The network structure and eco-evolutionary dynamics of CRISPR-induced immune diversification

Shai Pilosof[1,2], Sergio A. Alcala-Corona[2], Tong Wang[3,4], Ted Kim[4,5], Sergei Maslov[3,4], Rachel Whitaker[4,5,*], and Mercedes Pascual[2,6,**]

[1] *Department of Life Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel*

[2] *Department of Ecology and Evolution, University of Chicago, 1103 E 57 st, Chicago, 60637, USA*

[3] *Department of Bioenginnering, University of Illinois at Urbana-Champaign,Urbana, Illinois, USA*

[4] *Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign,Urbana, Illinois, USA*

[5] *Department of Microbiology, University of Illinois at Urbana-Champaign,Urbana, Illinois, USA*

[6] *Santa Fe Institute, Santa Fe, New Mexico, USA*

[*] *rwhitaker@life.illinois.edu*

[**] *pascualmm@uchicago.edu*

Abstract

As a heritable sequence-specific adaptive immune system, CRISPR-Cas is a powerful molecular mechanism shaping strain diversity in host-virus systems. Nevertheless, the structure and dynamics of host-virus interactions associated with diversification remain largely unexplored. We quantified the network structures of infection (who infects whom) and immunity (who is protected from whom) in a stochastic Lotka-Volterra model of host and viral populations including evolution. The coevolutionary dynamics exhibit an alternation between periods of virus-host diversification and host control. In periods of diversification, infection networks are partitioned into modules of hosts and viruses, reflecting the emergence of niches within which viruses can grow as the result of negative frequency-dependent selection. Acquisition of immunity closes available niches and builds a weighted-nested immunity network, representing redundant protection and causing a shift to a host-controlled regime. The nested structure enforces an orderly virus extinction which in turn increases the potential for an escape mutation in viruses and another transition to a virus-diversification regime. These dynamics and structures are not obtained under neutral scenarios lacking specific immunity. FInally, the immunity networks in three empirical systems also exhibit weighted nestedness, a pattern our theory shows is indicative of host control. Our findings emphasize the role that network structure plays in CRISPR-induced host-virus coevolution, providing one explanation for existing host/viral diversity in natural and empirical systems.

Introduction

The structure of complex ecological communities is clearly non-random. It is generally argued that interaction structure affects the stability of communities to perturbations (*1*–*3*) and arises from coevolutionary dynamics between interacting partners (*4*–*7*), yet the structure-dynamics nexus remains a long-standing central question in community and network ecology.

Host-parasite infection networks, in which interaction structure represents 'who infects whom', have been typically described as modular, nested, or both (*8*, *9*). Modularity concerns patterns of specificity, in which the network is partitioned into modules of hosts and parasites that interact densely with each other but sparsely with those outside the group. Nestedness concerns instead patterns of specialization, and describes a network structure in which specialist hosts are infected by subsets of parasites that in turn infect the more generalist hosts (*8*, *9*).

Common to all host-parasite network studies is a dominant focus on patterns of infection. In-

fection structure should critically depend however on the genetic basis of resistance of hosts to pathogens (*10*). The emergence of network structure from immune selection has been shown in pathogens of humans such as *Plasmodium falciparum* (*11*, *12*). In particular, strain theory developed for pathogens with multilocus encoding of antigens posits that negative frequency dependent selection, mediated by competition for hosts through cross-immunity, can structure pathogen populations into clusters of strains with limited overlap of antigenic repertoires (*11–15*). Parasites with antigens that are novel to the host immune system are at a competitive advantage, while those with common ones are at a disadvantage. So far, these studies have only analyzed genetic similarity from the parasite perspective because information on host immunity is typically absent. Hence, how immunity structures host-parasite interactions, for example into modular, or nested topologies, is unknown.

Given the ubiquity and importance of microbes and their infectious viruses for all ecosystems on earth, a critical application of host-parasite networks is to bacteria-phage interactions. An analysis of a large assemblage of bacteria-phage infection networks found that these are predominantly nested rather than modular (*16*) but a later study suggested that structure depends on phylogenetic scale: modularity at large phylogenetic scales of species interactions and nestedness among interacting strains of the same species (*17*). Several hypotheses have been put forward to explain the emergence of modularity and nestedness in bacteria-phage infection networks from immunity-related coevolutionary dynamics (*18*, *19*). For example, Fortuna et al. (*20*) have shown that networks of phage and bacteria evolve nestedness under arms race dynamics, but not under fluctuating selection. No study has addressed however host-virus infection or immunity networks in the context of the explicit underlying molecular mechanisms and the links from genotype to immune phenotype.

Clustered regularly interspaced short palindromic repeats (CRISPR) and its CRISPR-associated (Cas) proteins is a heritable adaptive immune system conferring sequence-specific protection against viruses, plasmids, and other mobile elements (*21*, *22*). As such, it provides a direct sequence-based link between host and pathogen genotypes and immune phenotypes. It also allows consideration of concomitant structures for both infection and immunity, and from both host and parasite perspectives. While the diversity of CRISPR alleles in natural, experimental and simulated populations has been explored (*23–27*), the structure of the emerging host-virus interaction networks, and how they are assembled and disassembled has not. Advancing this predictive framework for CRISPR-Cas in microbial populations is essential for effective experimental manipulation of microbial systems,

including applications to health and engineering (*28*, *29*), and can open new frontiers in studies of immune diversity and, more generally, tighten the link between microbiology and evolutionary biology(*30*).

Results

The CRISPR system functions as an adaptive immune system by incorporating DNA segments called 'protospacers' of infecting viruses into host genomes as 'spacers' that constitute sequence-specific immunity and memory (*22*). From this perspective, strain diversification of both hosts and pathogens emerges from their frequency-dependent co-evolutionary dynamics, and involves the large combinatorics of spacers and protospacers. Previous work has shown the possibility of diverse populations, containing many strains, with distributed immunity of the host, whose dynamics can exhibit two alternating major regimes, dominance of microbial hosts and escape and diversification of the viruses, respectively (*31*). We extended the model to allow for larger host and virus richness, and to track the co-evolutionary history of both hosts and viruses through time (Methods). The timing of the switching between the two regimes is defined and identified here based on the dynamics of virus abundance (Methods). In the 'virus-diversification regime' (VDR), virus strains proliferate and diversify while in the 'host-controlled regime' (HCR) host strains are able to constrain virus diversification and lead to their extinction (Fig. 1). During the former, viruses and hosts coexist with fluctuating abundances, whereas during the latter, hosts reach carrying capacity and viruses exhibit declining abundances and richness (Figs. S3, S4, S5).

Variation in viral or host abundance alone cannot drive the transitions between these two regimes, as these do not exist in the corresponding Lotka-Volterra dynamics under neutral conditions of hosts not acquiring specific immunity (SI D,E). An alternative explanation is that the structure of strain diversity, emerging from the eco-evolutionary dynamics via specific immunity, explains the transitions between these dynamical regimes. To investigate the structure of diversity in this complex system, we consider two complementary bipartite networks, the 'immunity' and 'infection' graphs, through time. In these networks a node represents a strain of a virus (unique combination of protospacers) or a host (unique combination of spacers), and edges represent a given type of interaction, either infection or protection from infection (Fig. S6).

At the start of the simulations and during VDRs, the infection network is built over time (by addition of host and virus strains) developing a modular structure in which infections are

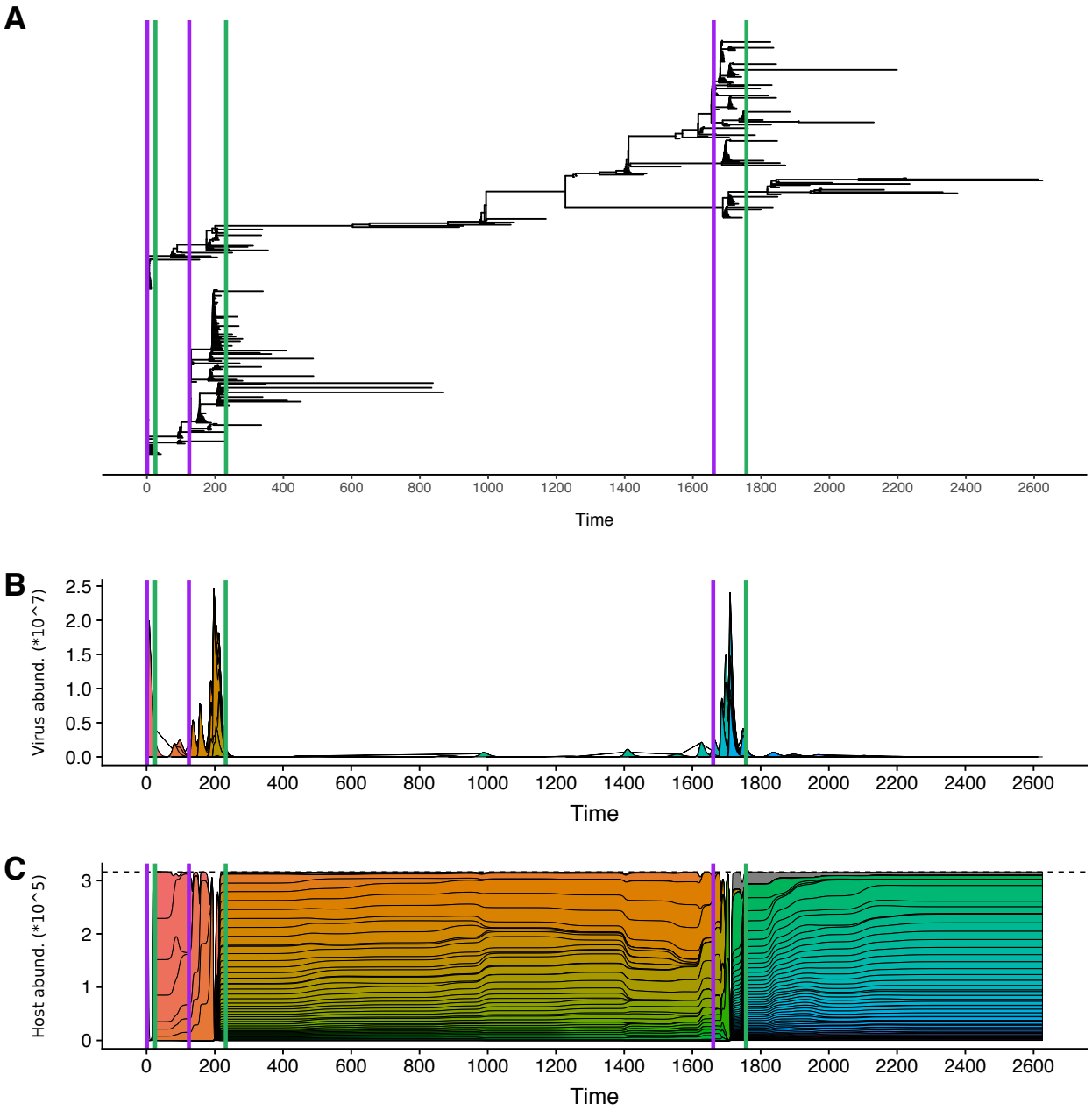


Fig. 1 An example of typical dynamics with regime shifts. Virus-diversification regimes (VDRs) start at purple vertical lines and end in the green vertical lines, when the host-controlled regimes (HCRs) begin. **(A)** Viral phylogenetic tree shows virus diversification during VDRs and virus extinctions during the HCRs. Intermittent virus diversification in the later part of HCRs eventually leads to an escape and the initiation of a new VDR. **(B)** and **(C)** are the abundance profiles of viruses and hosts, respectively. The 100 most abundant strains are colored, the rest are aggregated and shown in gray. Note that virus growth does not imply a decrease in host richness while host dominance does decrease virus abundance and richness sharply. Virus diversification implies a temporary escape from host control and their rise in abundance, which in turn increases encounters with hosts. The resulting rise in per-capita infection rates of hosts leads to their concomitant diversification because new host strains are generated by acquiring at least one new protospacer. Therefore, despite an initial decline in the abundance of hosts, their richness typically rebuilds (Fig. S3).

concentrated within modules of viruses and hosts, with more edges within than between these groups (Fig. 2A, Fig. S7).These modules reflect different niches (hosts as resources) for virus growth. As proposed by the strain theory for multilocus antigen pathogens (*13*, *32*, *33*), this niche structure arises in part from frequency-dependent competition among viruses for hosts. Specifically, viruses with rare escape mutations in protospacers to which hosts have not yet gained immunity exhibit a competitive advantage over those with abundant protospacers. Viruses with common protospacers are in turn at a disadvantage. The resulting frequency-dependent selection drives viruses to have little overlap in the hosts they can infect, and in so doing creates a modular structure which enables their coexistence and therefore their increasing diversity. The creation of niches is linked to the structured genetic diversity of the hosts, which is apparent in the also modular host-spacer network in which modules delineate groups of hosts that share immunity via the same spacers (Fig. S8).

The importance of immune selection in the formation of these niches can be demonstrated by asking whether clonal expansion alone could account for the modularity of the infection network. This is not the case, as shown by randomizations of structure that reorder strains according to phylogenetic relationships, in which the modules are lost (Fig. S9). Moreover, in the absence of frequency-dependent selection and associated host memory, diversification of the viruses and coexistence of different strains is not observed. This is demonstrated by a neutral model in which all the processes of the full system are retained except for the specific memory of the host (SI E).

In contrast to strain theory (*12*, *15*), here the persistence of the modular structure in the infection network and the coexistence of a diverse community of viral strains is only transitory. Diversification of viruses, enabled by the modular structure of the infection network, increasingly diversifies the host population (Fig. S5). In other words, escape from host control via mutation of protospacers allows higher abundances of particular virus strains, which therefore also experience higher encounter rates with hosts in general, leading to the acquisition of new spacers (through either the failure of their previous immunity or infection by an escape mutant) (see details on the model in Methods). Such red-queen co-evolutionary dynamics progressively adds spacers to hosts during the VDR, building the immunity network (Fig. 2B). In this network, edges indicate at least one spacer-protospacer match, and the weight of the edge encodes the number of different matches, protecting a given host from a given virus (Fig. S6C). By definition, a lack of an edge indicates infection, and an edge value larger than one indicates redundancy in immunity.

We find that this redundancy has a characteristic quantitative nested structure. We can order the network such that hosts with immunity to most viruses also have more matches to those viruses,

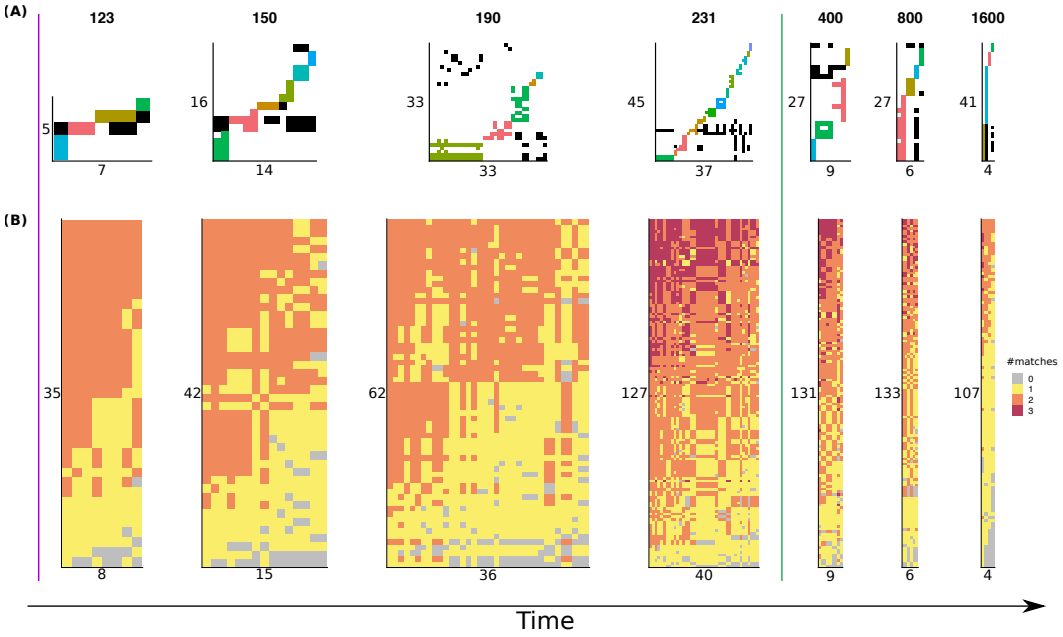


Fig. 2 Snapshots of network structure during the two regimes. Networks are defined as in Fig. S6. Each network is a snapshot of the population, with time steps depicted above the networks corresponding to those marked in Fig. 3. Numbers in the x and y axes indicate the number of viruses and hosts in the network, respectively. The purple and green lines depict the initiation of a virus-dominated and host-controlled regimes (VDR and HCR), respectively. **(A)** Modularity in the infection network. Colored interactions fall inside modules of virus strains that infect similar hosts (each module has a different color). Black interactions are those that fall outside all modules. The size of the network and the number of modules increase during the VDR (between purple and green vertical lines) and decline during the HCR (right of the purple vertical line). **(B)** Nestedness in immunity networks. Colors of interactions depict the number of matches between viruses and hosts. The network has a weighted-nested structure, which enforces an orderly viral extinction. Nestedness builds up during the VDR and declines during the HCR, as viruses go extinct. The extinction is orderly, with viruses to which many hosts has immunity via many spacers (those at the left), going extinct first.

and subsequent hosts (from top to bottom) are immune to subsets of viruses, via fewer matches (Fig. S6G, Fig. 2B). Similarly, each virus strain (from left to right) can infect a progressively smaller subset than the virus following it, also via fewer matches. The weighted nested structure is assembled by a complex interplay of temporal changes in abundances, associated encounter rates and selection pressures, in which host and viral age play an important role. How the matrix structure is woven from bottom to top, and left to right, with the respective addition of new hosts and viruses, is reflected in the relationship between the order of the rows and columns and strain age (Fig. S10).

The nested pattern in the redundancy of immunity, built during a VDR, enforces order and predictability in virus extinctions during the subsequent HCR. Viruses for which most hosts have

acquired immunity via multiple matches will tend to go extinct earlier than those with fewer matches (Fig. 2B, Fig. S11). Somewhat paradoxically, this orderly extinction will facilitate in turn a new viral escape and the initiation of another virus expansion cycle, as it reduces nestedness by preferentially removing viruses with a high number of matches. Specifically, the extinction process reduces the competition for hosts between viruses to which hosts have either no match (0-match) or a single match (1-match) (Fig. 3B) and therefore also increases their relative abundance in the populations. An increase in the proportion of 1-matches raises the effective mutation rate of the remaining viral population (SI C). Because viral escapes are associated with a particular tripartite structure in which hosts are immune to viruses via a single match (Fig. S12) , the increase in the frequency of viral strains with potential to escape is crucial for the ability to initiate a new VDR.

Epidemiologically, these two processes, viral growth rate and potential escape, can be captured via two modified measures of the basic reproductive number (*34*) (see SI B,C). The first, $R_j^0(t)$, is the number of offspring produced at time t by the virus strain j from infecting all hosts with no protection to it (0-matches). The second, $R_j^1(t)$, is the number of offspring that a virus strain j would produce by escaping protection from hosts via a single mutation (1-match) at time t. These two components can be added to obtain the 'potential reproductive number' of a virus strain j, $R_j^{pot}(t) = R_j^0(t) + R_j^1(t)$, which quantifies the contribution of a viral strain and its potential progeny to the population growth, conditional on escape.

During HCRs, the contribution of the second component R_j^1 can raise R_j^{pot} above the critical value of $R_j^{pot} = 1$ (Fig. 3C), indicating a viable escape. Importantly, as a higher proportion of the viral abundance becomes concentrated in strains with the highest R_j^{pot}, escape mutations must occur on those viruses that would eventually initiate a new VDR (Fig. 3D). Towards the end of an HCR, small viral outbreaks are observed of increasing frequency anticipating this critical transition (Fig. 1B). The small epidemics locally generate some virus diversification which counterbalances extinctions. These opposite effects mean that towards the end of the HCR, nestedness inverts its decline and starts growing again. Extinctions and births of virus strains change both the size of the immunity network and the distribution of protection weights among the links. These two effects are reflected in the nestedness index (*35–37*).

To ensure the generality of our results given the stochastic nature of the dynamics, we repeated our analysis for 100 simulations. Across these simulations, we find the same characteristic differences between the regimes in dynamics and network structure as in our main example. In particular, VDRs are shorter than HCRs, have higher values of weighted nestedness, have more

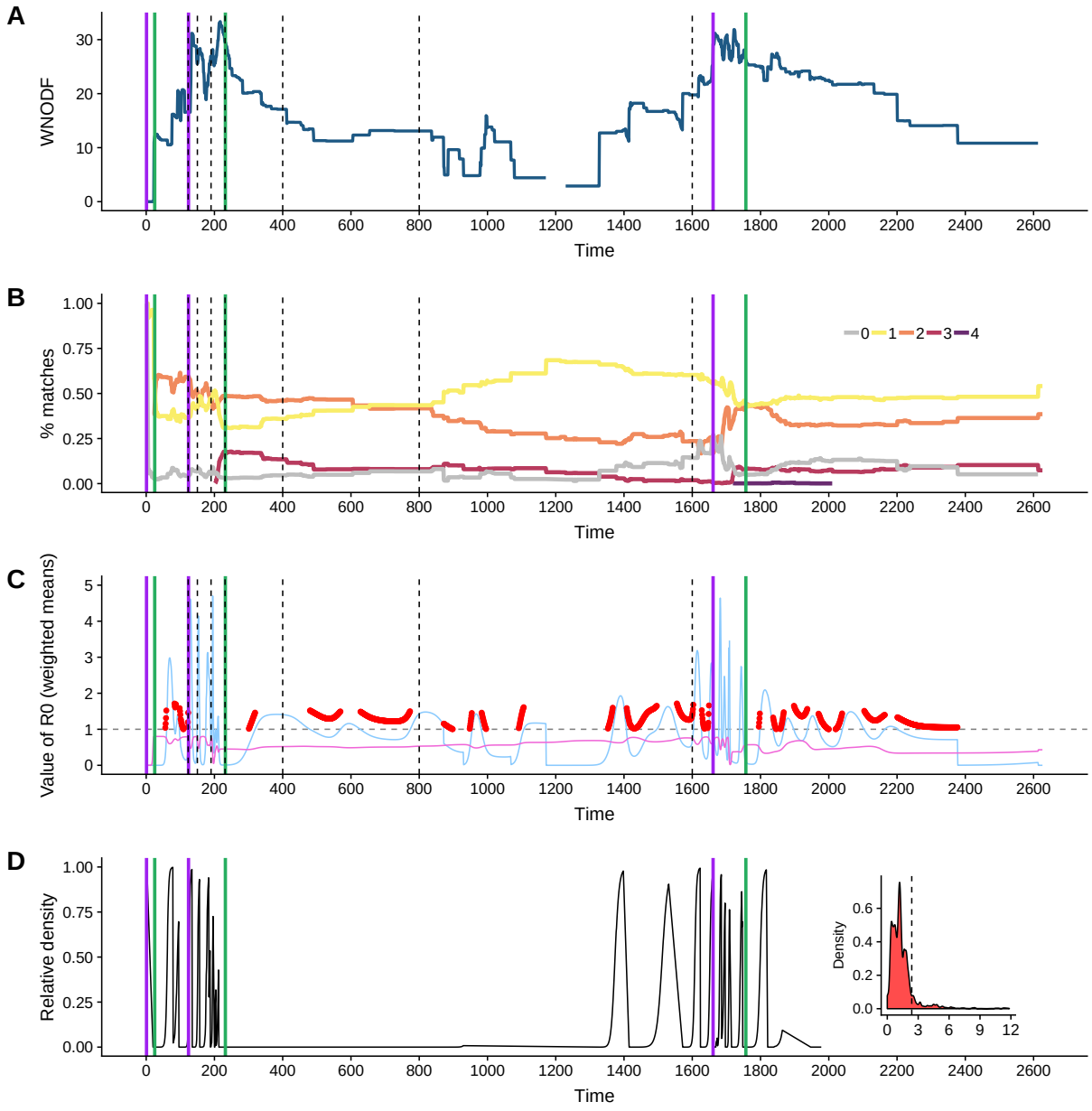


Fig. 3 Processes leading to regime shifts. **(A)** Weighted nestedness increases during VDRs and decreases during HCRs. **(B)** The disassembly of the network during HCRs increases the proportion of 0- and 1-matches, particularly towards the end of these periods. **(C)** During HCRs the average reproductive number R_j^0 (light blue line) is approximately 1 (horizontal dashed line). Although the mean R_j^1 is always < 1 (pink), R_{pot} can be > 1. During these times, depicted with red points, a mutation would allow the viral population to grow. **(D)** Even when $R_{pot} > 1$ in the HCR, an escape is likely only when the mutation hits an abundant virus. The plot shows the viral abundance corresponding to high values of R_{pot}. Specifically, this abundance is obtained by thresholding the distribution of R_{pot} values across the simulation (shown in the inset) at its 90% quantile (dashed vertical line), and summing the abundance of all viruses whose reproductive number exceeds this threshold. Approaching the end of the HCR, high values of R_j^{pot} (i.e., above the 90% quantile) are concentrated in virus strains that are highly abundant due to the extinction of most other viruses. Because mutations are more likely to occur in highly abundant virus strains, it becomes likely that a mutation hits a virus with a high potential for virus growth, and therefore, escape. Vertical dashed lines indicate the time points for the network snapshots in Fig. 2.

modules in the infection networks, and have higher values of R_j^0 (Fig. S13). Together, these results indicate a general pattern of viral diversification via creation of host niches, the corresponding buildup of immunity in VDRs, and the breakdown of weighted nestedness that eventually leads to escape during HCRs.

We also examined the effect of a five-fold acceleration in the speed of overall diversification of viruses, still within the general order of magnitude of plausible mutation rates. These "accelerated" simulations, which match the parameters used in (*31*) are also characterized by alternating regimes. As expected, they exhibit higher richness of hosts and viruses, and episodes of virus diversification can be longer than for lower mutation rates (Fig. S14). Their infection networks are modular and the immunity networks are nested (Fig. S15). Although viral richness still declines sharply at the beginning of each HCR, the earlier and faster generation of new strains via small outbreaks maintains a higher richness than before throughout this regime. Thus, nestedness appears more consistent through time. Higher mutation rates also make more apparent a general feature of the dynamics (already present at lower values) that requires further investigation in the future; namely, trends of diversification superimposed on the alternating regimes. Specifically, subsequent HCRs (VDRs) exhibit a higher number of host (viral) strains, culminating in long episodes of fluctuating richness for both players. This temporal but long trend appears to be transient, in the sense that the end of one of the high-richness episodes re-initiates the alternating regimes at a low diversity of both hosts and viruses (not shown).

Finally, we analyzed three unique empirical data sets. We have previously established profiles of diversity in naturally occurring microbial populations (*27*, *38*, *39*) but have not tested their structure in relation to immunity and infection networks. There are very few examples of virus and host populations resolved at an individual strain level and connected in space and time. We use empirical data from three data sets in which both CRISPR alleles and virus strains have been carefully and manually assembled (*27*, *39*). These represent lytic, chronic and temperate virus lifestyles and two different microorganisms from the two domains of life where CRISPR occurs (archaea and bacteria). The first two data sets compare genomes resolved to individual strains of the thermoacidophilic crenarchaeon *Sulfolobus islandicus* sampled at a single time point from two different locations. The third data set is from the gamma proteobacteria *Pseudomonas aeruginosa* (*39*), isolated from the sputum of CF patients in a high resolution longitudinal dataset collected by Marvig et al. (*40*), with virus genomes obtained from the mu-like viruses from the genomic database because these are the most highly targeted temperate phage genomes in P.

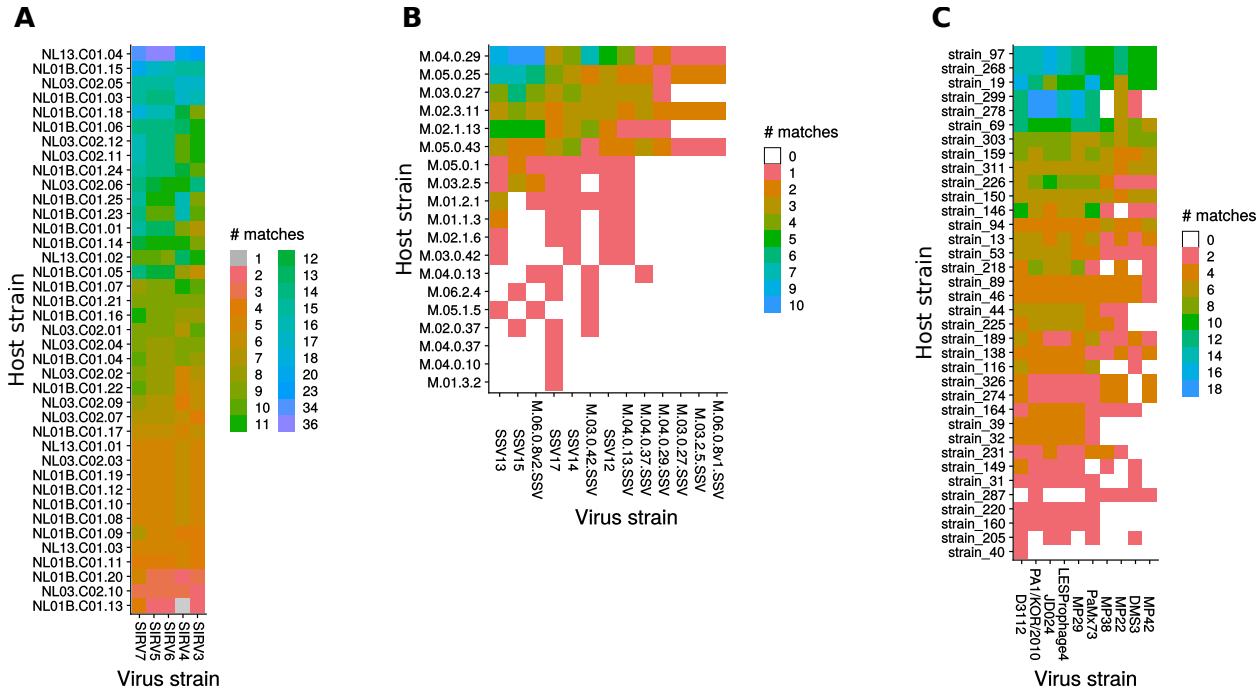


Fig. 4 Weighted nestedness of empirical immunity networks. The matrices depict the number of shared spacers and protospacers between hosts (rows) and viruses (columns). Each network is ordered by column and row sums and is nested in the quantity of matrix cells. The networks come from 3 different systems: **(A)** *Sulfolobus islandicus* hosts from a single location in Yellowstone National Park compared to contemporary lytic SIRV viruses isolated from Yellowstone National Park. **(B)** *S. islandicus* hosts compared to contemporary chronic SSV viruses from the Mutnovsky Volcano in Russia, 2010. (**C**) *Pseudomonas aeruginosa* hosts from Copenhagen compared to temperate mu-like viruses. To test the hypothesis that for a given distribution of matches in the population of host strains, the observed network is organized in a non-random, weighted-nested pattern, we shuffled networks by randomly distributing the interactions.

aeruginosa. From each data set, the empirical immunity matrices were constructed by comparing CRISPR arrays from each individual host strain to virus genomes. Because we did not have a time series, we assessed the statistical significance of nestedness for each empirical network by comparing the observed value of its nestedness index (for two different indices, Methods) to a distribution obtained from 10,000 shuffled networks. We find that the probability that a shuffled network will be more nested than the observed one is effectively nil (i.e., p-value$<$ 0.0001) in all three empirical networks (Fig. 4, Fig. S16). In light of our theoretical results, these findings suggest that in these three systems hosts control populations of viruses via distributed immunity that is redundant in the number of matches. Moreover, we find that, as in our theoretical results, the empirical host-spacer networks are also modular (Fig. S16), suggesting that the mechanism by which immunity is obtained is similar to the one we describe here.

Discussion

We have shown that the eco-evolutionary dynamics of this host-pathogen system influences, and is influenced by, the network structure of strain diversity. In particular, modularity of the infection network and weighted nestedness of the immunity network interact to influence the transient nature of alternating dynamical regimes. By promoting pathogen diversification, the former builds the latter. In turn, weighted nestedness contains the seeds of its own unraveling by implying a certain order in viral extinction, and in so doing, creating the conditions for a new viral escape.

Modularity has been described for host-parasite networks and attributed to a variety of mechanisms such as phylogeny, specificity and co-evolution (*9*, *41*, *42*). It was also detected in a host-virus system, where it was suggested to arise via negative frequency-dependent selection (*17*). Its emergence from immune selection has not been shown before. It is consistent with existing strain theory developed mostly for human infections and pathogens with multilocus encoding of antigens, when evolution and stochasticity are considered explicitly (*11*, *12*, *15*, *43*). Immune selection acts as a form of balancing selection creating groups of pathogens with limiting overlap in antigenic space. In the absence of data on immunity of hosts, this pattern has been described as modules, or clusters, in networks that depict genetic similarity between the pathogens themselves (*11*, *12*). This outcome is conceptually analogous to groups of species or microbes coexisting under frequency-dependent competition for resources (*44*). We are able to consider here the immune genotype of pathogens and hosts simultaneously, which allows us to move beyond pathogen genetic similarity to consider actual infection patterns. One major difference with previous strain theory is that modularity and associated coexistence in our model are only transient and not accompanied by stable or meta-stable population dynamics. Future work should examine whether this difference and the associated shifts in dynamical regimes arise from the heritable nature of CRISPR-based immunity, which by definition cannot produce fully susceptible offspring, needed to sustain transmission.

Weighted nestedness of protection networks has not been reported before as the focus of host-virus network studies has been on infection patterns (*18*). It remains an open question whether it can arise under non-CRISPR immunity or non-heritable immunity. For CRISPR-induced immunity, this structure reflects the co-evolutionary diversification of hosts in response to that of viruses, and the resulting redundancy of immune memory, which allows hosts to keep viruses in check despite their occasional escape. Thus, observation of weighted nestedness in nature would be

an indication of such control. Variation of this network property over time contains valuable information on the relative pace of viral diversification and host acquisition of immunity. Observing the temporal course of nestedness in both dedicated experiments and in natural environments should shed light on its role for controlling virus populations in relation to host and virus diversification. Quantitative extensions should also investigate appropriate ways to normalize weighted nestedness (sensu (*37*)), so that effects of its change can be isolated for network size and distributed redundancy of protection.

Transience of alternating dynamical regimes in the CRISPR system is not an impediment to the formation of rich structures by co-evolution; it is here its natural consequence. It should be recognized, however, that the alternation of dynamical regimes occurs in the model in a defined region of parameter space that allows for high diversification, and is absent in many earlier simulations (e.g. (*31*)). Identifying the conditions in terms of critical parameter combinations that give rise to host-control periods is of practical importance. Future theory should more broadly investigate the dynamics-structure nexus over parameter space, in particular for quantities that set the pace of co-evolution. Protospacer mutation rates and spacer acquisition probabilities have already been recognized as key to the outcome of viral extinction (*25*, *31*, *45*). Protospacer number should also be critical as the mutation rate per protospacer ultimately sets the speed of diversification and co-evolution. Although higher numbers should be examined for more realistic complexity, we expect our results to hold and provide a valuable point of reference. The nature of transitions between dynamical regimes is another open area for further theory. The observation of increasing frequency and size of small viral outbreaks at the end of host-control periods, suggests the critical nature of these transitions and their possible prediction.

The demonstration of significant weighted nestedness in the empirical networks suggests the applicability of our modeling results to predict these kinds of transitions. The next step is to make this connection. Knowledge on the stability of diverse host and viral populations can be applied to control of microbes in food and industrial sciences, infectious disease emergence and treatment with phage therapy, microbiome dynamics, agriculture, and environmental engineering.

Methods

The model

The model implements the formulation by Childs et al. (*23*), which combines three main components: ecological population dynamics, stochastic co-evolution generating diversity, and molecular identity of hosts and viruses defining CRISPR immunity. Diversification events (spacer acquisition by the host and protospacer mutation of viruses) are modeled stochastically, whereas population dynamics of different sub-populations (i.e., strains) of hosts and viruses are represented deterministically with Lotka-Volterra differential equations (eqns. 1,2).

$$\frac{dN_i}{dt} = rN_i\left(1 - \frac{\sum_i N_i}{K}\right) - \left[(1-q)\sum_j (1-M_{ij})V_j + p\sum_j M_{ij}V_j\right]\phi N_i \tag{1}$$

$$\frac{dV_j}{dt} = \beta\phi\left[(1-q)\sum_i (1-M_{ij})N_i + p\sum_i M_{ij}N_i\right]V_j - \left(\phi\sum_i N_i + m\right)V_j \tag{2}$$

Each strain of host i and virus j is defined by a *unique* genomic state of their spacer and protospacer sets S_i and G_j, respectively. In the ecological population dynamics, each host strain i has abundance N_i (the carrying capacity of all strains is K) and reproduces at a per-capita rate of r. Each viral strain j has abundance V_j, which increases due to infection and lysis of hosts and decays at a density-independent rate m. Extinction of any host or viral strain occurs when these fall below a critical threshold ρ_c. Viruses infect at a constant 'adsorption rate' ϕ either hosts that do not have protection or those whose protection fails (see below).

Host immunity to a virus is defined in the molecular component of the model and is based on genomic sequence matches between the spacer and protospacer sets. Specifically, CRISPR immunity is defined using the function $M_{ij} = M(S_i, G_j)$, which equals 1 if there is at least 1 match between the sets: $|S_i \cap G_j| \geq 1$, or 0 otherwise. The CRISPR immune mechanism is not perfect and can fail. When $M_{ij} = 1$, there is a probability p that the host strain is lysed and correspondingly, $1-p$ that it survives and the virus is eliminated. On the other hand, when $M_{ij} = 0$, there is a probability q that the virus strain is eliminated, resulting in the acquisition of a protospacer by the host, and $1-q$ that it is lysed by the virus. Both p and q are small (p, $q \ll 1$).

The above Lotka-Volterra dynamics are implemented in between any two stochastic events concerning evolution. Specifically, errors in viral replication can result after successful infection of a host, leading to the replacement of a random protospacer with mutation rate μ (per protospacer

per viral replication). During an unsuccessful infection attempt by a virus (regardless of host immunity), there is also a probability q of acquiring a new spacer by incorporating a protospacer and integrating it into the host's CRISPR system at its leading end. The maximum number of spacers per hosts strain is constant. If the maximum number of acquired spacers is reached, the addition of a spacer to the leading end is accompanied by the deletion of a spacer at the trailing end. In our simulations, the length of the spacer cassette is set to a sufficiently large value to avoid loss of acquired immune memory.

We numerically implemented the model in *C++* to increase computational efficiency. This enables consideration of a larger number of spacers/protospacers, and hence host and virus richness, for longer simulation times, than in the original MATLAB code (*23*)). Our implementation combines: (1) a Gillespie algorithm to determine the time between two stochastic events and to randomly select a virus/host strain to mutate a protospacer/acquire a new spacer, and (2) a numerical ODE solver using Euler's method. The code includes the following features to facilitate subsequent network (and other) analyses: (1) all data related to virus or host strains (e.g., identity of protospacers/spacers, abundance values) at each time are written to files during simulation; (2) the parent ID of each newly generated virus or host strain is tracked to generate phylogenetic trees; (3) a checkpoint implementation for running longer simulations. Details on the implementation are in Supplementary Information.

We used the parameters summarized in Table S1 based on (*23*, *31*). Host growth rate corresponds to a doubling time of an hour within the range of observed values for *P. aeruginosa* (20 mins) and *Sulfolobus* (8 hrs). The adsorption rate also falls within typical range values of 10^{-8} to 10^{-9} ml min^{-1}. The order of magnitude of our mutation rate is consistent with standard mutation rates for DNA based organisms, if we assume that mutation at a single site enables protospacer escape from a spacer match. The burst size of 50 is on the low side with values of 200 observed in *P. aeruginosa*. This is motivated by a lower protospacer number than currently documented in nature. Specifically, the speed of virus escape ultimately depends on the effective rate of mutation per protospacer, which increases linearly with the product of burst size and the mutation rate parameter μ and decreases inversely with the number of protospacers. Therefore, our parameter set should generate similar speeds of protospacer evolution than for higher burst sizes and also higher numbers of protospacers (e.g. $\beta = 200$ and $N_p \times 4$). Our numerical implementation allows consideration of a higher number of protospacers than before; for simplicity, in our study we used a number of protospacers that is still below those typically observed but have compensated with a

lower value of offspring produced (burst size; see SI.)

Definition of regimes

Our general approach to define regimes was to classify each point in the virus abundance time series to either a HCR or a VDR. We did so by detecting changes in relative virus abundance, defined as the total virus abundance at any given time t, $V_T(t)$, divided by the maximum abundance in all the time series, V_T: $A(t) = V_T(t)/max(V_T)$. Using relative abundance allows for comparisons and analysis across multiple simulations. Using the non-relative abundance does not change the regime definition (Fig. S17). A HCR is a sequence of points in which the total virus relative abundance changes very little, which can be captured in the second derivative of A. The first derivative is $A' = A(t_x) - A(t_{x-1})$ (we calculated in consecutive time steps so the denominator of the derivative is 1). The second derivative is $A'' = A'(t_x) - A'(t_{x-1})$. Values close to 0 in A'' would indicate no change in virus abundance and we therefore classified each point in A'' as belonging to the HCR if its value was smaller than a threshold $A''_t < 0.001$, and to a VDR otherwise. This procedure creates a sequence of points $C(t)$ with each point classified into HCR or VDR (Fig. S17). There may be cases however in which there is some momentary virus growth during host dominance, without a complete escape that leads to regime switch. To include such events within the HCR we calculated a threshold f defined as the 75% quantile value of the distribution of VDR lengths in C. Any sequence of VDR points shorter than f was converted to a HCR. The last condition for classification to a HCR was that a sequence of points had to be longer than the longest-lasting virus outbreak (i.e., the longest HCR length in C).

While the values for the thresholds we have used may seem, to some degree, arbitrary, this would be the case for any other algorithm for classification because a 'regime' is not naturally defined by the dynamical system (i.e., the ODE system), but rather detected in the emerging time series. As an independent corroboration for our method, we imposed the regime boundaries calculated using virus abundance time series to time series of host abundance and virus diversification. The regimes are evident in those time series as well (e.g., Fig. S3). For example, in the virus diversification time series, the rate of diversification is much steeper in the VDR than in the HCR (Fig. S5). This additional verification increases the confidence we have in our ability to classify these regimes. We also provide 100 figures, corresponding to 100 simulations with identified regimes for readers interested in further judging our method.

Network construction

The networks we use are bipartite networks, which contain two sets of nodes (e.g., hosts and spacers or hosts and viruses). Bipartite networks are mathematically represented using incidence matrices. A graphical overview of the networks (and associated matrices) and how we constructed them can be found in Fig. S6. Here, we provide details on network construction.

Networks of genetic composition At each time t, we defined a *host-spacer network*, $\mathcal{S}(t)$ (and associated matrix $\mathbf{S}_{xi}(t)$) as a bipartite network in which each edge is drawn between a host strain i and a spacer x (Fig. S6A,E). We analogously defined a *virus-protospacer network* $\mathcal{P}(t)$ (and associated matrix $\mathbf{P}_{yj}(t)$) as a bipartite network in which each edge is drawn between a virus strain j and a protospacer y (Fig. S6B,F). Hence, in each network a host or virus strain's genome is the set of its neighboring spacer or protospacer nodes, respectively.

Immunity network We used $\mathcal{S}(t)$ and $\mathcal{P}(t)$ to define an *immunity network* at a time t, $\mathcal{I}(t)$ (and associated matrix $\mathbf{I}_{ij}(t)$) in which edges are drawn between a virus strain j and a host strain i. Edge weights were defined as the number of matching spacers and protospacers between $\mathcal{S}(t)$ and $\mathcal{P}(t)$ for any given host-virus pair (Fig. S6. That is: $\mathbf{I}_{ij}(t) = \sum_x^{S_i} \sum_y^{G_j} M(S_i^x, G_j^y)$, where M has a value of 1 if spacer x in the spacer set of host i, S_i, is the same as protospacer y in the protospacer set of virus j, G_j and 0 otherwise.

Infection network We defined an *infection network* at a time t, $\mathcal{N}(t)$ (and associated matrix $\mathbf{N}_{ij}(t)$) as a subset of the immunity network in which $\mathbf{I}_{ij}(t) = 0$ (Fig. S6D,H). That is, edges in $\mathbf{N}_{ij}(t)$ were drawn between hosts and viruses that did *not* have an edge in $\mathcal{I}$. We weighted the edges of $\mathbf{N}_{ij}(t)$ by a normalized measure of encounter between virus strain j will and a host i, based on their abundance values: $\mathbf{N}_{ij}(t) = \frac{V_j(t)N_i(t)}{N_T(t)}$, where $V_j(t)$ and $N_i(t)$ are the abundances of a virus strain j and a host strain i at time t and N_T is the total abundance of hosts.

Network analysis

Weighted nestedness We evaluated the nestedness of $\mathbf{I}_{ij}(t)$ using WNODF (*36*). Briefly, the index ranges from 0 to 100, with the maximum 100 representing perfect nestedness. Perfect nest-

edness occurs when all 2x2 sub-matrices of the form

$$\begin{array}{c} \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cc} v_1 & v_2 \end{array} \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{array}$$

satisfy the conditions for host b_1 to be immune to the two viruses via more matches than b_2 ($a > c, b > d, a > d$) and for v_2 to have less matches to the two hosts than v_1 ($a > b, c > d$). We calculated WNODF with the `networklevel` function in the bipartite package (version 2.11) in R. We calculated WNODF when at least 2 strains of both hosts and viruses were present.

In empirical data we evaluated nestedness in two ways. First, we used WNODF. Second, we calculated the largest eigenvalue of $\mathbf{I}_{ij}(t)\mathbf{I}_{ij}^T(t)$, ρ, as suggested by (*46*). We did not use ρ in the simulations because preliminary analyses indicated that this measure is highly dependent on network size and therefore cannot be used to compare between networks. However, it is suitable for comparing an observed network to its shuffled counterparts because for a given distribution of weights an network size high values of ρ indicate a more quantitatively nested structure (*46*). WNODF and ρ are the only two measures for quantitative nestedness we are aware of. WNODF cannot handle networks which are fully connected (i.e., density of 1) and we therefore did not use it for our Yellowstone data set (Fig. S16A).

Community detection To find 'communities', or as commonly referred to, 'modules', we used the Map Equation objective function to calculate the optimal partition of the network (*47*, *48*). Briefly, the Map Equation is a flow-based and information-theoretic method (implemented with Infomap), which calculates network partitioning based on the movement of a random walker on the network (see for details). In any given partition of the network, the random walker moves across nodes in proportion to the direction and weight of the edges. Hence, it will tend to stay longer in dense areas representing groups of, for example, viruses and hosts with high interaction density. These areas can be defined as 'modules'. The time spent in each module can be converted to an information-theoretic currency using an objective function called the Map equation and the 'best' network partition corresponds is the one that minimizes the Map Equation (*47*, *48*). For convenience we use the term 'modules', as it is commonly used to refer partitions of networks across different disciplines, but Infomap does not calculate a modularity function (sensu (*49*)). We chose the map equation over the more commonly used modularity objective function because of the

computational efficiency of its implementation (Infomap) (*50*), given that we needed to analyze hundreds of thousands of networks.

Our networks were bipartite and the modules we detect contain both sets of nodes. The input to Infomap was an edge list of the original bipartite network (as was done in other bipartite implementations of Infomap; (*51*)). Specifically, we ran Infomap using: `Infomap filename -N 20 --tree -2 --seed 123 -i link-list`. Note that a unipartite representation of the same bipartite network obtained by using a block matrix—for example, using $\left(\begin{smallmatrix} \mathbf{S} & 0 \\ 0 & \mathbf{S}^T \end{smallmatrix}\right)$ for the host-spacer network—would generate the same results.

Epidemiological measures

We derived two modified measures of the basic reproductive number (see derivation in SI). The first, $R_j^0(t)$, is the number of offspring produced at time t by virus strain j from infecting all hosts with no protection to it (0-matches) and is defined as:

$$R_j^0(t) = \frac{\beta\phi(1-q)}{\phi N_T(t) + m} \sum_i^{I(t)} (1 - M_{ij}^0) N_i(t) \tag{3}$$

The delta-function M_{ij}^0 equals 1 (and 0 otherwise) when there are no matches between the set of spacers of host i, S_i and the set of protospacers of virus j, G_j. $I(t)$ is the set of hosts in time t. The second, $R_j^1(t)$, is the number of offspring that a virus strain j would produce by escaping protection from hosts via a single mutation (1-match) at time t and is defined as:

$$R_j^1(t) = \frac{\beta\phi(1-q)}{\phi N_T(t) + m} \cdot \frac{1}{g_j} \sum_i^{I(t)} (1 - M_{ij}^1) N_i(t) \tag{4}$$

Here, M_{ij}^1 equals 1 (and 0 otherwise) when there is exactly 1 matching spacer-protospacer pair in the set of spacers of host i, S_i and the set of protospacers of virus j, G_j. g_j is the length of the protospacer cassette of virus j and quantifies the probability that a mutation will hit a particular protospacer. Adding these two quantitites together we obtain the 'potential reproductive number' of a virus strain j, which quantifies the contribution of a viral strain and its potential progeny to population growth, conditional on escape:

$$R_j^{pot}(t) = \frac{\beta\phi(1-q)}{\phi N_T(t) + m} \left[\sum_i^{I(t)} (1 - M_{ij}^0) N_i(t) + \frac{1}{g_j} \sum_i^{I(t)} (1 - M_{ij}^1) N_i(t) \right] \tag{5}$$

Empirical data

The first data set represents a single time point from three adjacent hot springs (NL01, NL10 and NL13) in the Nymph Lake region of Yellowstone National Park. These three springs have shown to share a single well-mixed population of *S. islandicus hosts* (*52*). Viruses were collected from contemporary populations and CRISPR spacers and virus isolates suggest that these are dominated by the lytic virus SIRV (*53*). The second data set is from *S. islandicus* strains isolated from several springs in the Mutnovsky Volcano in Kamchatka Russia that also were shown to share populations of hosts. The predominant viral population in these hosts is the chronic Sulfolobus Spindle Shaped Virus (SSV) (*27*). The third data set consists of longitudinal sampling of human-adapted *P. aeruginosa* isolates from sputum samples of Cystic Fibrosis patients collected at a hospital in Copenhagen, Denmark and a global set of temperate mu-like viruses from *P. aeruginosa* viruses extracted from NCBI to substitute for a lack of sequenced contemporary viruses (*39*, *40*). See detailed methods in SI.

Shuffling of empirical data matrices

We shuffled the matrices of the empirical data by randomly distributing the interactions (function `r00_samp` in package `vegan` (version 2.5-4) in R). This procedure maintains the density of the network and the distribution of weights while shuffling the structure. Therefore, it tests the hypothesis that the for a given distribution of matches the observed network is non-randomly structured in a quantitatively nested way.

Acknowledgments

We thank Whitney England and Matthew Pauly for data processing and collection, and Qixin He for guidance on the construction of the phylogenetic trees from model outputs. RW acknowledges the support of the Cystic Fibrosis Foundation (CFF C2480) and an Allen Distinguished Investigator Award from the Allen Frontiers Institute; MP, that of the University of Chicago. We are also grateful for the access to the computer cluster of the Research Computing Center (RCC) of the University of Chicago.

References

1. D. B. Stouffer, J. Bascompte, en, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 3648–3652 (2011).
2. R. P. Rohr, S. Saavedra, J. Bascompte, en, *Science* **345**, 1253497 (2014).
3. S. Allesina, S. Tang, en, *Popul. Ecol.* **57**, 63–75 (2015).
4. J. M. Olesen *et al.*, *Ecology* **89**, 1573–1582 (2008).
5. J. Bascompte, D. B. Stouffer, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 1781–1787 (2009).
6. S. L. Nuismer, P. Jordano, J. Bascompte, *Evolution* **67**, 338–354 (2013).
7. D. S. Maynard, C. A. Serván, S. Allesina, en, *Ecol. Lett.* **21**, 324–334 (2018).
8. D. P. Vázquez, R. Poulin, B. R. Krasnov, G. I. Shenbrot, *J. Anim. Ecol.* **74**, 946–955 (2005).
9. M. A. Fortuna *et al.*, en, *J. Anim. Ecol.* **79**, 811–817 (2010).
10. S. Pilosof *et al.*, *Nat. Commun.* **5**, 5172 (2014).
11. Q. He *et al.*, en, *Nat. Commun.* **9**, 1817 (2018).
12. S. Pilosof *et al.*, en, *PLoS Biol.* **17**, e3000336 (2019).
13. S. Gupta, K. P. Day, *Parasitol. Today* **446**, 3737–3742 (1994).
14. C. O. Buckee, M. Recker, E. R. Watkins, S. Gupta, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15504–15509 (2011).
15. Y. Artzy-Randrup *et al.*, *Elife* **2012**, e00093 (2012).
16. C. O. Flores, J. R. Meyer, S. Valverde, L. Farr, J. S. Weitz, en, *Proc. Natl. Acad. Sci. U. S. A.* **108**, E288–97 (2011).
17. S. J. Beckett, H. T. P. Williams, en, *Interface Focus* **3**, 20130033 (2013).
18. J. S. Weitz *et al.*, *Trends Microbiol.* **21**, 82–91 (2013).
19. J. Gurney *et al.*, *Network structure and local adaptation in co-evolving bacteria-phage interactions*, 2017, DOI: 10.1111/mec.14008.
20. M. A. Fortuna *et al.*, *Evolution* **73**, 1001–1011 (2019).
21. S. J. Labrie, J. E. Samson, S. Moineau, en, *Nat. Rev. Microbiol.* **8**, 317–327 (2010).
22. J. van der Oost, E. R. Westra, R. N. Jackson, B. Wiedenheft, en, *Nat. Rev. Microbiol.* **12**, 479–492 (2014).

23. L. M. Childs, N. L. Held, M. J. Young, R. J. Whitaker, J. S. Weitz, en, *Evolution* **66**, 2015–2029 (2012).

24. D. Paez-Espino *et al.*, en, *MBio* **6**, DOI: 10.1128/mBio.00262-15 (2015).

25. S. van Houte *et al.*, en, *Nature* **532**, 385–388 (2016).

26. R. A. Daly *et al.*, en, *Nat Microbiol* **4**, 352–361 (2019).

27. M. D. Pauly, M. A. Bautista, J. A. Black, R. J. Whitaker, en, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180093 (2019).

28. R. Barrangou, R. A. Notebaart, en, *Trends Microbiol.* **27**, 489–496 (2019).

29. D. Bikard, R. Barrangou, en, *Curr. Opin. Microbiol.* **37**, 155–160 (2017).

30. E. R. Westra, A. J. Dowling, J. M. Broniewski, S. van Houte, *Annual Review of Ecology, Evolution, and Systematics* **47**, 307–331 (2016).

31. L. M. Childs, W. E. England, M. J. Young, J. S. Weitz, R. J. Whitaker, en, *PLoS One* **9**, e101710 (2014).

32. S. Gupta, K. P. Day, *Parasite Immunol.* **16**, 361–370 (1994).

33. S. Gupta, N. Ferguson, R. Anderson, *Science* **280**, 912–915 (1998).

34. H. Chabas *et al.*, en, *PLoS Biol.* **16**, e2006738 (2018).

35. W. Ulrich, M. Almeida-Neto, N. J. Gotelli, *Oikos* **118**, 3–17 (2009).

36. M. Almeida-Neto, W. Ulrich, *Environmental Modelling & Software* **26**, 173–178 (2011).

37. C. Song, R. P. Rohr, S. Saavedra, *J. Anim. Ecol.* **86**, ed. by A. Eklöf, 1417–1424 (2017).

38. N. L. Held, A. Herrera, H. Cadillo-Quiroz, R. J. Whitaker, en, *PLoS One* **5**, DOI: 10.1371/journal.pone.0012988 (2010).

39. W. E. England, T. Kim, R. J. Whitaker, *mSystems*, DOI: 10.1128/mSystems.00075-18 (2018).

40. R. L. Marvig, L. M. Sommer, S. Molin, H. K. Johansen, en, *Nat. Genet.* **47**, 57–64 (2015).

41. C. Fontaine *et al.*, *Ecol. Lett.* **14**, 1170–1181 (2011).

42. B. R. Krasnov *et al.*, *Am. Nat.* **179**, 501–511 (2012).

43. D. Zinder, T. Bedford, S. Gupta, M. Pascual, *PLoS Pathog.* **9**, e1003104 (2013).

44. R. D'Andrea, A. Ostling, *Am. Nat.* **187**, 130–135 (2016).

45. J. Iranzo, A. E. Lobkovsky, Y. I. Wolf, E. V. Koonin, en, *J. Bacteriol.* **195**, 3834–3844 (2013).

46. P. P. A. Staniczenko, J. C. Kopp, S. Allesina, en, *Nat. Commun.* **4**, 1391 (2013).

47. M. Rosvall, C. T. Bergstrom, en, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1118–1123 (2008).

48. M Rosvall, D Axelsson, C. T. Bergstrom, en, *Eur. Phys. J. Spec. Top.* **178**, 13–23 (2010).

49. M. E. J. Newman, M Girvan, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69**, 026113 (2004).

50. A. Lancichinetti, S. Fortunato, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **80**, 056117 (2009).

51. D. Edler, T. Guedes, A. Zizka, M. Rosvall, A. Antonelli, en, *Syst. Biol.* **66**, 197–204 (2017).

52. K. M. Campbell *et al.*, *Sulfolobus islandicus meta-populations in Yellowstone National Park hot springs*, 2017, DOI: 10.1111/1462-2920.13728.

53. M. A. Bautista, J. A. Black, N. D. Youngblut, R. J. Whitaker, en, *Viruses* **9**, DOI: 10.3390/v9050120 (2017).

54. D. T. Gillespie, *J. Comput. Phys.* **22**, 403–434 (1976).

55. *Extracting data from BLAST databases with blastdbcmd* (National Center for Biotechnology Information (US), 2008).

56. S. F. Altschul, W Gish, W Miller, E. W. Myers, D. J. Lipman, en, *J. Mol. Biol.* **215**, 403–410 (1990).

Supplementary Material

A Implementation of dynamical model

We used a Gillespie algorithm to introduce stochasticity in spacer acquisition and protospacer mutation. The Gillespie algorithm was originally conceived for chemical reactions (*54*) and is now widely used in epidemiology and ecology (e.g. (*11*, *14*, *15*, *23*)). The basic algorithm consists of three steps: (1) determination of the random time to the next reaction (also termed 'event'); (2) random selection of a specific event from a set of events using weighted sampling (events with higher rates are more likely to be chosen); and (3) updating the number of chemical molecules (in our case, host or virus strains) involved in the selected event. Mathematically, consider N stochastic events with corresponding rates $a_1, a_2, \ldots, a_N$. The random time to an event is given by $t_{\text{event}} = \frac{1}{a}\ln(\frac{1}{r}))$, where $a = \sum_{i=1}^{N} a_i$ and r is a random number uniformly distributed in the range $[0, 1)$. A specific event is selected with probabilities proportional to its rate, and the corresponding state numbers concerning this event are updated.

In the CRISPR coevolutionary model, we consider two stochastic events: spacer acquisition and protospacer mutation. To apply the Gillespie algorithm, we first calculate the spacer acquisition rates for all host strains and the mutation rates for all protospacers in all virus strains. We use these rates to then determine the time to the next stochastic event, to select a virus or host strain for the given change, and to add the newly generated host (in case of spacer acquisition) or virus (in case of a mutation event) strain to the system.

Mathematically, the spacer acquisition rate for host strain i is $a_i = q\phi \sum_j N_i V_j$ and the protospacer mutation rate for protospacer y of virus strain j is $\mu_{jy} = \mu\beta\phi(1-q)\sum_i N_i V_j(1 - M_{ij}) + \mu\beta\phi p \sum_i N_i V_j M_{ij}$. The random time to the next stochastic event is $t_{\text{event}} = \frac{1}{a}\ln(\frac{1}{r}))$ where $a = \sum_i a_i + \sum_j \sum_y \mu_{jy}$ and r is a random number uniformly distributed in the range $[0, 1)$. With probability $\frac{\sum_i a_i}{a}$, this process randomly selects a specific host to acquire a spacer from a random virus strain (with probabilities proportional to their spacer acquistion rates). Otherwise, this process randomly selects a specific protospacer in a specific virus strain to be mutated (with probabilities proportional to their mutation rates), generating a new virus strain.

This simulation scheme adopts the deterministic treatment of population dynamics in between stochastic events, and follows the protocol proposed in (*23*), where more details can be found. Although, in principle, population dynamics could also be implemented stochastically (*45*), this

approach would have been more time-consuming given the large numbers of viruses and host strains.

B R_0 expression

We derive here the general expression for the basic reproductive number (R_0) of the system. The growth rate of a virus strain j is given by:

$$\frac{dV_j}{dt} = (1-q)\beta \sum_i \left(1 - M(S_i, G_j)\right) \phi_{ij} N_i V_j \; + \; p\beta \sum_i M(S_i, G_j) \phi_{ij} N_i V_j \; - \; \sum_i \phi_{ij} N_i V_j \; - \; m V_j$$

where $M(S_i, G_j)$, or more simply M_{ij}, is the number of matches between the set of spacers of bacteria i and the set of protospacers of virus j.

The condition for the spread of this virus strain $\frac{dV_j}{dt} > 0$ becomes

$$\left[(1-q)\beta \sum_i \left(1 - M(S_i, G_j)\right) \phi_{ij} N_i V_j \; + \; p\beta \sum_i M(S_i, G_j) \phi_{ij} N_i V_j \; - \; \sum_i \phi_{ij} N_i V_j \; - \; m V_j\right] > 0$$

or

$$\left[\beta \sum_i \left((1-q)(1-M_{ij}) + pM_{ij}\right) \phi_{ij} N_i V_j \; - \; \sum_i \phi_{ij} N_i V_j \; - \; m V_j\right] > 0$$

By assuming that $\phi_{ij} = \phi$ is constant and the same for all strains,

$$\left[\beta\phi \sum_i \left((1-q)\left(1 - M_{ij}\right) + \; pM_{ij}\right) N_i\right] > \phi \sum_i N_i + m$$

Then,

$$\frac{\beta\phi}{\phi \sum_i N_i + m} \left[\sum_i \left((1-q)\left(1 - M_{ij}\right) + \; pM_{ij}\right) N_i\right] > 1$$

With the total abundance of bacteria $N_T = \sum_i N_i$,

$$R_0^j \equiv \frac{\beta\phi}{\phi N_T + m} \left[\sum_i \left[(1-q)\left(1 - M_{ij}\right) + \; pM_{ij}\right] N_i\right] > 1 \qquad \text{(S1)}$$

Thus,

$$R_0^j = \begin{cases} \frac{\beta\phi}{\phi N_T + m}\left((1-q)(\sum_i N_{i|_{M_{ij}=0}})\right) & if \quad M_{ij} = 0 \\ \\ \frac{\beta\phi}{\phi N_T + m}\left(p(\sum_i N_{i|_{M_{ij}>0}})\right) & if \quad M_{ij} > 0 \end{cases}$$

We can call the first term of equation S1 the effective R_0, or $R_{0_{eff}}$, resulting from feasible infections of bacteria that do not have immunity to the given virus. The second term can be neglected since the probability of CRISPR failure $p \ll 1$.

C Potential R_0 resulting from an escape mutation

We define the potential reproductive number of a virus as the number of offspring it would produce from a single escape mutation. For this, we need to consider the contribution to this number from infections of bacteria that are protected by a single match, R_{0_1}. We let q_j denote the number of protospacers per strain and write

$$R_0^{j_{pot}} = R_{0_{eff}} + R_{0_1}$$

Then,

$$R_0^j = \frac{\beta\phi(1-q)}{\phi N_T + m}\left[\sum_i (1 - M_{ij}^0)N_i + \frac{1}{g_j}\sum_i M_{ij}^1 N_i\right] \tag{S2}$$

where M_{ij}^1 is 1 only when there is a single match between the pair of virus j and bacteria i, and $\frac{1}{q_j}$ is the probability that the mutation falls on a particular protospacer of virus j.

D Lotka-Volterra mean-field dynamics

The simplest model whose dynamics can be compared to those of the full model corresponds to the mean-field differential equations describing the bacteria-virus interactions when the structure of who is protected from whom is randomized. That is, in this model, the emergent structure of the protection matrix is randomized, and the immunity matrix is reduced to a fixed density of edges immunity. Starting from the full model from Childs et al.,

$$\frac{dN_i}{dt} = r_i N_i \left(1 - \frac{\sum_i N_i}{K}\right) \; - \; (1-q) \sum_j \left(1 - M(S_i, G_j)\right) \phi_{ij} N_i V_j \; - \; p \sum_j M(S_i, G_j) \phi_{ij} N_i V_j$$

$$\frac{dV_j}{dt} = (1-q)\beta \sum_i \left(1 - M(S_i, G_j)\right) \phi_{ij} N_i V_j \; + \; p\beta \sum_i M(S_i, G_j) \phi_{ij} N_i V_j \; - \; \sum_i \phi_{ij} N_i V_j \; - \; m V_j$$

and simplifying some notation, we have

$$\dot{N}_i = r N_i \left(1 - \frac{\sum_i N_i}{K}\right) \; - \; \left[(1-q) \sum_j (1 - M_{ij}) V_j \; + \; p \sum_j M_{ij} V_j\right] \phi N_i$$

$$\dot{V}_j = \beta\phi \left[(1-q) \sum_i (1 - M_{ij}) N_i \; + \; p \sum_i M_{ij} N_i\right] V_j \; - \; \left(\phi \sum_i N_i + m\right) V_j$$

We assume that the entries of the immunity matrix M_{ij} (defined by the matches between the set of spacers (S_i) of the host i and the set of protospacers (G_j) of the virus j) take a constant value in time, $M \equiv \langle M_{ij} \rangle$. We further rewrite the system for the total abundance of virus V and bacteria N as the following general predator-prey system:

$$\dot{N} = rN \left(1 - \frac{N}{K}\right) \; - \; \left[(1-q)(1 - \langle M_{ij} \rangle) \; + \; p \langle M_{ij} \rangle\right] \phi V N$$

$$\dot{V} = \beta\phi \left[(1-q)(1 - \langle M_{ij} \rangle) \; + \; p \langle M_{ij} \rangle\right] N V \; - \; (\phi N + m) V$$

With $\mathbb{M} \equiv [(1-q)(1-\langle M_{ij}\rangle) \; + \; p\langle M_{ij}\rangle]$, we can rewrite the system as:

$$\dot{N} = rN\left(1-\frac{N}{K}\right) \; - \; \mathbb{M}\phi VN \tag{S3}$$

$$\dot{V} = \beta\phi\mathbb{M}NV \; - \; (\phi N + m)\, V \tag{S4}$$

D.1 Equilibrium points

The above two-dimensional system (S3 and S4) has 3 equilibrium points $C_i = (N^*, V^*)$:

- The trivial point $C_1 = (0, 0)$.
- The point where hosts have reached their carrying capacity and the viruses have gone extinct, $C_2 = (K, 0)$.
- The positive coexistence point defined by the intersection of the nullclines of the system

$$C_3 = \left[\frac{m}{\phi(\beta\mathbb{M}-1)}, \frac{r}{\phi\mathbb{M}}\left(1-\frac{m}{\phi K(\beta\mathbb{M}-1)}\right)\right] \tag{S5}$$

These nullclines are both a function of $\mathbb{M}$ and are given by:

$$N^* = \frac{m}{\phi(\beta\mathbb{M}-1)}, \qquad V^* = \frac{r}{\phi\mathbb{M}}\left(1-\frac{N^*}{K}\right). \tag{S6}$$

The local (asymptotic) stability of these equilibria can be determined based on the Jacobian matrix of the system :

$$\mathbb{J} = \begin{bmatrix} r\left(1-\dfrac{2N}{K}\right) - \mathbb{M}\phi V & -\mathbb{M}\phi N \\ \\ \beta\mathbb{M}\phi V & \beta\mathbb{M}\phi N - (\phi N + m) \end{bmatrix}$$

Together the expressions for the nullclines and for the Jacobian show that the location of the equilibria in phase space and their local stability depend on $\mathbb{M}$ and therefore, on $M = \langle M_{ij}\rangle$. The change in the equilibria as a function of M is illustrated in Figure S1. These dynamics can be compared to those of the full system for the aggregated abundance of all viruses and hosts S2.

E Neutral model without explicit immunity (Randomization of the immunity matrix)

By construction, the above mean-field simplification of the system does not explicitly include the diversity of hosts and viruses. To determine the importance of both specific immune memory and the structure associated with it, we consider here a neutral model in which the ability to the system to generate diversity is not affected (that is, both the mutation of viruses and the acquisition of spacers by the bacteria are still included) but the specificity of immune memory is removed. We specifically randomize every Δt the matches $M(S_i, G_j)$ between the protospacers set of viruses (G_j) and the spacers set in hosts (S_i). For every virus-host strain pair, we randomize their match by choosing a random number r and comparing it with the *average density* of the immunity network, ρ. Then, we establish a match for this pair if $r < \rho$. This preserves the average density of matches but randomizes their identity.

Under this null model, the system shows neither the regime of virus diversification (VDR) nor that of host dominance (BDR) (results not shown). This comparison to the full model established that the structure of specific immunity generated by the CRISPR system is necessary to create the observed dynamics of the full model.

F Detailed methods for empirical data collection

Our data consists of three data sets:

Yellowstone This dataset was collected in hot springs in Nymph Lake at Yellowstone National Park and consists of a population of *Sulfolobus islandicus* and its contemporary lytic (Sulfolobus islandicus rod-shaped viruses: SIRVs).

Russia This dataset consists of a set of *S. islandicus* strains isolated from Kamchatka, Russia (in 2000) and sympatric chronic viruses (Sulfulobus spindle-shaped viruses: SSVs) (*53*).

Pseudomonas This dataset consists of longitudinal sampling of human-adapted *Pseudomonas aeruginosa* isolates from sputum samples of Cystic Fibrosis patients collected at a hospital in Copenhagen, Denmark and a global set of temperate and lytic *P. aeruginosa* viruses extracted from NCBI to substitute for a lack of sequenced contemporary viruses (*38–40*). Viruses were

grouped based on nucleotide similarity into families known as clusters, and these clusters were assigned a number identifiers which have been described in a previous study (4). In this study we only used viruses from cluster 3 to avoid a false positive result in which we obtain a nested structure due to immune patterns that depend on the cluster (phylogeny).

Illumina sequenced reads from samples were quality filtered using prinseq with the following arguments: `-derep 1245 -lc_method entropy -lc_threshold 50 -trim_qual_right 30 -trim_qual_left 30 -trim_qual_type min -trim_qual_rule lt -trim_qual_window 5 -trim_qual_step 1 -trim_tail_left 5 -trim_tail_right 5 -min_len 66 -min_qual_mean 30 -ns_max_p 1 -verbose.` Spacers were extracted from quality filtered sequencing reads using an in-house bioinformatic pipeline (in preparation, code available upon request). These scripts utilize known repeats from *S. islandicus* and *P. aeruginosa* and BLASTn to extract spacers from reads located between any repeats. BLASTn cutoffs for repeats against reads were based on an e-value of 0.001 with the `-task blastn-short` argument. After spacer extraction, spacers are grouped based on a hamming distance cutoff by comparing spacers as strings of nucleotides and using a sliding window across each string of basepairs (in preparation, code available upon request). We define unique spacers that have 100% nucleotide identity to one another, using a hamming distance of 0 between nucleotide sequences. Unmatched overhanging base pairs between spacers were considered mismatches since these are likely independently acquired spacers from sequential protospacers with different PAM sequences. Unique spacers were mapped to strains in order to determine the spacer set per strain. *Sulfolobus islandicus* isolates in Nymph Lake contained on average 256 spacers per strain ranging from 62 to 520 spacers. *Sulfolobus islandicus* isolates from Kamchatka contained on average 181 spacers per strain ranging from 20 to 795 spacers. *P. aeruginosa* isolates contained 34 spacers on average with a range of 4 to 64 spacers. *P. aeruginosa*, Nymph Lake *S. islandicus*, and Mutnovsky *S. islandicus* isolates contained 40, 40 and 50 total alleles respectively.

Spacer matches to protospacers were initially found using BLASTn with a `-task blastn-short` argument, with an e-value minimum of 0.01. The *P. aeruginosa* database contained 6,231,702 total bp, with 98 viral genomes ranging from 3,588 bp to 309,208 bp. The SSV BLASTn database contained 34 genomes containing 514,147 total bases with the longest sequence being 18,548 bp and the shortest being 11,323 (*27*, *39*). The SIRV BLASTn database was composed of 10 genomes containing 347,896 total bases with the longest sequence being 32,308 bp and the shortest being 32,308. Protospacer BLAST matches were extended to 3 base pairs longer than the length of the spacer and retrieved with blastdbcmd tool from the blast+ package to retrieve the PAM sequences (*55*,

Table S1 Description and values of model parameters used in simulations, following (*23*), except for a higher numbers of spacers and protospacers.

Parameters	Description	Value
r	Growth rate	$1h^{-1}$
K	Carrying Capacity	$10^{5.5}$ mL^{-1}
β	Burst size	50
ϕ	Adsorption rate	10^{-7} mL/h
m	Viral decay rate	0.1 h^{-1}
μ	Viral mutation rate per protospacer and replication	$1 - 5 \times 10^{-7}$
ρ_c	Density cutoff	0.1 mL^{-1}
p	CRISPR failure probability per infection	10^{-5}
q	Spacer acquisition probability per infection	10^{-5}
N_p	Range of protospacers in a virus strain	$15 - 75$
N_s	Range of spacers in a host strain	$10 - 50$

56). Gaps found between alignments were added to these extended protospacer matches. Gaps, or insertion/deletion events, were considered as mismatch when comparing along the entire length of the aligned protospacer and spacer.

We analyzed our data with two criteria. The first criterion was a perfect match with zero mismatches between the entire length of the aligned spacer and protospacer (100% alignment) and with correct PAM sequences in the correct orientation based on the 3bp extensions. The range in number of protospacers found in *S. islandicus* viruses was 39-41 in SIRVs with and 11-47 in SSVs at the most stringent criterion. The range of protospacers found in viruses was 0-32 with *P. aeruginosa* viral database. In our second criterion, we allowed 4 mismatches and do not require a specific PAM match. The number of protospacer matches to viruses ranged from 138 to 167 with SIRVs and 6-24 in SSV for 4mm criterion. Protospacer matches ranged from 0-76 with *P. aeruginosa* viruses. Our results were not qualitatively affected by the choice of criterion and we present results for the second one.

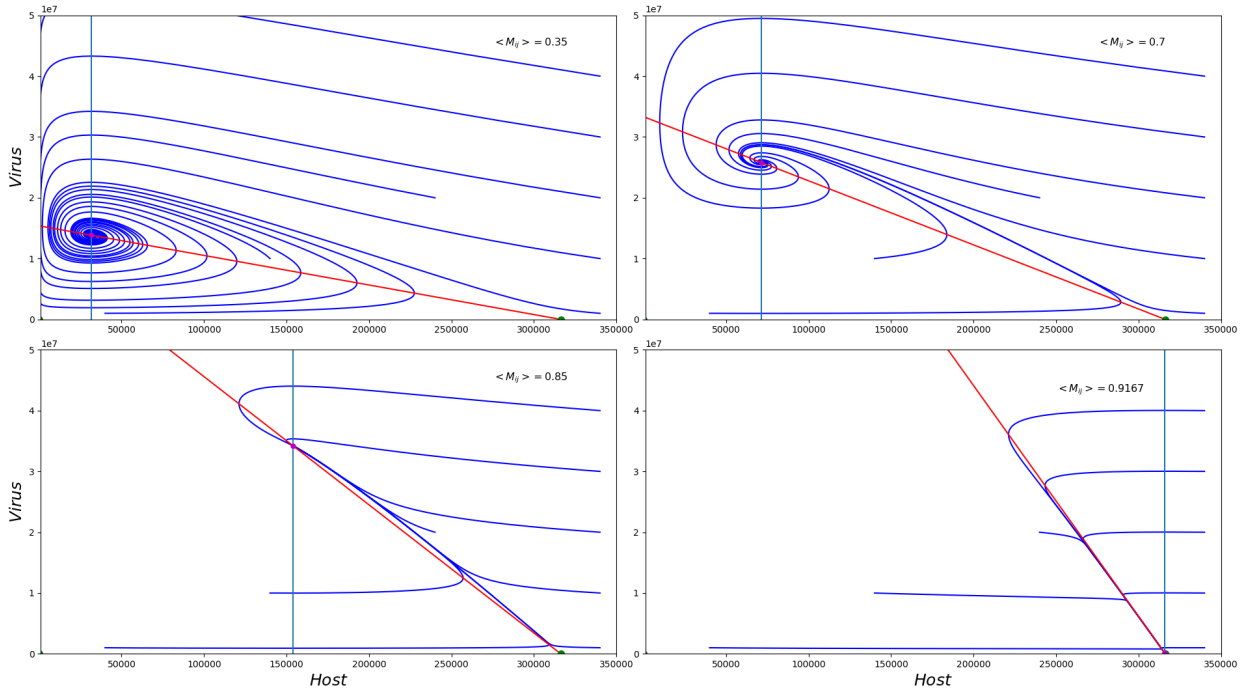
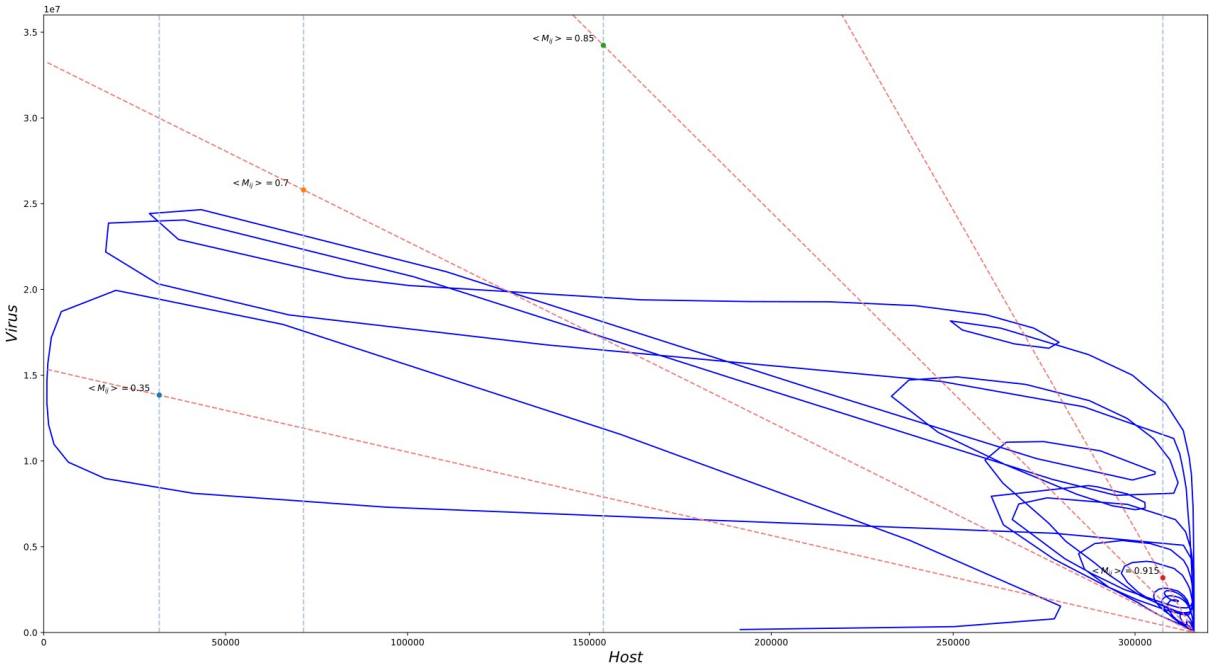


Fig. S1 Phase portraits of the mean-field ODE version of the system for increasing values of the average fraction of matches M. Top panels from left to right: $M = 0.35$ and $M = 0.7$; bottom panels (also from left to right): $M = 0.85$ and $M = 0.9167$. As M grows, the fixed point C_3 transitions from a spiral sink to a degenerate nodal sink as it approaches C_2 and finally collapses onto it.



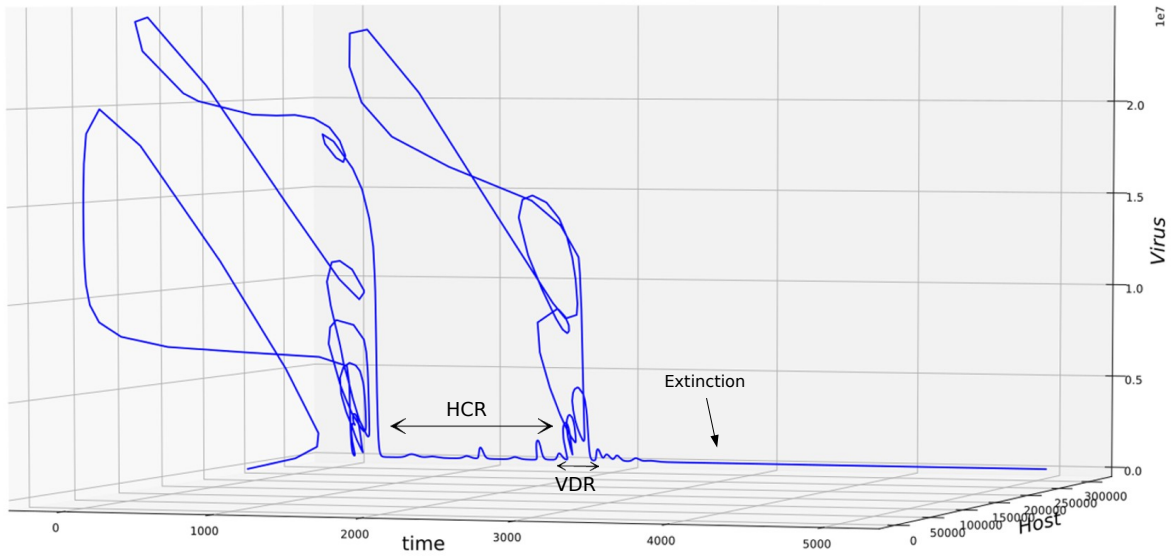


Fig. S2 Phase portrait of the full system for total abundances of hosts and viruses. The dynamics of the full system differ from those of the two-dimensional mean-field ODE as expected from the transitions from coexistence (during the VDR) and the host dominance at carrying capacity (during the HCR) (top panel). During the VDR, the system's trajectory spirals around a moving coexistence point, closest to the fixed points C_3 for increasing M (color points). During the HCR, the system is close to the degenerate sink, collapsing onto C_2 after virus extinction. These different parts of the dynamics become more evident when including time explicitly in the phase portrait (bottom panel).

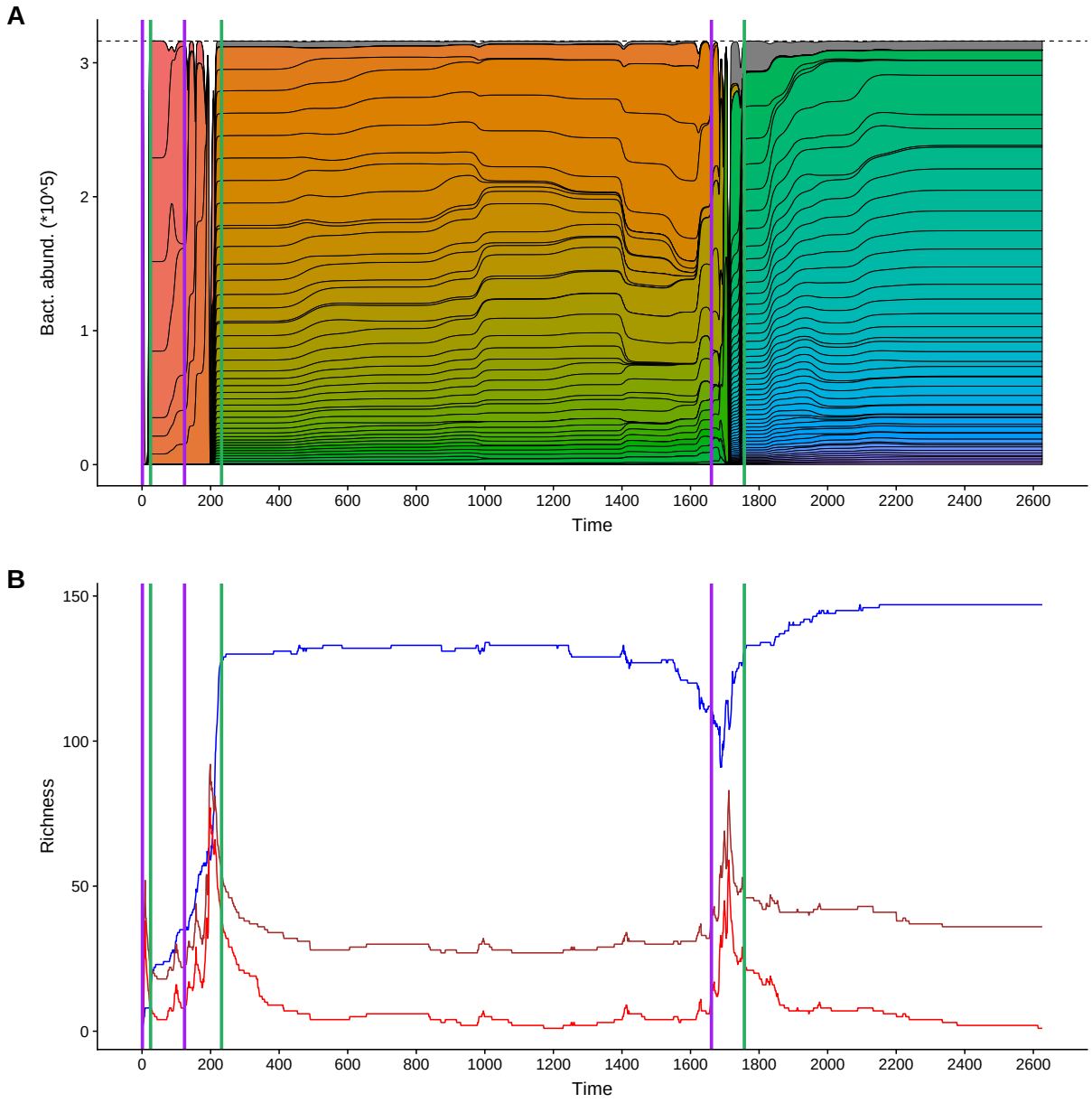


Fig. S3 Viral and host abundance and richness. **(A)** Host abundance. The 100 most abundant strains are colored, the rest are aggregated and shown in gray. **(B)** Richness (i.e., number of unique strains) of hosts (blue) and viruses (red). The number of unique spacers (spacer richness) is depicted in brown. During VDRs the abundance of both hosts and viruses fluctuates. As a response to virus diversification, host richness eventually increases despite possible declines at the beginning of the VDR resulting from the initial viral attack.

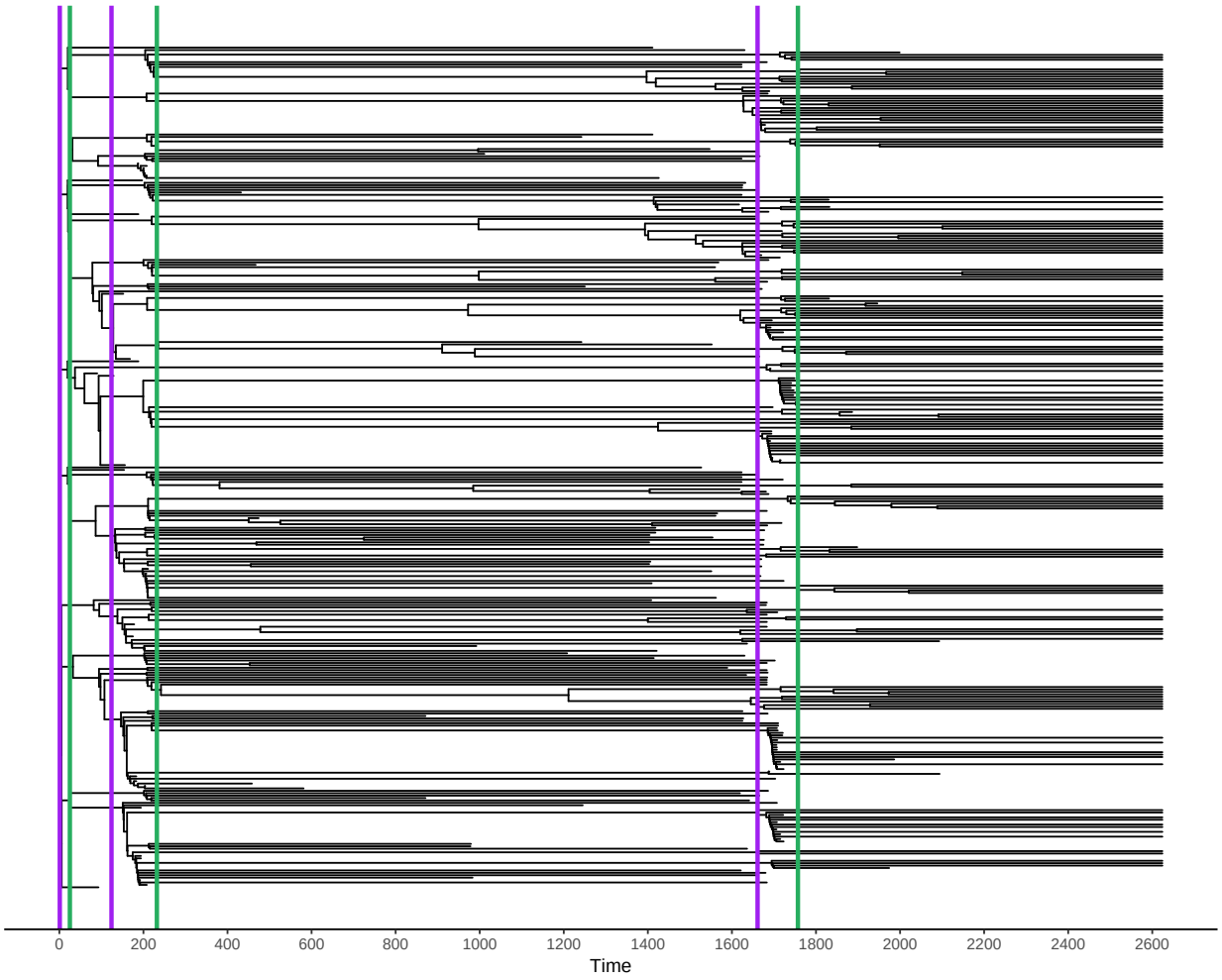


Fig. S4 Host phylogenetic tree. The tree is not inferred, but rather drawn based on exact genealogical data (which strains descends from which) collected during the simulation. Branch length indicates the lifetime of any given host strain. Hosts diversify and go extinct primarily during VDRs, resulting in strain replacement but strains can also persist from one VDR to the next.

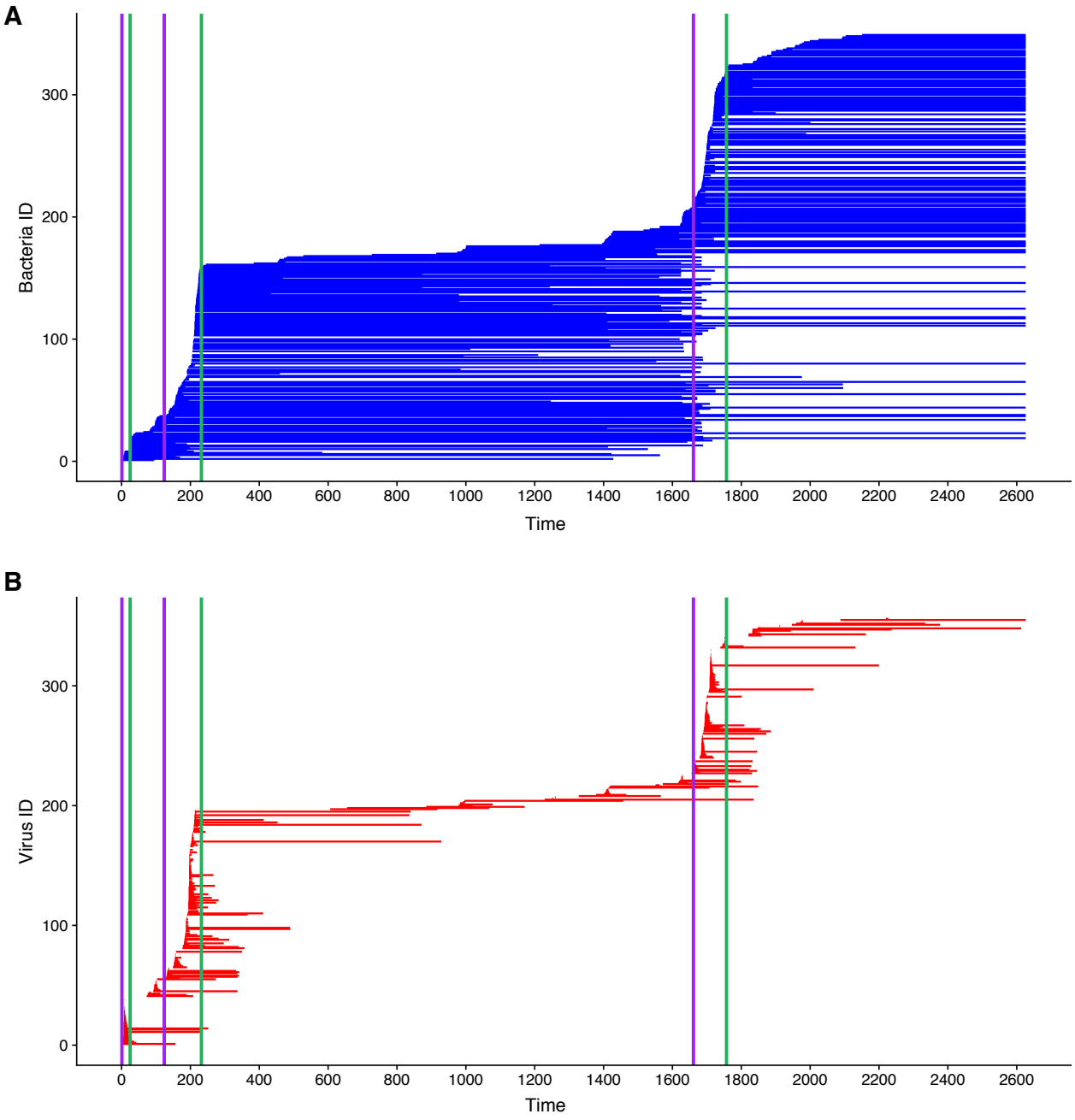


Fig. S5 Host and virus diversification and extinction. Each host **(A)** or viral **(B)** strain is plotted with a line, starting at the time when the strain was generated and ending when the strain went extinct. During VDRs the rate of diversification of both viruses and hosts is higher than during HCRs. While viruses have relatively short persistence (shorter line lengths), hosts have long persistence and can persist during an entire VDR.

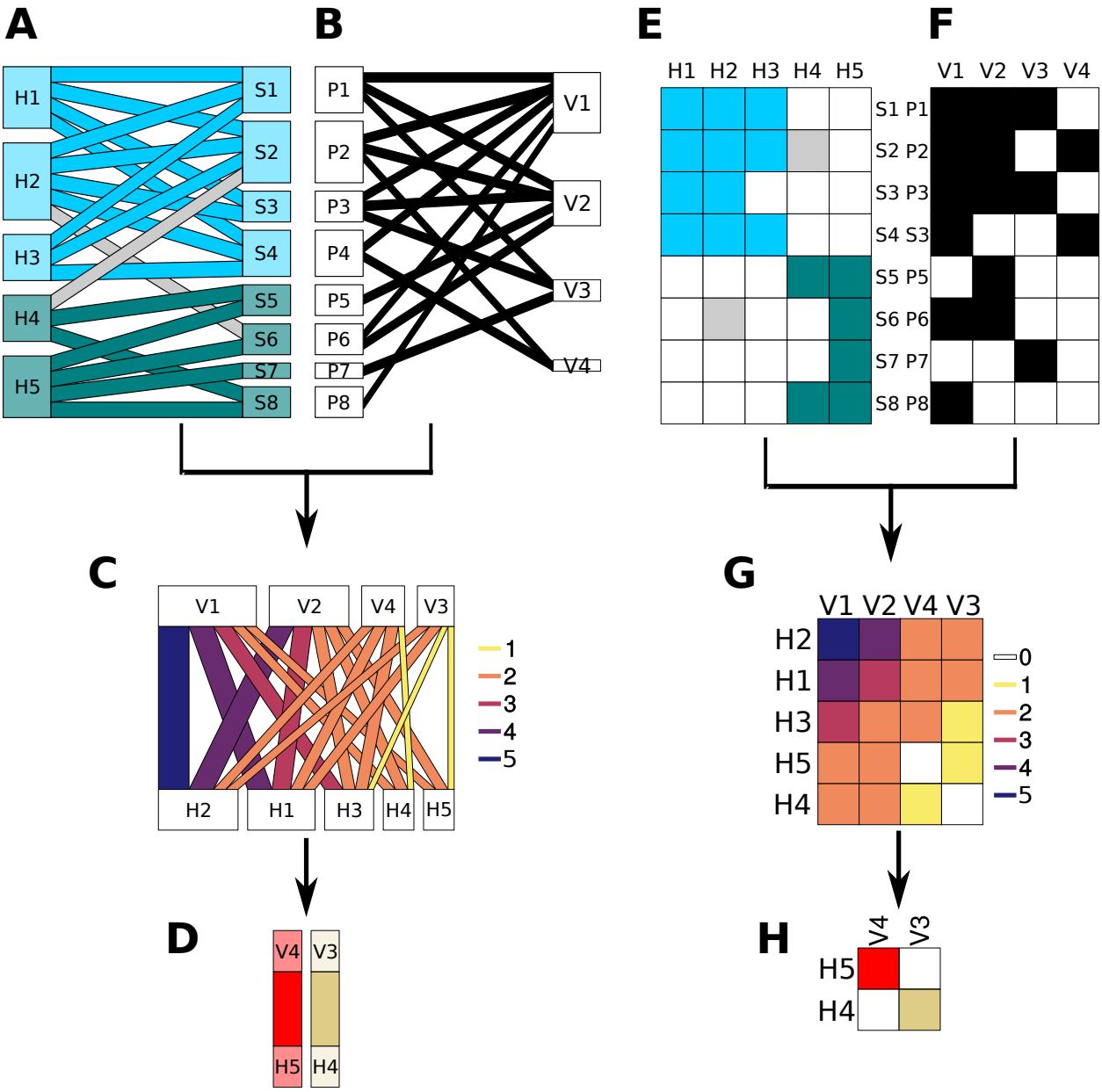


Fig. S6 Structures of diversity. Diagrams illustrating the two different kinds of networks (left column) and associated matrices (right column) used in this study and how they are built. The toy example has 5 hosts (H1-H5), 4 viruses (V1-V4), 8 spacers (S1-S8) and 8 protospacers (P1-P2). **(A)** A bipartite network depicting the spacer composition of hosts. Hosts are affiliated to one of two modules (depicted in light blue and dark green), and interactions can fall either within a module (colored) or outside the module (gray). **(B)** A bipartite network (not modular) depicting the protospacer composition of viruses. **(C)** The immunity network is created by counting the number of shared spacers and protospacers between pairs of hosts and viruses (i.e., matches). Interactions in the network are weighted by the number of matches, here depicted by different colors and width. **(D)** The infection network is created by considering unrealized interactions in the immunity network (equivalent to 0-matches in (G)). Here, only two such interactions exist, between 2 viruses and 2 hosts. There are two modules, depicted in colors, with interactions occurring within the modules only. **(E)** The counterpart matrix of the network in (A). Interactions (matrix cells) depict the occurrence of a spacer in a host strain, and are colored as in (A). **(F)** The counterpart matrix of the network in (B). Matrix cells depict the occurrence of a protospacer in a virus strain. **(G)** The counterpart matrix of the network in (C), with colors depicting the number of spacer matches. The matrix is organized by the sum of columns and rows and is quantitatively nested. This is the matrix we use in Fig. 2B . **(H)** The counterpart matrix for the network in D. The two interactions correspond to the empty matches in (G). This is the matrix we use in Fig. 2A.

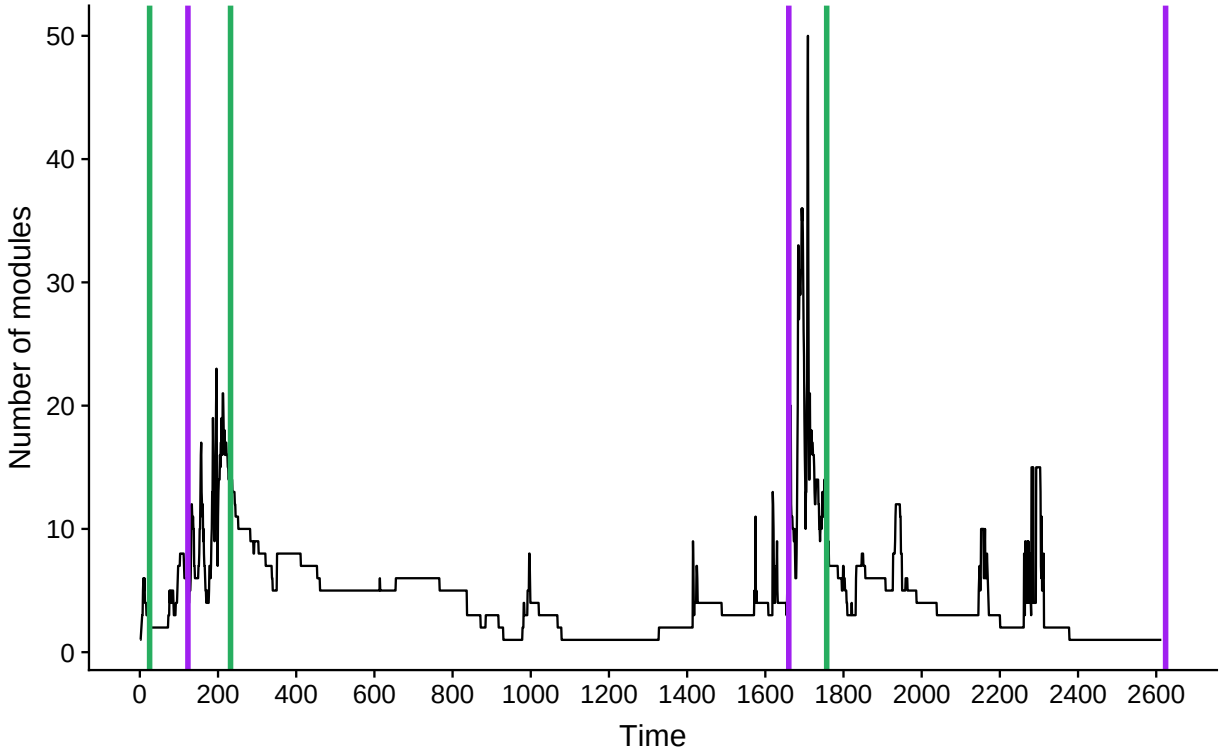


Fig. S7 Modules in the infection network. A time series of the number of modules in the infection network for a single simulation. Modularity enables diversification since it allows the temporary coexistence of different groups of viruses.

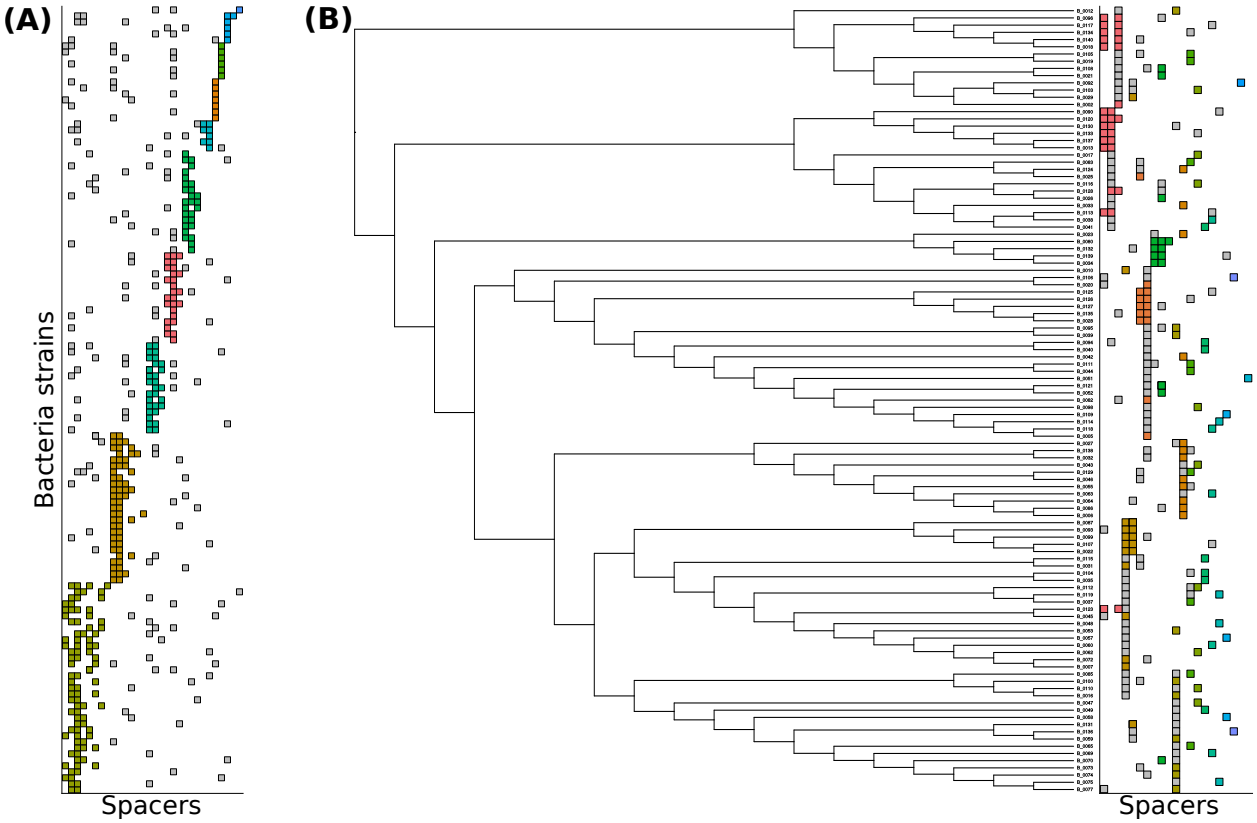


Fig. S8 Modularity in host-spacer networks. **(A)** Each interaction in the network (matrix cells) indicates the occurrence of a spacer in the genome of a host. Colored interactions fall inside modules of hosts containing similar spacers. Each module has a different color. Gray interactions are those that fall outside all modules. **(B)** When hosts are ordered by phylogeny the modules are shuffled. This indicates a lack of phylogenetic signal in the genome of hosts, which does not emerge due to clonality.

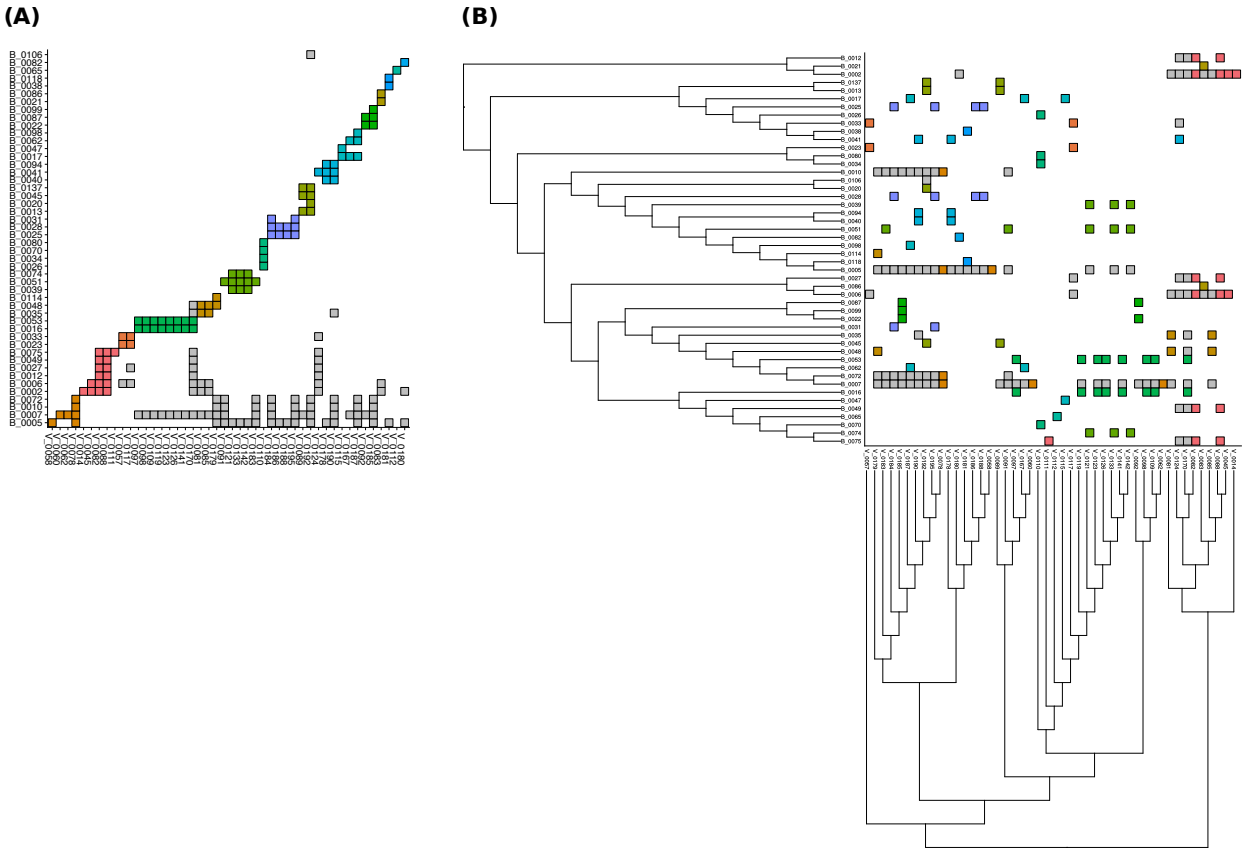


Fig. S9 Lack of phylogenetic signal in infection network. **(A)** Each interaction in the network (matrix cells) indicates infection of a host strain to by a virus strain. Colored interactions fall inside modules (depicted by different colors) of virus strains that infect similar hosts. Gray interactions are those that fall outside all modules. **(B)** When hosts and viruses are ordered by phylogeny the modules are shuffled, indicating a lack of phylogenetic signal in the modular structure.

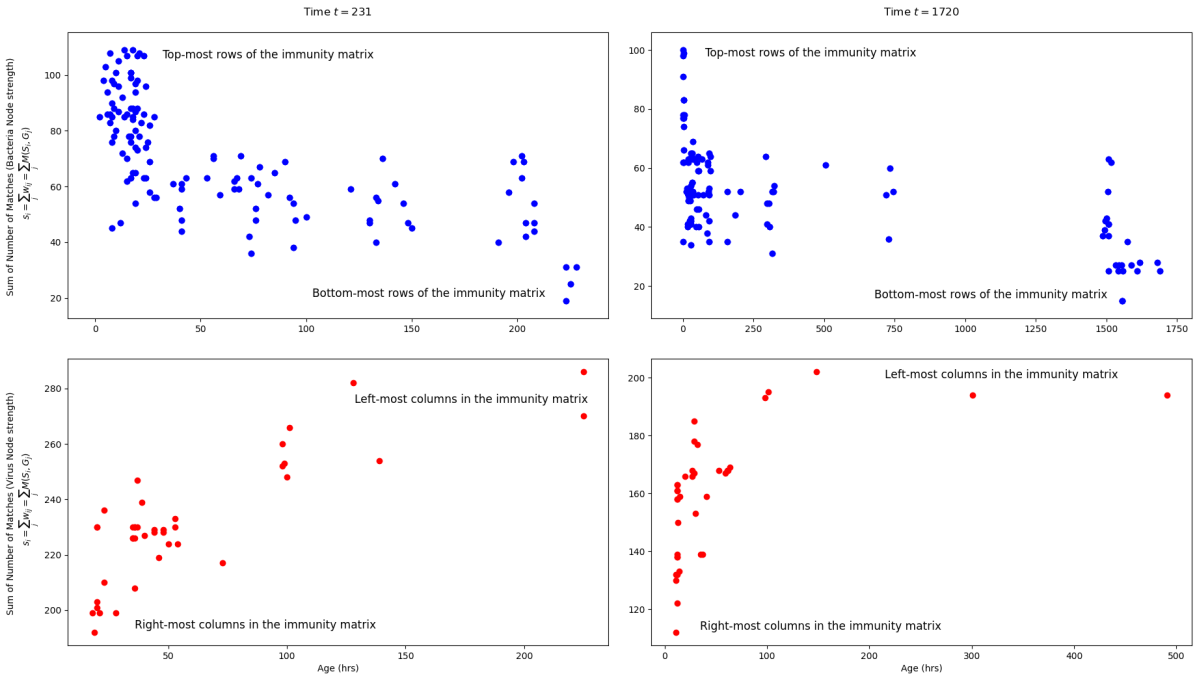


Fig. S10 Host and virus rankings in the weighted nested immunity matrix as a function of age. The plots show node strength (sum of the corresponding number of matches in the immunity network) of hosts (top) and viruses (bottom) against their age measured from the time of their birth, for two selected times before the start of an HCR (left, t=231 and right, t=1720). Each data point represents a host (top) or virus (bottom) strain. On the different plots, selected groups of youngest and oldest strains are indicated. The oldest host strains occupy the lower rows (low node strength), and their rankings tend to decrease with age. This is because descendants of a given host inherit all of its spacers and add a new one, which always results in an increase in total matches (host node strength). Because they have acquired additional protection, they can grow in abundance and through resulting enhanced encounters and infections, the failure of existing spacers can add redundancy (more than one spacer to the same virus), further contributing to their ranking. The oldest viruses occupy the leftmost columns, with the highest column sums of matches to hosts, since longer lifetimes provide the opportunity for many encounters, and therefore for both the failure of existing spacers (which adds redundancy to a given entry) and the addition of new spacers (which distributes immunity throughout entries). A successful offspring will have mutated a protospacer that confers escape from a given match of the parent; thus, successful descendants exhibit one less match and are placed to the right. It is worth noting that there is considerable variation around the general trends with age, reflecting the complex interplay of the stochastic acquisition of spacers and protospacers with the abundance dynamics which affect both encounter and mutation rates.

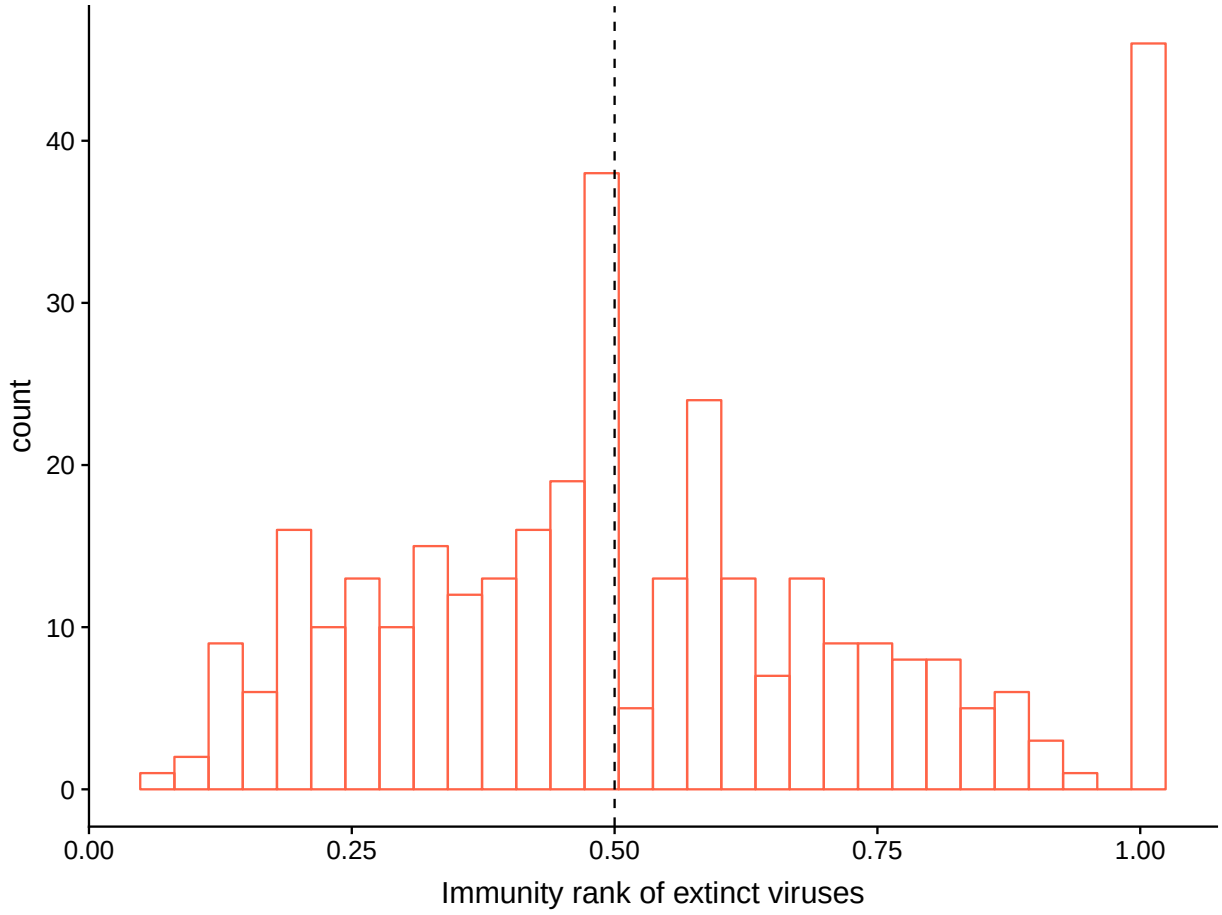


Fig. S11 Order of extinctions. We tested for "orderly" extinctions in which extinction preferentially happens from the viruses to which hosts have most immunity to those that can infect many hosts. For any virus that went extinct we calculated an 'immunity rank'. Specifically, for a given time step, we calculated the strength of all n virus nodes in the immunity network, $s = (s_1, s_2, \ldots, s_j, \ldots, s_n)$, where s_j is the node strength of virus j (i.e., the sum of the columns in Fig 2B). Viruses with higher values of s_j are those that are more to the left in Fig 2B, and to which hosts have high immunity. We removed duplicate values in s (to avoid ties) and ordered it in ascending order to obtain s'. We then calculated the relative position of s_j in s'. A rank of 1 means that the virus that went extinct was highly ranked (e.g., position 5 out of 5 values will render a rank of 1). 50% of viruses (median indicated by a vertical dashed line) had an extinction rank of 0.5.

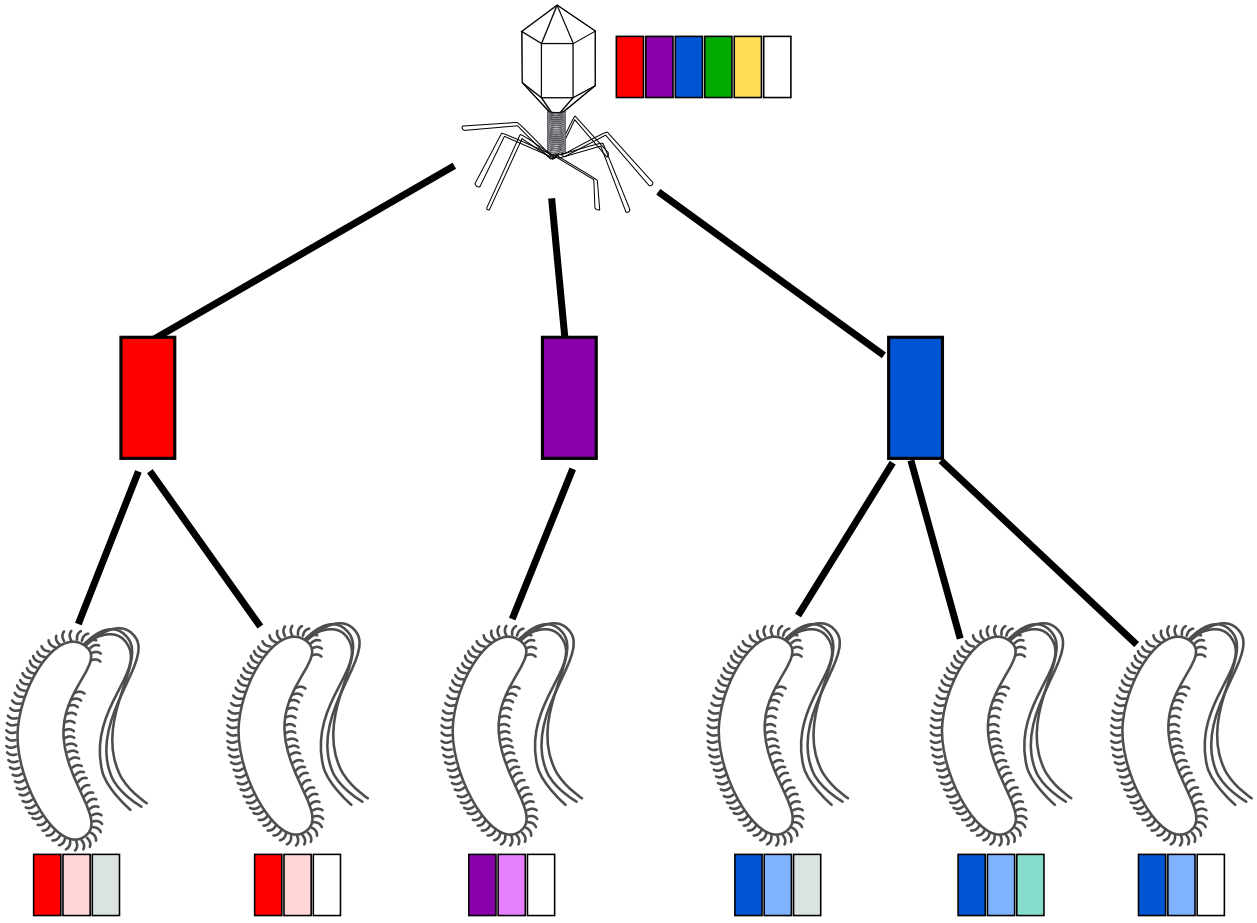


Fig. S12 Viral escape via 1-matches. A tripartite virus-protospacer-host network depicting escape routes for a single virus. Each host is connected to a single protospacer (colored boxes). The spacer composition of strains is shown. Escape occurs through matching colors.

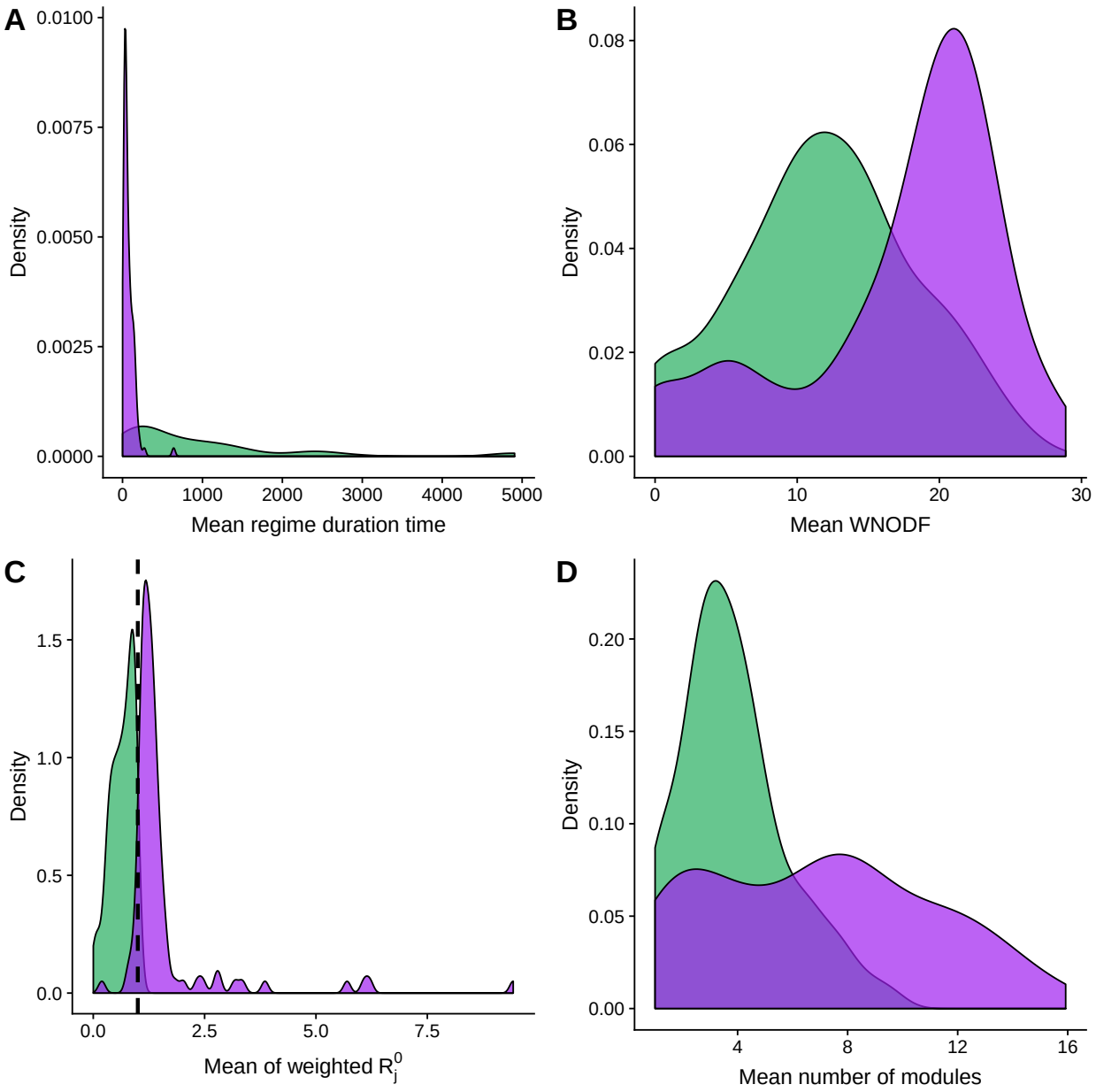


Fig. S13 Results for multiple simulations comparing VDRs to HCRs. Results are summarized using distributions of the main characteristics of the regimes and their corresponding network structures. **(A)** Regimes length: VDRs are shorter than HCRs. **(B)** Nestedness is higher during VDRs than HCRs. **(C)** The basic reproductive number is higher during HCRs, indicating that virus reproduction is higher. During HCRs the basic reproductive number is generally small than one (dashed vertical line). **(D)** The infectino network has more modules during VDRs due to virus diversification. In all plots the average was first taken for each regime type (VDR or HCR) within each simulation. Then, these averages were plotted as a distribution of 100 simulations. VDRs and HCRs are in purple and green, respectively.

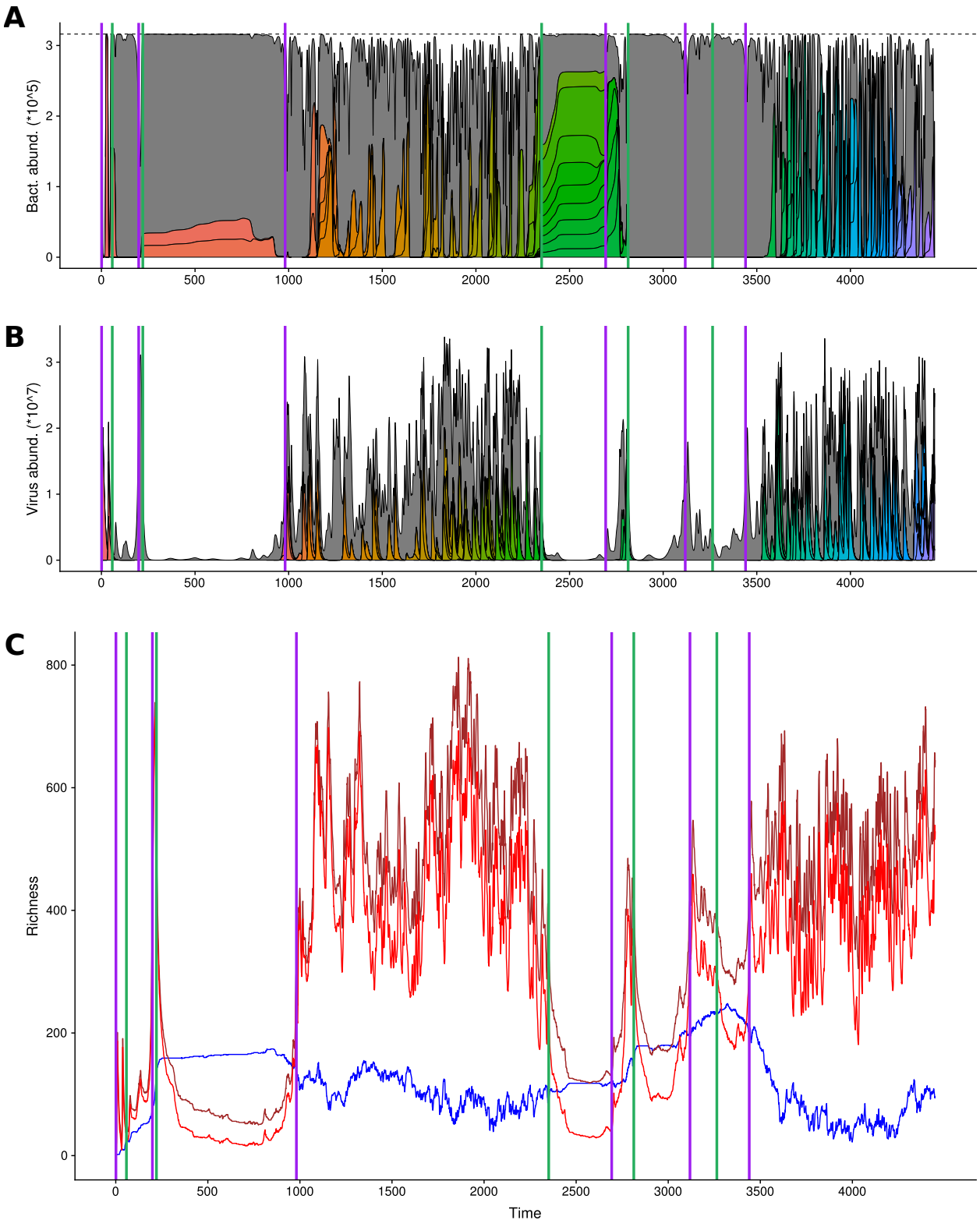


Fig. S14 Abundance and richness of hosts and viruses in high mutation rates. **(A)** Abundance of hosts. Dashed horizontal line indicates carrying capacity. **(B)** Abundance of viruses. **(C)** Richness of viruses (red), hosts (blue) and spacers (brown). In (A) and (B), the 100 most abundant strains (separately for viruses and hosts) are colored, and the rest are in gray.

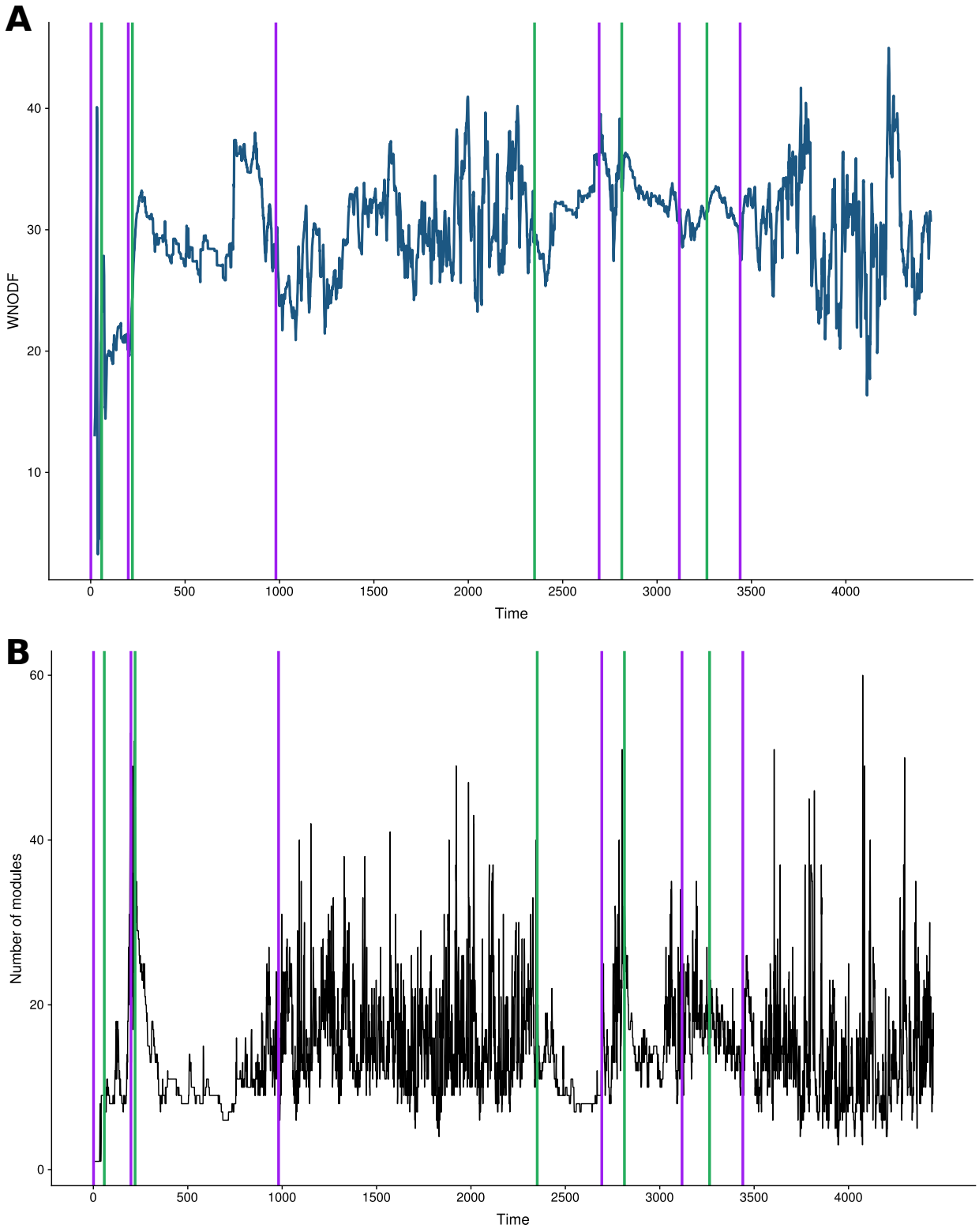


Fig. S15 Structure of immunity and infection networks in accelerated dynamics. **(A)** Nestedness of the immunity network. **(B)** Number of modules in the infection network. The fluctuations in nestedness and the number of modules during HCRs reflects the constant generation of viral and host strains.

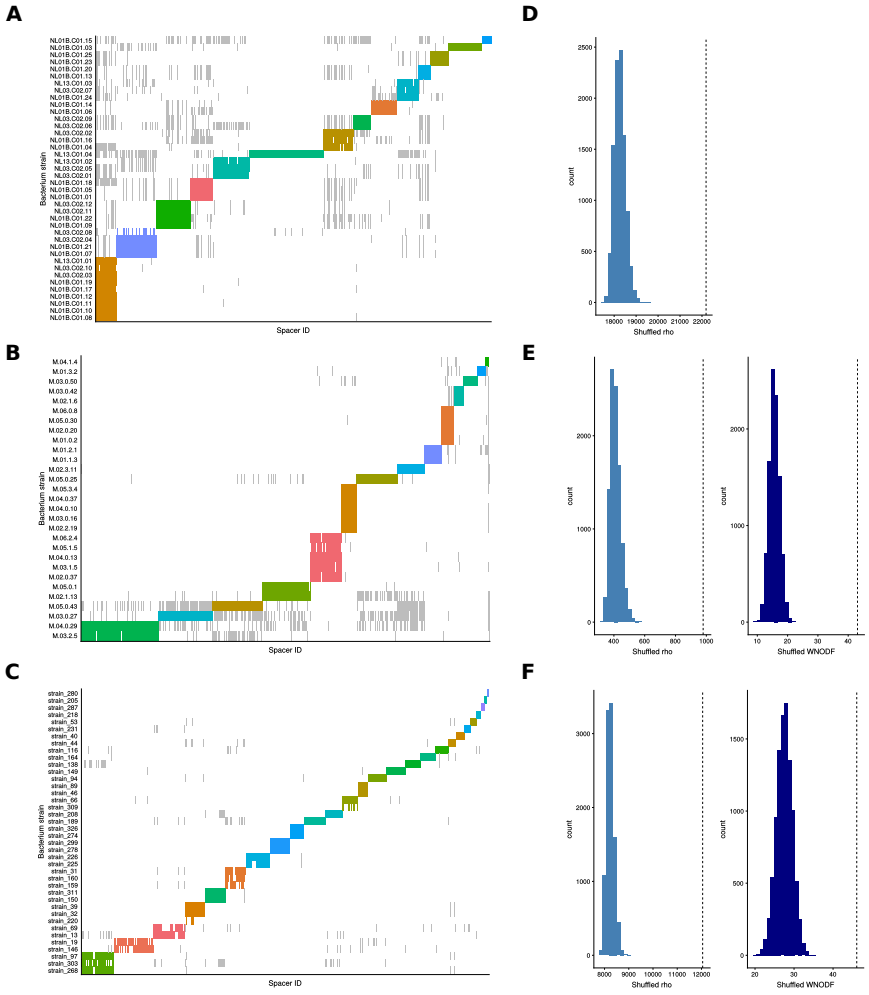


Fig. S16 Additional information on empirical data analysis. Each row represents a different data set: *Sulfolobus islandicus* hosts compared to contemporary lytic SIRV viruses (Top). *S. islandicus* hosts compared to contemporary chronic SSV viruses from the Mutnovsky Volcano in Russia, 2010 (middle). *Pseudomonas aeruginosa* hosts from Copenhagen compared to temperate mu-like viruses (bottom). Panels **(A)**, **(B)** and **(C)** are host-spacer networks in which interactions within host-spacer modules are colored (as in Fig. S8A). Panels **(C)**, **(D)** and **(E)** are distributions of either the leading eigenvalue (ρ), or WNODF of 10,000 immunity networks shuffled by randomly distributing interactions. Value of the observed ρ or WNODF is depicted with a vertical dashed line. We did not calculate WNODF in the Yellowstone data set because this index cannot handle networks which are completely full (i.e., with a density of 1).

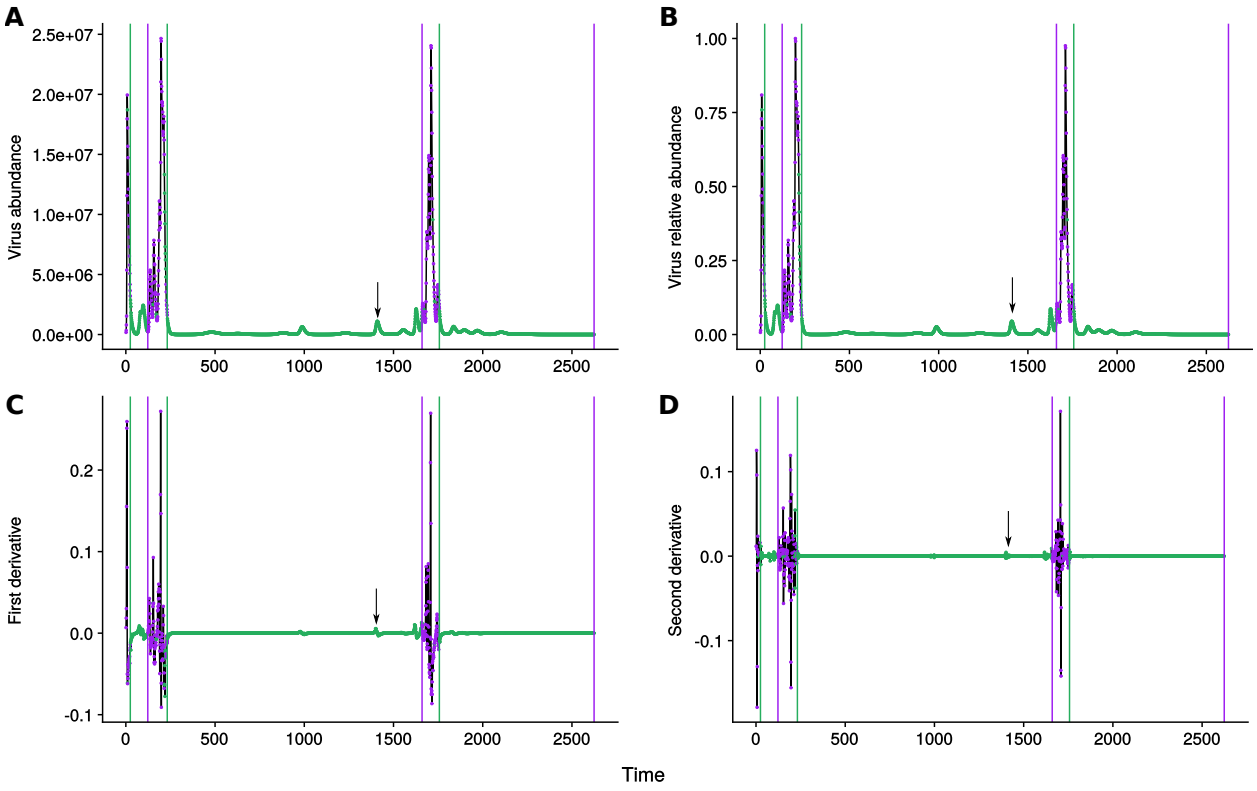


Fig. S17 Regime definition. Each point in the virus abundance time series in panel **(A)** is first converted to relative abundance (panel **(B)**) and then classified into a HCR (green) or VDR (purple). This classification is based on the second derivative (panel **(D)**) (for comparison we also show the first derivative in panel **(C)**). Momentary virus growth periods (marked with an arrow) are not classified as VDR. The final classification is shown using vertical lines. HCRs start with a purple line and VDRs with a green line.