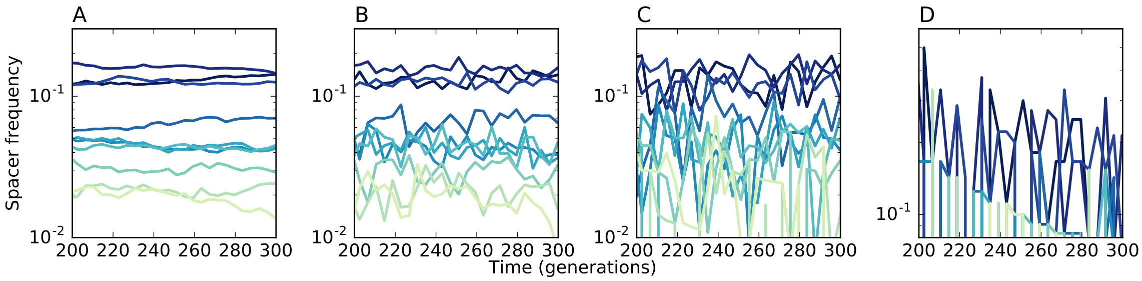

Figure 17: A comparison of the top 10 spacer types from the original simulation (A) with a randomly sampled subset of the simulated data (B, C, D). Spacers are sampled without replacement at every 50th simulation time point. Data is undersampled by a factor of 10^2 (B), 10^3 (C) and 10^4 (D).

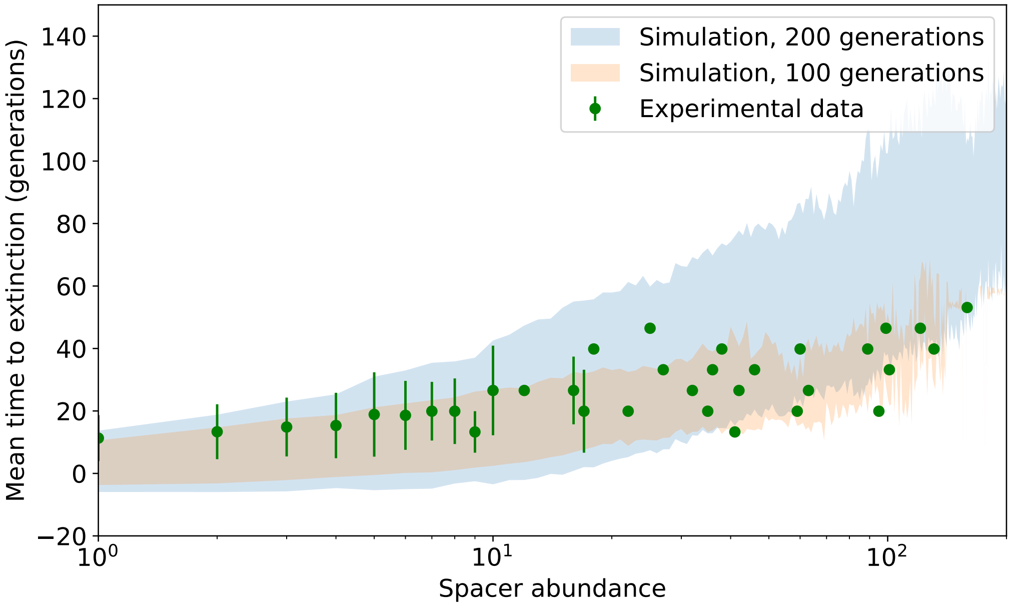

Figure 18: Standard deviation of mean time to extinction for a short and long simulated dataset with $\eta = 10^{-5}$ and $e = 0.387$ (shaded areas), and mean time to extinction for experimental data (green points). Errorbars for experimental data are standard deviation of mean time to extinction. Time in generations for the experimental data is time in days ×6.64, assuming exponential growth between daily 100-fold dilutions.

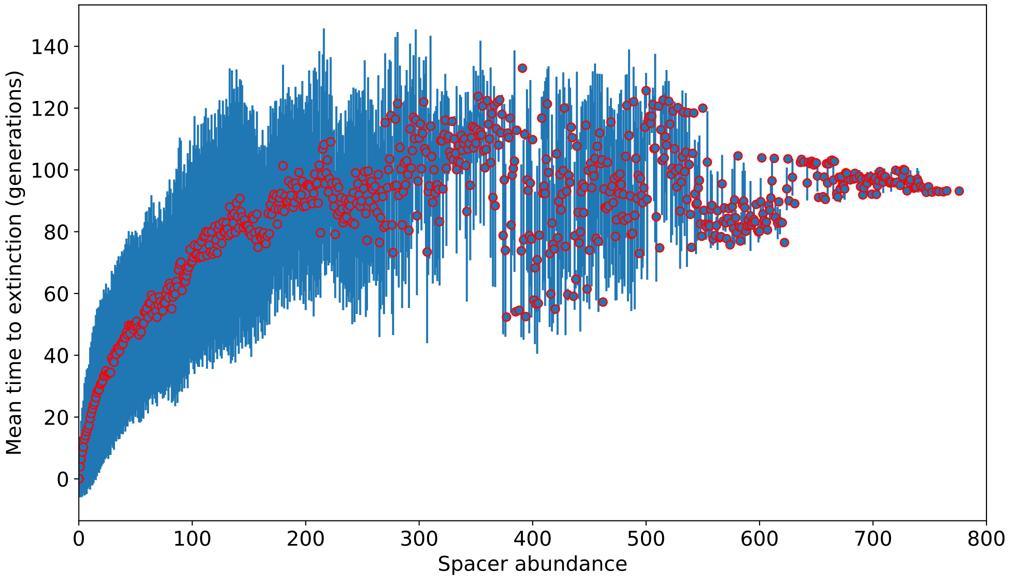

Figure 19: Mean time to extinction for simulated data with $\eta = 10^{-5}$ and $e = 0.387$ (red points). Errorbars (blue lines) are standard deviation of mean time to extinction.

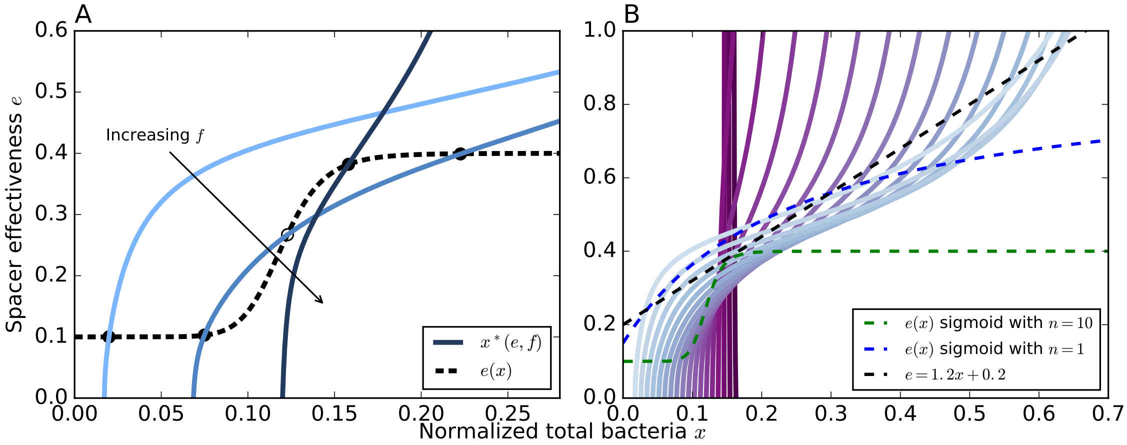

Figure 20: (A) The original dependence of bacterial population size at steady-state on spacer effectiveness e and normalized flow rate f is plotted for three values of f (solid blue lines). We model upregulation from quorum sensing by introducing a density-spacer effectiveness (dashed black line), $e(x) = e_{min} + (e_{max} - e_{min})\left(\frac{x^n}{x^n + x_0^n}\right)$, so that spacer effectiveness is no longer a constant parameter. Any intersection of the dashed line with a solid line is a fixed point; fixed points are indicated with solid circles (stable) and open circles (unstable). (B) Spacer effectiveness e vs. bacterial population size at steady-state for different values of f (solid lines). Line colour darkens as f increases. Three different choices of $e(x)$ are plotted (dashed lines), all of which intersect some of the solid curves in three places, indicating bistability.

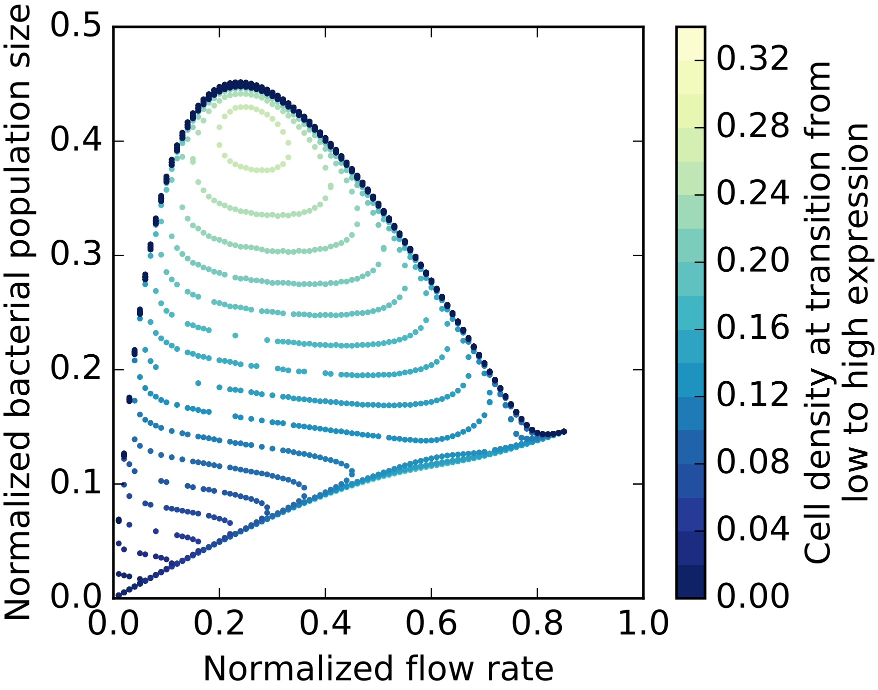

Figure 21: Fixed points (bacterial population size) as a function of flow rate f for different values of the transition point between low and high expression. As the transition point increases (lighter colours), the bistability changes from an 'S' shape to a circle and a line. This bifurcation happens at a transition point of approximately $x = 0.15$.

the existence of bistability, but it does cause an interesting bifurcation. SI Figures 21 and 22 show in two and three dimensions what happens to the fixed points as the transition point x_0 is scanned from 0 to 0.3. For a transition point at low cell density, the unstable fixed points are adjacent to the low expression stable fixed points at one end and the high expression stable fixed points at the other end, making hysteresis possible. However, as the transition point increases to higher cell density, the two ends meet and form a closed loop with just the high expression state. In this situation, bistability still exists, but the system can never jump from the low expression state to the high expression state without being placed there since there is one continuous low expression stable state across the entire range of f.

5.2 Bistability across system variables

Bistability affects all four dynamical variables in our model. SI Figure 23 shows each variable at steady state vs. flow rate f in a regime with bistability.

5.3 Adding regulation to acquisition, loss, and growth rate

We model CRISPR-Cas regulation by making spacer effectiveness density-dependent, but it is reasonable that up-regulation of CRISPR-Cas would affect other system parameters as well. In particular, spacer acquisition rates would likely increase since acquisition relies on the Cas protein machinery as does interference [16]. Additionally, spacer loss is thought to happen by homologous recombination and to occur in tandem with acquisition [17, 18].

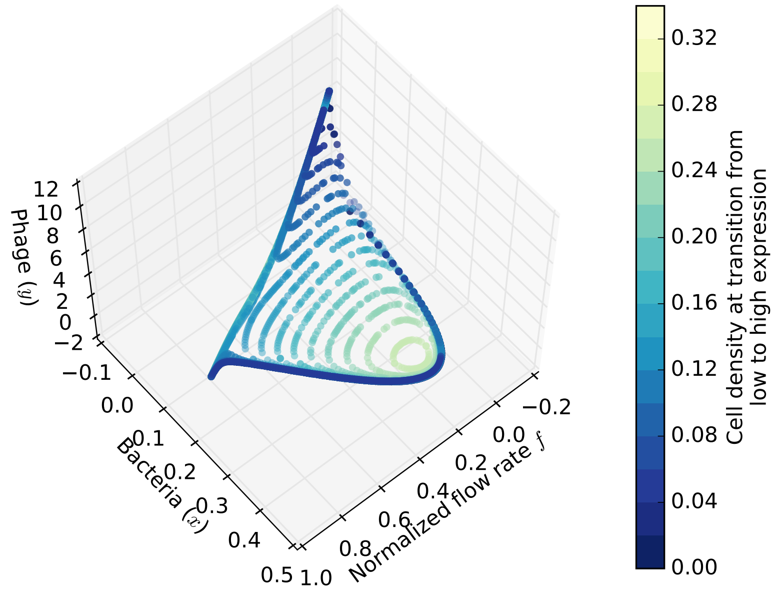

Figure 22: Fixed points (bacterial population size and phage population size) as a function of flow rate f for different values of the transition point between low and high expression.

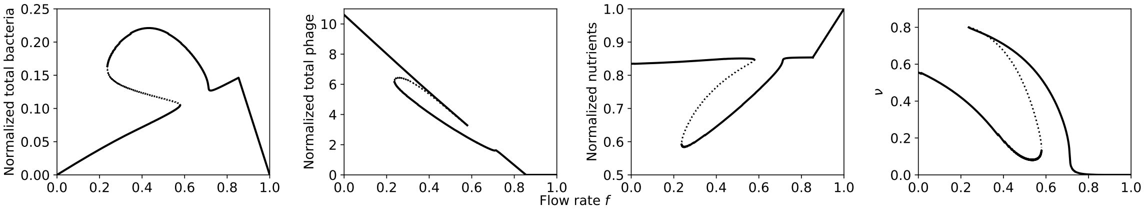

Figure 23: Bacteria x, phages y, nutrients z, and the fraction of bacteria with spacers ν as a function of f in a parameter regime showing bistability. The solid black lines indicates a stable fixed point and the dashed black line indicates an unstable fixed point.

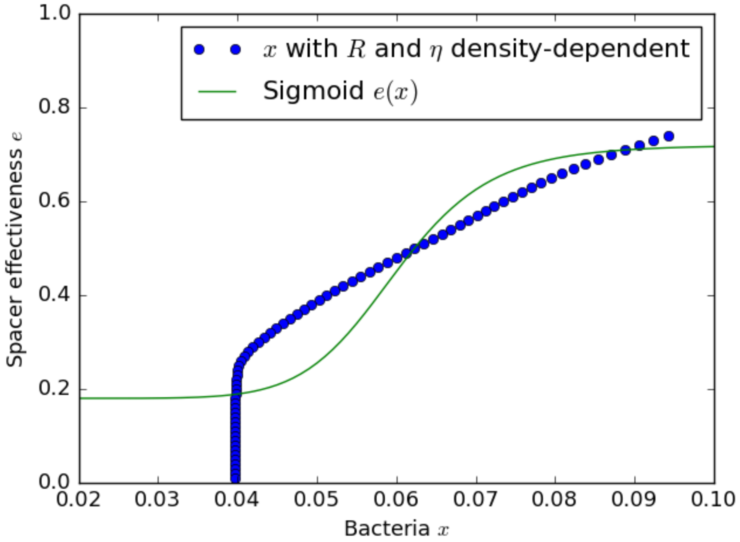

Figure 24: The dependence of bacterial population size x at steady-state on spacer effectiveness e when r and η are both sharp functions of density (blue dots). A monotonic function for spacer effectiveness as a function of x (green solid line) can still only intersect in at most three places, qualitatively giving the same bistability result.

We added a sharp sigmoidal density dependence to both spacer acquisition probability and spacer loss rate. SI Figure 24 shows the resulting steady-state bacterial population size as a function of spacer effectiveness. The result is still monotonically increasing, which means that a monotonic function for spacer effectiveness as a function of x can still only intersect in at most three places, qualitatively giving the same bistability result.

Measurements of the fitness cost of CRISPR in *Streptococcus thermophilus* identified Cas protein expression as having a fitness cost [19], making it reasonable that bacteria would down-regulate Cas expression in times when CRISPR is not needed. [19] measured a selective advantage of 0.11 for S. *thermophilus* with a *cas9* or *csn2* gene knockout in direct competition with wild type but did not observe a difference in maximum growth rate. This definition of selective advantage corresponds to the difference in average exponential growth rate per hour for each strain. We incorporated a Cas-expression-dependent decrease in bacterial growth rate in our model and investigated its effect on bistability. Here we model Cas expression as a theta function (discrete 'off' and 'on' states) with the switch occurring at $x_C = 0.06$ (arbitrarily chosen):

$$g(x) = \begin{cases} g_1 & x \leq x_C \\ g_0 & x > x_C \end{cases} \tag{48}$$

The growth rate gC_0 depends on the Cas expression state with g_1C_0 being the growth rate per minute without Cas expression and g_0C_0 being the growth rate with Cas expression, where $g_0 < g_1$. A selective advantage of 0.11 gives $g_0 = g_1 - 0.11/(60C_0)$.

SI Figures 25 and 26 show the resulting change in steady-state bacterial population size as a function of spacer effectiveness for two different growth rate dependences on expression. For even a 50 percent reduction in growth rate at high Cas expression, the resulting curves are not qualitatively altered, and as before, a monotonic curve for $e(x)$ can intersect in at most three places to give bistability.

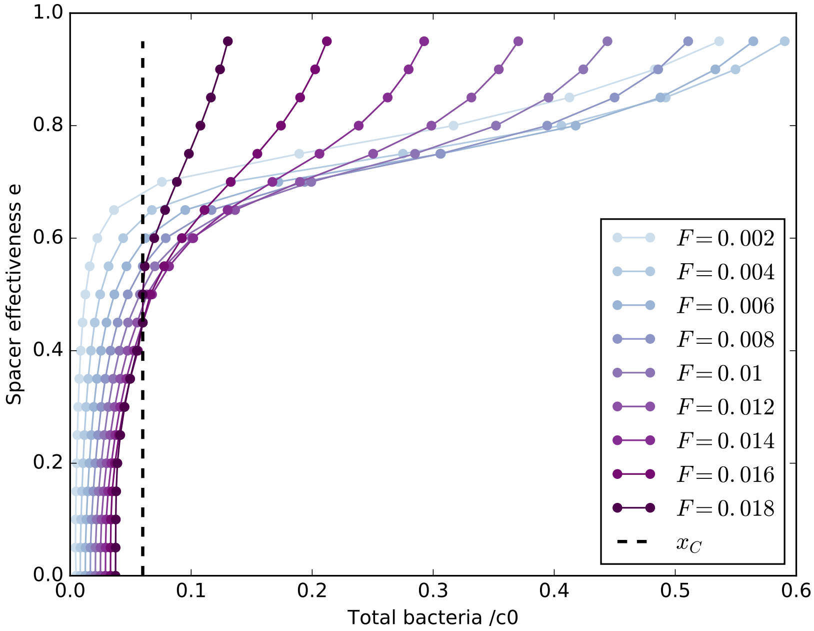

Figure 25: The dependence of bacterial population size x at steady-state on spacer effectiveness e when g is a sharp function of cell density x. The value of x at which regulation is turned on or off is indicated by the black dashed line. Lines are plotted for F instead of $f = F/(gC_0)$ because f depends on g. Plotted is bacterial population size at steady state where the growth disadvantage for Cas expression is $g_0 = g_1 - 0.11/(60C_0)$, calculated from the measured selection coefficient in [19].

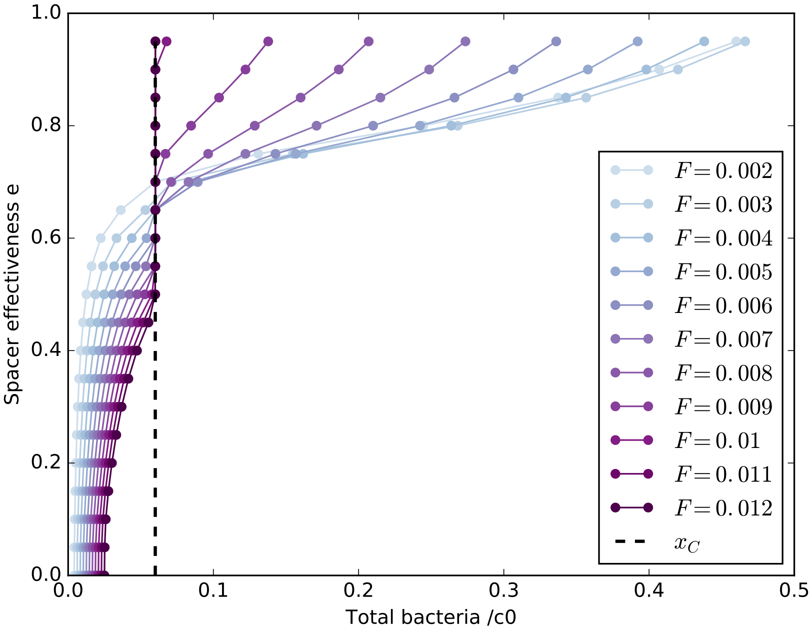

Figure 26: The dependence of bacterial population size x at steady-state on spacer effectiveness e when g is a sharp function of cell density x. The value of x at which regulation is turned on or off is indicated by the black dashed line. Plotted is bacterial population size at steady state where the growth disadvantage for Cas expression is $g_0 = 0.5g_1$.

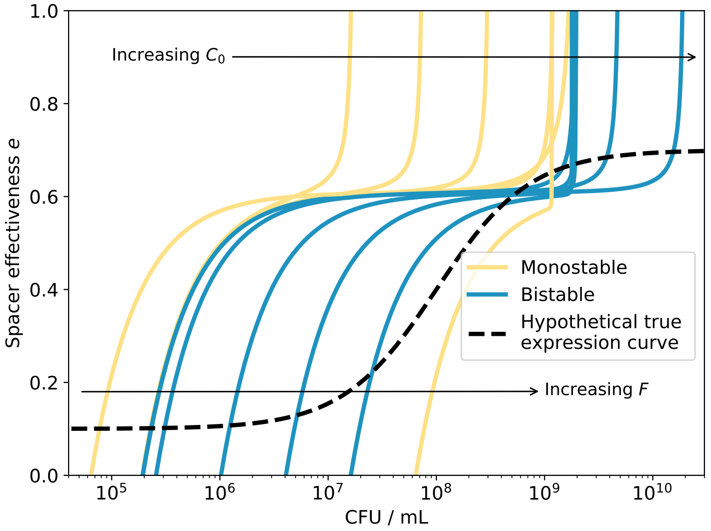

Figure 27: The dependence of bacterial population size x at steady-state on spacer effectiveness e in the model for different values of F and C_0 (solid lines). For a given measured dependence of *cas* expression on cell density (black dashed line, for example), F and C_0 can tune whether the system is monstable or bistable by changing the number of intersections between the two curves.

5.4 Experimentally measuring regulation

While we chose parameters that are reasonable for S. *thermophilus*, it is unlikely that our quantitative results match experimental conditions for different organisms. Our prediction is that in an appropriate parameter range, an experiment measuring bacterial population density as a function of flow rate may exhibit hysteresis as the flow rate is first increased and then decreased, allowing the bacteria-phage population to reach steady state after each change in flow rate. It is easy to imagine however that the transition determining high or low Cas expression may not automatically align with the cell densities in the chemostat. The first experimental step is to measure the true Cas expression as a function of cell density for *Pseudomonas*, as done in [20]. In their experiment, *cas3* expression increased by a factor of about 10 for a 10-fold increase in cell density (from $\approx 8 \times 10^7$ to $\approx 8 \times 10^8$ CFU/mL).

Next, the concentration of nutrients in the inflow medium C_0 can be used to tune the cell density to one at which CRISPR would naturally be highly expressed at a high flow rate. Then the flow rate F can independently tune the position along the bifurcation diagram in SI Figure 4. In this way an experimental population of *Pseudomonas* can be tweaked to qualitatively align with our model.

SI Figure 27 shows the steady-state bacterial concentration vs. spacer effectiveness in our model as C_0 and F are varied. Provided the true Cas expression is a sharp enough function of density and that the low expression state is below the plateau in effectiveness in SI Figure 27, it will be possible to choose C_0 and F such that the system is bistable. The position of the plateau in effectiveness at which the bacterial density changes sharply is controlled by Bp_V: as Bp_V increases, the plateau moves to higher effectiveness. B and p_V are properties specific to the phage and may change with the particular phage species used.

5.4.1 Significance of regulation in natural populations

In natural populations, multiple states may define different ecological niches as seen in structured populations from microbial mats [21] to the human microbiome [22]. Biofilms are an example of both dense and structured communities of bacteria and are found in many natural environments such as hot springs [21] and acid mine drainage [23] and in many clinically relevant environments such as medical implants, lungs of cystic fibrosis patients, and dental plaques [24]. Because of their protective polysaccharide coating, biofilms are often difficult to treat with antibiotics [24], and phage therapy has been proposed as a potential treatment for antibiotic-resistant bacterial colonies. Høyland-Kroghsbo *et al.* posited that upregulation of CRISPR-Cas could pose a challenge to potential phage therapies for biofilms [20]. If such a biofilm-bound population is in the bistable regime we find, there may be a way to prime the population in way that pushes it to the low CRISPR-Cas expression state to utilize phage therapy effectively. More broadly, in a resource-limited environment, for example, a bacterial population may do better to maintain a low density and avoid phage predation while repressing the expression of Cas proteins, but consequently may lose their CRISPR-Cas system entirely. These ecological constraints may shed light on why CRISPR-Cas is neither universal nor uncommon in the microbial world.

References

[1] David Paez-Espino, Wesley Morovic, Christine L Sun, Brian C Thomas, Ken-ichi Ueda, Buffy Stahl, Rodolphe Barrangou, and Jillian F Banfield. Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nature communications*, 4:1430, jan 2013.

[2] André Gilles, Emese Meglécz, Nicolas Pech, Stéphanie Ferreira, Thibaut Malausa, and Jean-françois Martin. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, 12(1):245, 2011.

[3] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Efficient step size selection for the tau-leaping simulation method. *J. Chem. Phys*, 124(4):044109, 2006.

[4] Sacha Lucchini. *Genetic Diversity of Streptococcus thermophilusPhages and Development of.* PhD thesis, Swiss Federal Institute of Technology Zurich, 1999.

[5] M Delbrück. Adsorption of bacteriophage under various physiological conditions of the host. *J. Gen. Physiol.*, 23(5):631–42, 1940.

[6] F. Vaningelgem, M. Zamfir, T. Adriany, and Luc De Vuyst. Fermentation conditions affecting the bacterial growth and exopolysaccharide production by Streptococcus thermophilus ST 111 in milk-based medium. *J. Appl. Microbiol.*, 97(6):1257–1273, 2004.

[7] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry.* Elsevier, third edition, 1981.

[8] Brian J. McGill, Rampal S. Etienne, John S Gray, David Alonso, Marti J Anderson, Habtamu Kassa Benecha, Maria Dornelas, Brian J. Enquist, Jessica L. Green, Fangliang He, Allen H. Hurlbert, Anne E. Magurran, Pablo A. Marquet, Brian A. Maurer, Annette Ostling, Candan U. Soykan, Karl I. Ugland, and Ethan P. White. Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10(10):995–1015, 2007.

[9] B. Dennis and G. P. Patil. The gamma distribution and weighted multimodal gamma distributions as models of population abundance. *Mathematical Biosciences*, 68(2):187–212, 1984.

[10] S. Engen and R. Lande. Population dynamic models generating the lognormal species abundance distribution. *J. Theor. Biol*, 132(2):169–183, 1996.

[11] O. H. Diserud and S. Engen. A general and dynamic species abundance model, embracing the lognormal and the gamma models. *The American Naturalist*, 155(4):497–511, 2000.

[12] Joshua B. Plotkin and Helene C. Muller-Landau. Sampling the species composition of a landscapre. *Ecology*, 83(12):3344–3356, 2002.

[13] R. A. Fisher, A. Steven Corbet, and C.B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecology*, 12(1):42–58, 1943.

[14] Ryan A Chisholm and Stephen W Pacala. Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proc Natl Acad Sci USA*, 107(36):15821–15825, 2010.

[15] Melissa B. Miller and Bonnie L. Bassler. Quorum Sensing in Bacteria. *Annual Review of Microbiology*, 55(1):165–199, 2001.

[16] Rodolphe Barrangou and Luciano A. Marraffini. CRISPR-cas systems: Prokaryotes upgrade to adaptive immunity. *Molecular Cell*, 54(2):234–244, apr 2014.

[17] Hélène Deveau, Rodolphe Barrangou, Josiane E. Garneau, Jessica Labonté, Christophe Fremaux, Patrick Boyaval, Dennis a. Romero, Philippe Horvath, and Sylvain Moineau. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.*, 190(4):1390–1400, feb 2008.

[18] Ariel D Weinberger, Christine L Sun, Mateusz M Pluciński, Vincent J Denef, Brian C Thomas, Philippe Horvath, Rodolphe Barrangou, Michael S Gilmore, Wayne M Getz, and Jillian F Banfield. Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comp. Biol.*, 8(4):e1002475, jan 2012.

[19] Sylvain Gandon, Pedro F Vale, Guillaume Lafforgue, Francois Gatchitch, Rozenn Gardan, and Sylvain Moineau. Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. *Proc Biol Sci*, 282(1812):20151270, 2015.

[20] Nina M Høyland-Kroghsbo, Jon Paczkowski, Sampriti Mukherjee, Jenny Broniewski, Edze Westra, Joseph Bondy-Denomy, and Bonnie L Bassler. Quorum sensing controls the Pseudomonas aeruginosa CRISPR-Cas adaptive immune system. *Proc Natl Acad Sci USA*, 114(1):201617415, 2016.

[21] David M Ward, Mary M Bateson, Michael J Ferris, M. Kuhl, Andrea Wieland, Alex Koeppel, and Frederick M Cohan. Cyanobacterial ecotypes in the microbial mat community of Mushroom Spring (Yellowstone National Park, Wyoming) as species-like units linking microbial community composition, structure and function. *Philos. Trans. R. Soc. B: Biol. Sci.*, 361(1475):1997–2008, 2006.

[22] Curtis et al. Huttenhower. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, jun 2012.

[23] Anders F Andersson and Jillian F Banfield. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science (New York, N.Y.)*, 320(5879):1047–50, may 2008.

[24] Michael T. Madigan and John M. Martinko. *Brock biology of microorganisms*. Pearson Prentice Hall, 11 edition, 2006.