# Abundance measurements reveal the balance between lysis and lysogeny in the human gut microbiome

Jaime Lopez[1,2], Saria McKeithen-Mead[1,3], Handuo Shi[3], Taylor H. Nguyen[1], Kerwyn Casey Huang[1,3,4,†], Benjamin H. Good[2,4,5,†]

[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

[2]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA

[3]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA

[4]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

[5]Department of Biology, Stanford University School of Medicine, Stanford, CA 94305, USA

[†]Correspondence: kchuang@stanford.edu (K.C.H.) and bhgood@stanford.edu (B.H.G.).

**Abstract**

The human gut contains diverse communities of bacteriophage, whose interactions with the broader microbiome and potential roles in human health are only beginning to be uncovered. Here, we combine multiple types of data to quantitatively estimate gut phage population dynamics and lifestyle characteristics in human subjects. Unifying results from previous studies, we show that an average human gut contains a low ratio of phage particles to bacterial cells (~1:100), but a much larger ratio of phage genomes to bacterial genomes (~4:1), implying that most gut phage are effectively temperate (e.g., integrated prophage, phage-plasmids, etc.). By integrating imaging and sequencing data with a generalized model of temperate phage dynamics, we estimate that phage induction and lysis occurs at a low average rate (~0.001-0.01 per bacterium per day), imposing only a modest fitness burden on their bacterial hosts. Consistent with these estimates, we find that the phage composition of a diverse synthetic community in gnotobiotic mice can be quantitatively predicted from bacterial abundances alone, while still exhibiting phage diversity comparable to native human microbiomes. These results provide a foundation for interpreting existing and future studies on links between the gut virome and human health.

## Introduction

The human gut harbors a complex community of bacteria, viruses, and microbial eukaryotes that plays important roles in human health (1–3). Previous studies have largely focused on the bacterial portion of this community, but in recent years the bacteriophage ("phage") that infect these bacteria have started to draw more attention. Advances in DNA sequencing and anaerobic culturing have led to extensive databases of gut phage genomes (4,5), as well as increasing numbers of phage isolates that can be propagated in the lab for mechanistic investigation (6,7). Phage can influence the microbiome in multiple ways. They can directly kill their bacterial hosts through lytic infection (7,8) or by inducing lysis from a temperate state (8,9). Temperate phage can also serve as important vectors of horizontal gene transfer (10), carrying cargo genes that enhance the metabolic or defense capabilities of their bacterial hosts (11). These interactions with gut bacterial ecology and evolution have been hypothesized to impact human health. Cohort studies have revealed numerous associations between the composition of the gut virome and various health-related states, including cancer treatment efficacy (2) and lifespan (12). Transplants of sterile phage-containing fecal filtrates from healthy donors can help resolve and protect against infections (13,14) or exacerbate disease phenotypes (15), phenomena potentially mediated by bacteria-phage interactions. Phage particles can also interact directly with the human immune system (16). These results suggest that quantitative characterization of gut phage communities is likely critical for understanding and engineering the gut microbiome.

However, while the individual members of the gut virome are becoming increasingly well characterized, much less is known about their ecological dynamics within a typical human and the effects they exert on the surrounding microbial community. In marine ecosystems, phage particles outnumber bacteria ~10:1 (17,18) and are estimated to kill ~20% of the bacterial population each day (18). Such high rates of lysis generate strong selection pressures for both bacteria and phage, leading to antagonistic co-evolution (19) and “kill-the-winner” dynamics of strain turnover (20–22). By contrast, estimates of the virus-to-microbe ratio (VMR) in the human gut vary widely across studies, from greater than 1:1 (23,24) to less than 1:10 (25,26). Furthermore, while some studies have suggested that the gut microbiome is dominated by temperate phage (8,27), little is known about rates of induction and lysis, and other studies have suggested that evasion of phage-mediated lysis is a major driver of bacterial evolution within human hosts (11,28). Inferring these ecological parameters is particularly challenging in the complex setting of the human gut, as it requires linking existing measurement approaches with quantitative models of phage population dynamics.

Here, we address this gap by combining mathematical modeling and publicly available data to obtain quantitative baseline estimates of gut viral populations sizes and induction rates in human hosts. Using a meta-analysis of gut viral population size measurements, we show that existing data can be unified into a coherent quantitative picture in which the gut microbiome has more phage genomes than bacterial genomes, but many fewer phage particles than bacterial cells. This suggests that the gut is dominated by temperate phage (here, “temperate” refers to all phage that reproduce

with their host genome, including classical lysogens, phage-plasmids, etc.). Building on this quantification, we develop a modeling framework that enables inference of mean gut phage induction rates from microscopy and metagenomic measurements. Our findings suggest that, in typical adults, gut phage are rarely induced and place a low mean fitness burden on their bacterial hosts. Finally, we show that similar ecological dynamics arise in gnotobiotic mice colonized with a synthetic community of >100 human gut bacterial isolates. As expected for a microbiome dominated by temperate phage, we find that the virome composition of these mice can be quantitatively predicted from the bacterial composition alone, while still exhibiting viral diversity comparable to a typical human stool microbiome. These results suggest that existing methods for predicting gut phage lifestyles drastically overestimate the fraction of lytic phage, indicating that many gut phage contain yet uncharacterized host-association genes

## Results

### The typical human gut microbiome contains fewer phage particles than bacterial cells

To determine the range of phage population sizes and virus-to-microbe ratios (VMRs) in the gut, we compiled measurements across multiple methodologies and studies (**Table 1**). Although the VMRs, initially appeared to vary across studies, we found that they could be unified into a coherent quantitative picture by employing a consistent calculation approach that accounts for key differences among existing phage quantification techniques (**Fig. 1A**, **Methods**).

Many approaches for estimating phage abundance involve the isolation of virus-like particles (VLPs) as representatives of the free phage particle population within stool samples. In the most common method, isolated VLPs are enumerated via epifluorescence microscopy using a DNA-binding dye (23,29). These microscopy-based methods estimate the concentration of free phage particles in the stool, although their accuracy is constrained by VLP isolation efficiency (23,30) and the presence of non-phage particles (31). Aggregating VLP enumeration data from multiple studies and age groups (**Fig. 1B**), we found that, apart from newborns in which VLP densities are often below the limit of detection (23), stool VLP density stabilizes after >1 month of age to a population average of $\sim 2\times10^{9}$ VLPs/g stool, which is maintained throughout adulthood. Combining these data with existing estimates of the density of microbial cells in stool in humans older than >1 month ($\sim 10^{11}$ cells/g stool; (32)) yields a VLP-to-microbe ratio

115 ~$10^{-2}$. This estimate is three orders of magnitude lower than the VMRs commonly
116 reported for surface seawater systems (20), hinting qualitatively different viral ecological
117 dynamics that we will explore in more detail below. We also find that the inter-individual
118 variation in VLP counts is similar in magnitude to that of bacterial counts, with post-
119 infancy VLP measurements exhibiting a population coefficient of variation (CV) of 0.61
120 versus 0.46 for bacterial counts (32). This suggests that the total gut phage population
121 does not undergo dramatic abundance fluctuations across hosts.
122
123 An alternative form of phage particle measurement utilizes a spike-in approach,
124 involving shotgun sequencing of amplified DNA from the VLP pool after adding a known
125 amount of a non-gut reference phage (25). The fraction of sequencing reads mapping to
126 reference versus non-reference phage can then be used to obtain an independent
127 estimate of absolute phage particle density (**Fig. 1A**). A recent application of this
128 approach to longitudinal samples from ~10 healthy adults yielded a ~5-fold higher
129 concentration than microscopy-based studies (mean of ~$1 \times 10^{10}$ VLPs/g, inter-
130 individual CV of 0.9, **Fig. 1B**). These data also provided an estimate of the temporal
131 variation, with monthly VLP estimates within individuals having a mean CV of 0.78,
132 suggesting that the total phage load in individual hosts does not undergo dramatic
133 fluctuations. The differences between this study and microscopy-based quantifications
134 may be due to underestimation of viral counts by imaging-based approaches relative to
135 sequencing/qPCR-based approaches (33). However, we found that the two
136 measurements are largely consistent after exclusion of reads mapping to the
137 *Microviridae* family of phage (**Fig. 1B**), which are thought to be disproportionately

enriched by the multiple displacement amplification (MDA) protocol commonly employed in VLP sequencing (34). Regardless, even the larger VMR estimates resulting from sequencing-based approaches including *Microviridae* reads (~$10^{-1}$) are still far lower than the ~10:1 ratios reported for surface seawater (20).

**The typical human gut microbiome contains more phage genomes than phage particles**

A third class of quantification methods estimates VMRs directly from metagenomic sequencing of stool samples (24). This approach has been enabled by the recent assembly of large databases of viral and prokaryotic genomes from the human gut (4,5,35), from which >98% of reads from a typical stool sample can be classified using taxonomic profilers like Phanta (24). By normalizing the ratio of phage to bacterial reads with corresponding phage and bacterial genome lengths, one can obtain an independent estimate of the VMR. Applying this approach to a collection of 255 previously sequenced adult gut metagenomes yields an average VMR of ~4:1 (inter-individual CV = 0.38), corresponding to an absolute density of ~$4 \times 10^{11}$ phage genomes/g after multiplying by the typical bacterial density in Ref. (32). These values are two orders of magnitude higher than the VLP-based estimates above.

The discrepancy between these estimates can be reconciled by the observation that bulk stool metagenomics measures the total number of viral genomes in a stool sample, including those encapsulated in bacterial cells (e.g., as prophage), while VLP-based methods only measure free viral particles. Hence, it is useful to distinguish between two

161 distinct abundance measures: the genomic VMR (gVMR), estimated from bulk stool
162 sequencing, and the particle VMR (pVMR), estimated from VLP-based approaches. The
163 two measures are roughly equivalent in environments like surface seawater where the
164 pVMR is much larger than one (and therefore particles dominate the gVMR). However,
165 they can dramatically diverge in ecosystems like the gut where the pVMR is much less
166 than one. In this case, the ~100-fold difference between the number of phage particles
167 and phage genomes in the gut suggests that the vast majority of gut phage are
168 temperate or otherwise attached to their bacterial hosts. These temperate phage may
169 not be traditional prophage that are integrated into their host's genome; many gut phage
170 do not contain recognizable lysogeny-associated genes and thus may utilize other host-
171 associated lifestyles, such as those of phage-plasmids (36).
172
173 Consistent with this temperate-dominated picture, we found that the ratio of phage
174 genomes to phage particles is also large for many individual viral species. While MDA
175 amplification biases make precise quantification difficult (34), comparisons between
176 matched VLP and bulk sequencing in infants and adults revealed three broad classes of
177 behavior. Some viral species are observed in the bulk metagenome but not in the
178 associated VLP pool (**Fig. 1C, bottom**). These species account for about half of the
179 total phage abundance in bulk stool samples (**Fig. 1D**) and could represent cryptic (37)
180 or inactive (27) prophage, as well as phage that are poorly amplified by MDA (34). A
181 second set of viral species are observed in VLP sequencing but not in the associated
182 bulk metagenome (**Fig. 1C, left**). These species account for about half of the total
183 phage abundance in the VLP pool, and could reflect both MDA amplification biases (34)

as well as phage that are poorly captured by bulk metagenomics. Finally, a third class of viral species is present in both the VLP and bulk metagenomes. Their abundances are broadly consistent with the aggregate particle to genome ratio above, with ~80% of phage-sample pairs having a ratio below 1:100 (**Fig. 1C, Fig. S1**), and contain a mixture of phage species classified as temperate and purely lytic (**Fig. S2**). These species account for the other half of the VLP and bulk phage populations (**Fig. 1D**). These data suggest that even with limitations imposed by MDA biases, a large fraction of gut phage exhibit a generalized form of temperance, with a small population of viral particles maintained by a much larger number of host-associated viral genomes.

**The phage particle to phage genome ratio provides a lower bound on the rate of phage induction**

While population sizes and lifestyles are important aspects of gut ecology, they provide only a static picture of the gut virome and its potential interactions with gut bacteria. To interpret these data and estimate the rates of phage induction and lysis in the human gut, we utilized mechanistic models of phage population dynamics over time (38–40).

We begin by considering a simplified model of phage ecology, which approximates each host gut as a well-mixed ecosystem with mass-action kinetics (**Fig. 2A**, **Methods**). For a single pair of bacteria and phage, this model can be described by a system of three differential equations for the concentrations of uninfected susceptible bacteria ($S$), infected bacteria or prophage ($P$), and free phage particles ($V$):

$$\frac{dS}{dt} = \underbrace{\mu_S(S,P)\cdot S}_{\text{Growth}} - \underbrace{\kappa SV}_{\text{Infection}} - \underbrace{\delta S}_{\text{Dilution}}, \quad (1)$$

$$\frac{dP}{dt} = \underbrace{\mu_P(S,P)\cdot P}_{\text{Growth}} + \underbrace{\kappa f_L SV}_{\text{Infection}} - \underbrace{\xi P}_{\text{Induction}} - \underbrace{\delta P}_{\text{Dilution}}, \tag{2}$$

$$\frac{dV}{dt} = \underbrace{B\xi P}_{\text{Induction}} + \underbrace{B(1-f_L)\kappa SV}_{\text{Direct lysis}} - \underbrace{\kappa SV}_{\text{Infection}} - \underbrace{r_e \kappa PV}_{\text{Failed infection}} - \underbrace{\delta V}_{\text{Dilution}}. \tag{3}$$

Here, $\mu_S(S,P)$ and $\mu_P(S,P)$ are the growth rates of susceptible and infected bacteria, $\delta$ is the overall dilution rate, and $\kappa$ is the infection rate. We assume that a fraction $f_L$ of infections result in the formation of prophage, while the remaining infections result in direct lysis of the cell with burst size $B$. Phage particles are also produced by induction of prophage at rate $\xi$. We assume that infected cells are immune to further infection by phage particles, with these failed infections resulting in loss of the infecting phage particle (e.g., to superinfection inhibition mechanisms (41)) with rate $r_e\kappa$. We also consider extensions of this model that account for dead cells, dead phage, and actively lysing cells (**Methods**).

Depending on the induction rate and lysogeny fraction, this minimal model can interpolate between a classic lytic lifestyle and a purely temperate phase in which phage primarily reproduce via lysogeny (38,42). In the latter case, the spontaneous induction of prophage can maintain a small population of phage particles (pVMR $\ll$ 1) while the ratio of phage to microbial genomes (gVMR) remains near one, similar to the distributions seen in **Fig. 1**. This prophage-dominated regime emerges for a broad range of model parameters, particularly when the cost of prophage carriage is low (**Methods**).

We can extend this basic model to larger numbers of phage and bacterial species, except that we must now allow for multiple prophage states in each bacterium (representing simultaneous infection by different combinations of prophage). By summing **Eq. 3** over phage species and integrating over time, one can derive an approximate equation relating the aggregate prophage and phage particle concentrations:

$$0 \approx \bar{\xi}^* \bar{B}^* \left(\frac{P^*}{V^*}\right) - \bar{\psi}^* - \delta^*, \tag{4}$$

where $\bar{x}$ and $x^*$ denote community- and time-weighted averages of the quantity $x$, respectively (**Methods**), and $\psi$ is the residual phage adsorption rate (e.g., due to failed infections of lysogens). **Eq. 4** assumes that over sufficiently long timescales, the fluxes controlling phage population sizes within an individual (induction, degradation, infection, etc.) are approximately balanced, even though day-to-day fluctuations could still be substantial (we discuss further details of our calculation assumptions in **Methods**). Based on the stability and moderate variance of the distribution of phage population densities (**Fig. 1B**), this assumption appears to hold in healthy humans >1 month of age.

Rearranging **Eq. 4** yields a relation for the average induction rate as a function of the other key model parameters:

$$\bar{\xi}^* = \left(\frac{1}{\bar{B}^*}\right)\left(\frac{V^*}{P^*}\right)(\bar{\psi}^* + \delta^*). \tag{5}$$

Consistent with intuition, **Eq. 5** predicts that the average induction rate is linearly proportional to the phage particle to genome ratio, $V^*/P^*$. It also increases with the

combined rate of particle removal (dilution rate $\delta^*$ and adsorption rate $\bar{\psi}^*$, **Fig. 2C**), and decreases with average burst size $\bar{B}^*$, since a smaller number of induction events are required to maintain the same density of phage particles. By combining **Eq. 5** with order-of-magnitude estimates of the other parameters, we can estimate the underlying induction rate. The ratio of phage particle and phage genome densities can be estimated from the population distribution in **Fig. 1** as $(V^*/P^*) \approx 10^{-2}$. The mean dilution rate $\delta^*$ is determined by the inverse of gut transit time, which can vary across humans but is approximately 1 day$^{-1}$ (43). The burst size $B$ can vary substantially across phage, with the model *Escherichia coli* phage λ having a burst size $B \approx 100$ (44), ΦCrAss001 having $B \approx 2.5$ (45), and crAssBcn isolates having $B \approx 50$ (46). Thus, we estimate the order of magnitude of $\bar{B}^* \approx 10$. The residual adsorption rate $\bar{\psi}^*$ is more difficult to estimate due to our limited understanding of infection rates and host ranges of gut phage *in vivo*. Nonetheless, setting this quantity to zero yields a lower bound on the induction rate,

$$\bar{\xi}^* \geq \left(\frac{1}{\bar{B}^*}\right)\left(\frac{V^*}{P^*}\right)\delta^* \approx 10^{-3}\ \mathrm{day}^{-1}. \tag{6}$$

This lower bound increases to $\bar{\xi}^* \geq 10^{-2}$ day$^{-1}$ when using the estimate of $V^*/P^* = 10^{-1}$ from VLP spike-in sequencing but is still two orders of magnitude lower than the dilution rate $\delta^*$. The bound is also relatively tight, with substantial deviations only possible if $\bar{\psi}^*$ is larger than $\delta^*$ (**Fig. 2C**). These results suggest that the gut phage particle pool, despite having ~1,000-fold higher density than the highly lytic surface seawater virome (47), can be maintained by a very low rate of induction per infected bacterium.

**The relative coverage of integrated prophage provides an upper bound on the rate of phage induction**

Prophage induction can be identified from metagenomic data by comparing the relative coverage of an integrated prophage genome and nearby regions of its bacterial host genome (48). Such methods have thus far been used to make binary determinations of prophage induction for individual phage-bacteria pairs (48), but they also provide information about the underlying induction rate. To extract this information, we use a generalized version of our model in **Eq. 1** to explicitly model activated prophage, representing the state between the start of induction and lysis (**Methods**). These activated lysogens contain $B_a \approx B$ additional copies of the phage genome that correspond to nascent phage particles. The relative coverage $R$ of the prophage and host genomes in metagenomic sequencing data is given by

$$R = 1 + B_a f_a, \tag{7}$$

where $f_a$ is the fraction of currently activated cells. If activated cells are produced from lysogens at rate $\xi$ and have a mean lysis time of $1/\gamma$, then the ratio of activated to non-activated cells will approach a steady-state value of $\sim\xi/(\gamma + \delta)$ (**Methods**). This result can be combined with **Eq. 7** to relate the mean induction rate to the mean relative coverage:

$$\xi^* \leq \frac{(\gamma^* + \delta^*)(R^* - 1)}{(B^* - 1) - (R^* - 1)}. \tag{8}$$

This estimate is robust to the confounding impact of dead cells and viruses contributing to $R$ (**Methods**). The relationship between the induction rate and the relative coverage critically depends on the characteristic lysis time of activated infected cell, $1/\gamma$. Prior studies suggest that lysis time scales with the bacterial host division time (49–51). This

scaling is consistent with estimates of phage burst energetics: a phage burst consumes a large fraction of the host bacterial energy budget (52), implying that production of a phage burst is limited by similar factors as host replication. The mean growth rate roughly matches the dilution rate $\delta^*$ in the parameter regime implied by **Fig. 2C**. Thus, for the following calculations we assume that $\gamma^*$ is of the same order of magnitude as $\delta^*$.

**Eq. 8** applies to the subset of phage that are detected within a contig of an assembled bacterial genome. While it in principle enables measurements of arbitrarily low induction rates (**Fig. 2D**), but in practice it is difficult to distinguish small values of $R$ from 1 due to noise and biases in sequencing. Indeed, in a previously published analysis of positive and negative controls, induction of individual prophage could only be reliably determined for $R > 2$, and the median number of such events across fecal metagenomes was zero (49). To establish a tighter upper bound of the induction rate, we take the "clipped" average of $R$ (i.e., setting values of $R < 1$ to $1$) across all adult samples analyzed in (48), yielding $R^* - 1 \approx 10^{-1}$. Substituting this value into **Eq. 8** with $B^* = 10$, and $\gamma^* \approx \delta^* = 1\ \mathrm{day}^{-1}$ yields an upper bound on the induction rate of

$$\xi^* \leq 10^{-2}\ \mathrm{day}^{-1}. \tag{9}$$

The clipped mean is larger in infants ($R^* - 1 \approx 0.4$), even after excluding infants exposed to antibiotics (**Table 2**), suggesting that the induction rate may be higher in infants. Combined with our other estimates, we can thus bound the average induction rate within the range $10^{-3} - 10^{-2}\ \mathrm{day}^{-1}$ for adults (**Fig. 2B**), with a somewhat higher upper bound for infants**.** Importantly, both estimates are substantially lower than the rate

of microbial growth and dilution from the gut, suggesting that gut phage impose a low mean fitness burden on their bacterial hosts.

**Similar virome properties arise in gnotobiotic mice colonized with a diverse synthetic community of human gut bacterial isolates**

We next examined the implications of our results for a synthetic gut community designed to mimic the complexity of a native human microbiome (53). This synthetic community is composed of 119 bacterial isolates from 48 prevalent genera and stably colonizes gnotobiotic mice for ≥2 months. We reasoned that the virome of hCom2 would be exclusively composed of temperate phage (at least initially), since it was constructed from axenic bacterial cultures (54). Our finding that the human gut is dominated by rarely inducing temperate phage makes two major predictions about the properties of the hCom2 virome and its relation to the human data above.

First, if the induction rates in hCom2 are as low as our model predicts (**Fig. 2**), we expect its viral composition in bulk metagenomic sequencing to be entirely predictable from the abundances of its bacterial members (since $R \approx 1$). It is usually difficult to test such a prediction in natural communities like the human gut, in which only a subset of phage can be directly linked to their bacterial hosts (48). Synthetic communities like hCom2 provide a unique opportunity to test this prediction, since their initial phage-bacteria associations can be inferred from the sequenced bacterial isolate genomes. To carry out this test, we generated *in silico* hCom2 metagenomes based on data from a recent experiment (53) using the sequenced genome of each bacterial strain in

341 proportion to their measured abundance in each sample (**Methods**). By construction,
342 these *in silico* datasets only contain phage sequences that were present within the
343 original bacterial genomes. We then compared the taxonomic composition of these *in*
344 *silico* datasets with their corresponding mouse metagenomes using the same pipeline
345 as above (**Fig. 3A,B**, **Methods**).
346
347 Consistent with previous observations in a smaller 15-member community (8), we found
348 that the abundances of individual phage species were highly correlated across the *in*
349 *silico* and *in vivo* datasets, with the representative sample in **Fig. 3A,B** having a
350 Spearman correlation of $\rho = 0.9$ for mutually detected phage, compared to $\rho = 0.97$ for
351 bacteria (as expected by construction). Similar results were obtained for other
352 compositional similarity metrics, like the Jaccard index or the total abundance of shared
353 species (**Fig. S3**). The similarity between the *in vivo* and *in silico* metagenomes was
354 maintained over time, and even after challenge with an undefined fecal sample (**Fig.**
355 **3C,D, Fig. S3**). These strong correlations confirm that the hCom2 virome is dominated
356 by temperate phage, and that the induction rates are consistent with our inferences from
357 the human data above.
358
359 A second – and much stronger – prediction of our prophage-dominated human gut
360 model is that the hCom2 stool virome should qualitatively resemble the stool virome of a
361 typical human. We tested this prediction by comparing the taxonomic composition of
362 hCom2-colonized mouse fecal samples (53) with that of a cohort of 245 healthy human
363 stool metagenomes (55). We reasoned that if hCom2, a synthetic community composed

364 of axenic bacterial cultures, was missing a large portion of the normal gut virome, then
365 feces from hCom2-colonized mice would have substantially lower virome diversity than
366 a typical human stool sample. However, we found that hCom2-colonized mouse feces
367 exhibited similar phage Shannon diversity as human stool samples, with the hCom2
368 samples falling between the 13$^{th}$ and 53$^{rd}$ percentiles of the observed human distribution
369 (**Fig. 4A**). We obtained a similar correspondence between hCom2 and human stool
370 using a metric of species richness (**Fig. 4B**, **Fig. S4**), as well as the overall ratio of
371 phage-to-bacterial genomes (**Fig. 4C**, **Fig. S4**). This similarity between hCom2 and
372 human stool viromes also holds at finer taxonomic levels, with 16 of the 20 most
373 prevalent phage genera within the human cohort found at >0.1% abundance in hCom2
374 samples (**Methods**). Thus, consistent with our estimates above, we find that human-like
375 levels of viral diversity can be achieved by a synthetic community of exclusively
376 prophage.
377
378 The striking similarities between the viromes of hCom2-colonized mouse feces and
379 human stool can shed light on other coarse-grained features of the human gut virome.
380 For example, computational tools have been developed to predict the lifestyles of phage
381 species from their genomes (56,57), enabling estimation of the ratio of virulent phage
382 (those that cannot stably replicate within their hosts) to temperate phage (24,57). We
383 used predictions from widely used tools to estimate the virulent to temperate ratio (VTR)
384 in hCom2-colonized mouse fecal samples (**Methods**). Since hCom2 was constructed
385 entirely from axenic bacterial cultures, it might be expected to provide a negative control
386 with a VTR of ~0. However, hCom2-colonized mouse feces metagenomes yielded a

387 VTR of ~0.5, similar to the typical values observed in human stool samples (**Fig. 4D**,
388 **Fig. S4, Fig. S5**). This result suggests that existing methods of phage lifestyle prediction
389 methods underestimate the number of phage that are capable of stable replication
390 within their bacterial hosts, consistent with previous observations from human stool
391 samples (**Fig. S2**) (25).

392 **Discussion**

393 Our results complement existing surveys of gut phage diversity (4,23,25,58) by
394 providing a quantitative assessment of phage population dynamics in typical human
395 hosts. Our updated estimates of the virus-to-microbe ratio show that the small number
396 of gut phage particles (pVMR ~ $10^{-2}$–$10^{-1}$) is accompanied by a much larger number of
397 phage genomes (gVMR ~ 4), implying that the vast majority of gut phage genomes are
398 replicating within their bacterial hosts. These results support the emerging view that
399 temperate phage lifestyles play a dominant role in the human gut (8,25,36,59,60), even
400 if they do not contain recognizable integrase genes (e.g. owing to utilization of novel
401 integrases or having non-integrative lifestyles) (25,36) (**Fig. 4D**). Our quantitative
402 framework extends this picture by providing new insights into the corresponding phage
403 induction rates. By integrating imaging and sequencing measurements with a
404 generalized model of temperate phage dynamics, we estimated that the average
405 induction rate in adults lies in the relatively low range of $10^{-3}$–$10^{-2}$ per bacterium per day,
406 imposing only a modest fitness burden on gut bacteria.

407

408 These results starkly contrast with well-studied examples like surface seawater, which
409 possesses a larger ratio of phage particles (pVMR ≈ gVMR~10) and a higher average
410 lysis rate (20). The reasons for this difference remain uncertain, but they may partially
411 stem from the distinct physical structures of the two ecosystems. Previous theoretical
412 and experimental studies have shown that increased spatial structure can select for
413 lower virulence and increased lysogeny (61,62), consistent with the dominance of
414 temperate phage in the more spatially structured gut ecosystem, although more work is

415 needed to quantify the strength of this effect in the gut. Regardless, our results establish
416 baseline expectations for the co-variation between phage and bacterial abundances
417 within the microbiome of a typical human. They imply that tight associations between
418 the bacterial and phage communities may not be driven by active predator-prey
419 interactions, but may instead be a simple consequence of their synchronized replication
420 within the same cells, in line with the “piggyback-the-winner” model (60,63,64). This
421 latter scenario suggests that phage may impact the gut microbiome primarily by acting
422 as genetic cargo, altering the behavior of their bacterial hosts in certain conditions
423 (11,65).
424
425 These results have substantial implications for future studies of the gut virome’s role in
426 human health. Many studies have sought to identify biomarkers and characterize
427 possible mechanistic links between gut virome composition and health states such as
428 lifespan (12), cancer treatment response (2), diabetes (66), metabolic syndrome (67),
429 and alcoholic hepatitis (68). Importantly, our results highlight confounding factors that
430 complicate such analyses of virome-health associations, particularly for studies focused
431 on bulk stool sequencing in which a high abundance of prophage will likely result in
432 strong statistical links between phage and bacterial composition if the number of VLPs
433 is low. In studies focused on VLP sequencing, similar correlations could emerge if the
434 VLP pool is largely a product of relatively uniformly induced prophage, a scenario
435 hypothesized by prior work (23) and supported by the substantial overlap between bulk
436 and VLP virome compositions (**Fig. 1C**) (24). These results suggest that methods

437 similar to phylogenetic regression (69) may be useful for dealing with these confounding
438 factors.
439
440 The quantities in our modeling framework represent averages over time, space, and
441 hosts that may mask important behaviors that are transient or localized to a particular
442 host microniche. In particular, averages over longer timescales may not capture shorter-
443 term variation in induction rates. While the VLP population appears to be broadly stable
444 over long timescales, the monthly CV of VLP abundances within individuals has a mean
445 of 0.77 (25). One explanation for such variation is phage induction driven by
446 environmental changes within the host, a hypothesis consistent with prior studies
447 showing increased lytic activity in response to perturbations such as bacterial/phage
448 invasion (8,11), inflammation (70), or exposure to certain dietary or pharmaceutical
449 compounds (71,72). In addition to such temporal and host variation, phage population
450 sizes and induction rates may also vary spatially within an individual gut (60,73–75), as
451 environmental conditions and bacterial densities change substantially along the
452 gastrointestinal tract (76). It remains possible that the low pVMRs observed in stool
453 could be produced by a very high induction rate in a smaller population of bacteria in the
454 proximal colon or small intestine. In the future, spatial variation could be investigated
455 using recently developed methods for spatially resolved sampling of the microbiome
456 (77) to measure population sizes and prophage copy numbers across the intestines.
457 Our modeling framework can be readily applied to such data to estimate local virome
458 induction rates.
459

460 Variation in lifestyle characteristics across phage is also expected, with some phage
461 effectively existing as mobile genetic elements that rarely lyse their host and others
462 being primarily lytic. While our current estimates average over multiple phage taxa, our
463 modeling framework can also be applied to measurements of individual phage species
464 to estimate species-specific properties. For example, if the particle-to-genome ratios of
465 an individual phage species can be more accurately measured (**Fig. 1C**), a species-
466 specific estimate of the induction rate can be obtained from **Eq. 5**. Applications of this
467 approach are currently limited by the known amplification biases of existing VLP
468 sequencing methods (34), but the adoption of sequencing protocols that do not involve
469 MDA (78) may enable such species-specific resolution in the future.
470
471 Beyond our modeling assumptions, there are also limitations in the phage quantification
472 methods used for experimental measurements. The process of VLP isolation may lead
473 to substantial loss of phage particles, particularly given the spatially structured nature of
474 stool and the potential for phage particles to adhere to large particulates. Additionally,
475 imaging-based quantification methods can both underestimate phage densities due to
476 loss of material during preparation (33), and overestimate due to the presence of cell
477 debris or other DNA-containing particles (31). Similarly, RNA phage cannot be
478 visualized using DNA-staining-based microscopy (23,29). Underestimation of phage
479 particle densities would imply a higher true pVMR, which would increase the
480 corresponding induction rate estimate from **Eq. 5.** Note, however, that for our pVMR
481 estimates to be comparable to that of surface seawater would require very large
482 differential loss rates (>99%), which could potentially be measured with appropriate

483 spike-ins. Our analysis framework can easily be applied to updated density estimates as
484 they become available.
485
486 Overall, our work motivates future experimental directions for the gut virome field. While
487 informative, our estimates of the mean induction rate still encompass 1-2 orders of
488 magnitude owing to limitations of current data. Given the noise intrinsic to metagenomic
489 sequencing, we expect that deeper bulk sequencing will have limited benefits for
490 estimating of the mean induction rate in the parameter regimes suggested by our
491 analysis. More accurate and direct estimation will likely be dependent on measurement
492 of rare induced cells. Single-cell bacterial sequencing (79) is a promising avenue to
493 achieve the needed detection power. Alternatively, measurement of *in vivo* phage
494 adsorption rates or the degradation rates of lysed cells (**Methods**) would enable
495 improved estimation of induction rates that we derived in **Fig. 2**. Our results also provide
496 guidance for the design of virome perturbation experiments, which should focus on
497 measuring increases in induction and horizontal gene transfer – a major avenue through
498 which prophage influence their hosts. Finally, the similarities between the estimated
499 VTR in hCom2-colonized mouse feces (**Fig. 4D**) and human stool metagenomes
500 highlights the current lack of knowledge regarding the genetic mechanisms enabling
501 bacterial host-association of gut phage. These results imply that many gut phage
502 currently computationally identified as virulent in fact contain unidentified and
503 uncharacterized host-association genes. This pool of genes represents a rich ground for
504 future phage molecular biology work.

**Acknowledgements**

The authors thank the Huang and Good labs, Ami Bhatt, Danica Schmidtke, Gabriel Birzu, Colin Hill, and Andrey Shkoporov for helpful discussions. The authors acknowledge support from NIH RM1 Award GM135102 and R01 AI147023 (to K.C.H.), NSF Awards EF-2125383 and IOS-2032985 (to K.C.H.), NIH R35 GM146949 (to B.H.G.), Alfred P. Sloan Foundation grant FG-2021-15708 (to B.H.G.), Human Frontier Science Program grant RGEC33/2023 (to B.H.G.), and a Friedrich Wilhelm Bessel Award from the Humboldt Foundation (to K.C.H.). B.H.G. and K.C.H. are Chan Zuckerberg Biohub Investigators. J.L. was supported by a Stanford PRISM Baker Fellowship. This work was also supported in part by the National Science Foundation under Grant PHYS-1066293 and the hospitality of the Aspen Center for Physics. We thank the Stanford Research Computing Center for use of computational resources on the Sherlock cluster.

## Figures

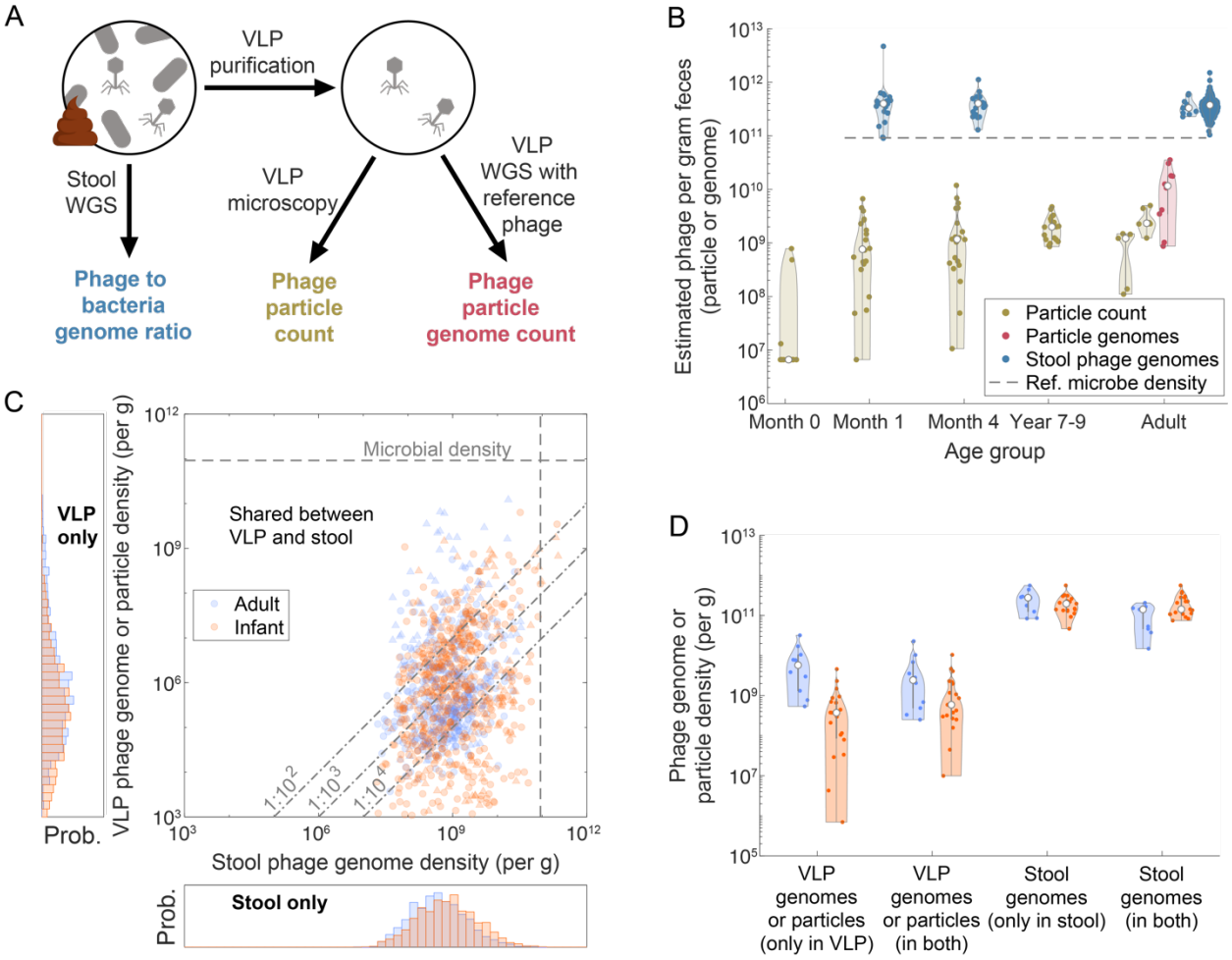



**Figure 1: Comparisons of absolute population densities suggest that the human gut virome is numerically dominated by prophage.** (A) Schematic overview of common gut viral population quantification methods. Metagenomic classification of bulk fecal samples yields phage-to-bacteria genome ratios, which can be combined with absolute bacterial densities to estimate the absolute density of phage genomes. Alternatively, virus-like particles (VLPs) can be extracted from the stool and quantified by epifluorescence microscopy or spike-in sequencing. (B) Gut virome population densities are approximately maintained across human life stages. Each violin plot represents quantification of one population using one measurement method in one

529 study (**Table 1**), with individual dots representing subjects. The gray line denotes the
530 average stool bacterial density reported in (32) ($0.92\times10^{11}$ bacteria/g stool). (C)
531 Species-level absolute abundance analysis of the overlap of phage communities found
532 using VLP- and stool-based quantification approaches. Data represent one healthy
533 adult population (25) (*n* = 10, VLP WGS absolute quantification) and one population of
534 4-month-old infants (23) (*n* = 19, VLP microscopy absolute quantification). Each point is
535 the absolute abundance of one phage species in a matched pair of VLP and stool
536 samples from one subject (**Methods**). Triangle markers denote species classified as
537 virulent and circle markers denote species classified as temperate (**Methods**).
538 Histograms show the distribution of absolute abundances of phage species found
539 exclusively in either the VLP or stool samples. (D) Relative distribution of phage
540 genomes or particles between VLP and stool WGS. Each violin plot represents the total
541 absolute abundances of phage genomes or particles found within only VLP samples,
542 only stool samples, or shared between VLP and stool. Individual points correspond to a
543 single subject. Underlying data are the same as in (C).

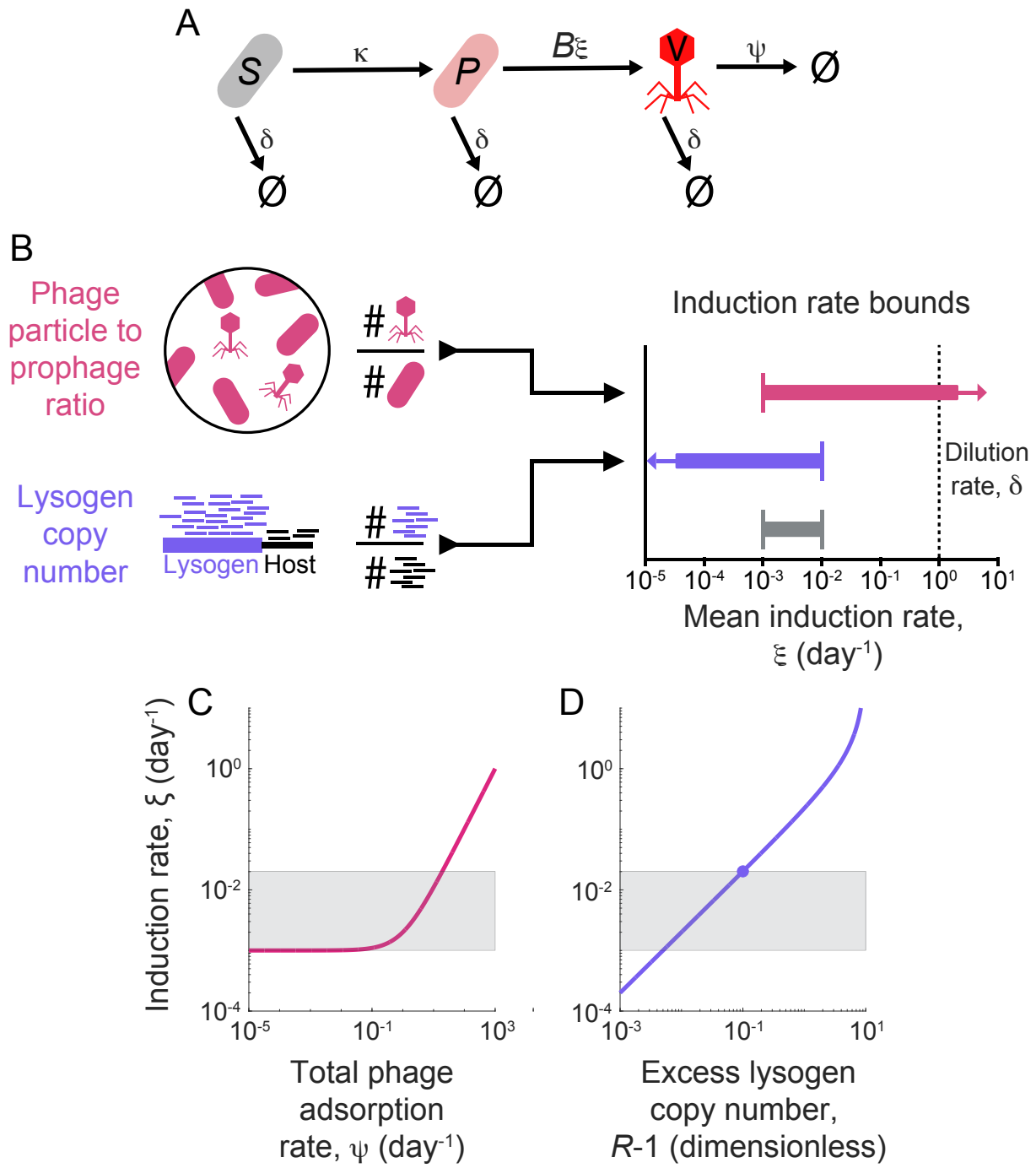



**Figure 2: Mathematical modeling of phage population dynamics enables estimation of the average phage induction rate in the human gut.** (A) Schematic of the minimal model of temperate phage dynamics represented by **Eq. 1-3**. Infection of susceptible bacteria produces prophage, which induce at rate $\xi$, lysing their host and

549 producing a burst of $B$ free phage particles. (B) Schematic representation of induction
550 rate estimates. We combine measurements of phage particle to genome ratio and
551 relative prophage copy number with the model in (A) to estimate upper and lower
552 bounds on the phage induction rate. Note that given the uncertainty in parameter
553 values, these estimates are only reported as approximate orders-of-magnitude, with the
554 combined bound illustrated in grey. (C) Estimated induction rate as a function of total
555 phage adsorption rate $\psi$ (i.e., all non-dilution phage particle removal mechanisms). The
556 solid line corresponds to **Eq. 5**, using the phage genome to particle ratio inferred from
557 **Fig. 1**. (D) Estimated induction rate as a function of the relative coverage of prophage,
558 $R$. The solid line corresponds to **Eq. 8** with $\gamma \approx \delta$. The solid circle is the mean relative
559 coverage in adults ($R - 1 \approx 10^{-1}$), using measurements from (48).

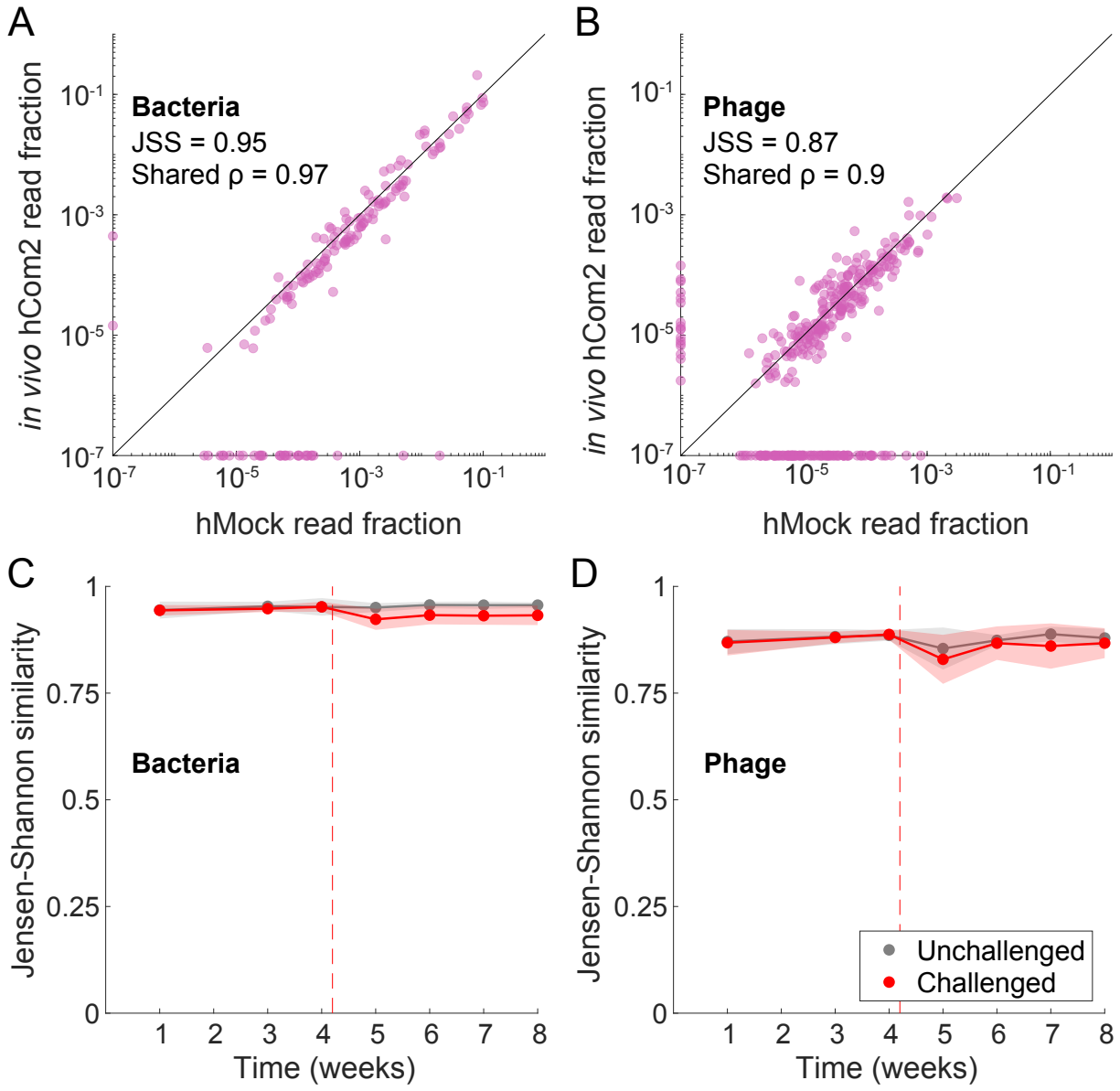



**Figure 3: Phage abundance dynamics in a diverse synthetic gut community can be predicted from bacterial abundances alone.** (A,B) Relative read abundances of bacterial (A) and phage species (B) in fecal samples from hCom2-colonized gnotobiotic mice, compared to *in silico* metagenomes ("hMock") generated from their corresponding bacterial genomes weighted according to the fecal bacterial microbiota composition (**Methods**). The example shown is for a single representative sample (mouse 3, week 1). JSS is the Jensen-Shannon similarity and shared $\rho$ denotes the Spearman

568 correlation computed from species observed in both samples. (C,D) Bacterial and
569 phage JSS between *in vivo* and *in silico* metagenomes over time and in response to
570 human stool challenge perturbation. Lines show mean JSS in either unchallenged mice
571 ($n$ = 5) or mice challenged with a human stool perturbation after week 4 ($n$ = 15) over
572 time. Shaded areas represent 1 standard deviation computed across mice at each time
573 point, and the dashed line denotes the time of fecal challenge.

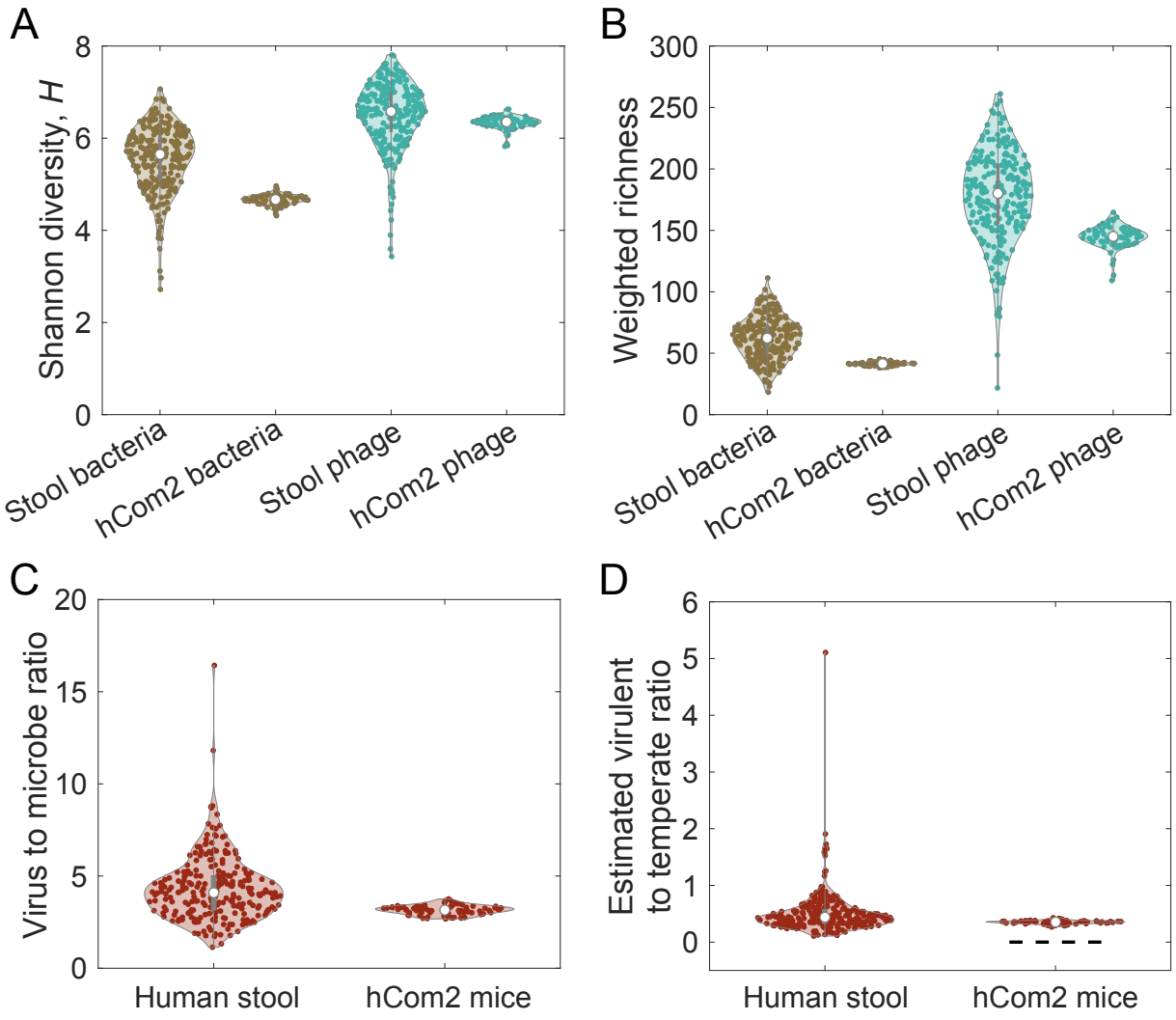



**Figure 4: Large-scale features of human stool viromes are recapitulated in a community constructed only of bacterial isolates.** (A-D) Comparison of Shannon diversity (A), weighted species richness (B), virus to microbe ratio (C), and virulent to temperate ratio (D) in fecal samples from hCom2-colonized gnotobiotic mice (53) compared to human stool samples. Violin plots labelled “stool” represent distributions of microbiome properties across $n$ = 245 healthy adults studied in (55). Violin plots labelled “hCom2” represent samples from 20 gnotobiotic mice colonized with the synthetic community hCom2 ($n$ = 77 total samples, all from unchallenged mice or pre-challenge).

583 Virulent to temperate ratios (VTRs) were estimated using the UHGV database phage
584 species virulence predictions (**Methods**). Dashed black line denotes the null
585 expectation of VTR = 0 for a community constructed from axenic bacterial cultures.

586 **Methods**

587

588 **Meta-analysis of gut phage quantifications**

589 Table 1 summarizes the studies used in our meta-analysis of gut phage abundances.

590 Each row corresponds to a single violin plot in **Fig. 1A**, with the order of the table rows

591 matching the order in which the datasets appear in the figure.

592

| Data origin | Secondary quantification | Method | Population |
|---|---|---|---|
| Liang *et al.* 2020 **(23)** | N/A | VLP EFM | Newborns |
| Liang *et al.* 2020 **(23)** | N/A | VLP EFM | 1-month-old infants |
| Liang *et al.* 2020 **(23)** | This study | Stool WGS | 1-month-old infants |
| Liang *et al.* 2020 **(23)** | N/A | VLP EFM | 4-month-old infants |
| Liang *et al.* 2020 **(23)** | This study | Stool WGS | 4-month-old infants |
| Bikel *et al.* 2021 **(80)** | N/A | VLP EFM | 7–10-year-old children |
| Kim *et al.* 2011 **(26)** | N/A | VLP EFM | Healthy adults |
| Hoyles *et al.* 2014 **(29)** | N/A | VLP EFM | Healthy adults |
| Shkoporov *et al.* 2019 **(25)** | N/A | VLP WGS | Healthy adults |
| Shkoporov *et al.* 2019 **(25)** | This study | Stool WGS | Healthy adults |
| Yachida *et al.* 2019 **(55)** | This study | Stool WGS | Healthy adults |

593 **Table 1: Studies represented in the quantification meta-analysis in Fig. 1A.** "Data

594 origin" column indicates the study that produced the original data, and "Secondary

595 analysis" denotes studies that performed additional bioinformatic analyses represented

596 in **Fig. 1A**. For all stool WGS quantifications, Phanta was used to estimate the genomic

597 virus-to-microbe ratio (gVMR) which was then multiplied by an estimate of gut microbial
598 abundance (32). For studies in which a table of quantifications was not explicitly
599 provided, counts were digitally extracted from figures using WebPlotDigitizer.
600
601 We now briefly describe the different measurement methodologies applied by the
602 studies analyzed here. The measurement methodologies for gut phage abundance fall
603 into three classes (**Fig. 1A**), which quantify different subsets of the gut phage
604 population. One method (labeled "Stool WGS" in **Table 1**) is based on metagenomic
605 sequencing of whole stool samples, from which the ratio of the abundance of phage
606 DNA to that of bacteria DNA can be computed (24). By normalizing this ratio by typical
607 phage and bacterial genome lengths, the ratio of phage to bacterial genome copies is
608 obtained (24), which combined with quantification of absolute bacterial density
609 generates an estimate of absolute phage genome density. This method captures both
610 prophage (e.g., lysogens, phage-plasmids, etc.) and the fraction of phage particles that
611 lyse during DNA extraction. The other two methods involve isolation of virus-like
612 particles (VLPs) as representatives of the phage particle population present within stool.
613 Isolation typically involves 0.2- or 0.45-μm filtration and DNAse/RNAse treatment,
614 among other steps. In one method, VLPs are stained with a DNA-binding dye and
615 enumerated via epifluorescence microscopy (23) (labeled "VLP EFM" in **Table 1**), while
616 in the other method the VLPs are mixed with a known quantity of a non-gut reference
617 phage and metagenomically sequenced, with the reference phage enabling absolute
618 quantification (25) (labeled "VLP WGS" in **Table 1**). In contrast to the method based on
619 bulk stool metagenomics, these VLP-based methods do not capture prophage by

design. Given the drastic methodological differences between bulk and VLP-based approaches, we define two separate VMRs: the genomic VMR (gVMR), based on bulk stool sequencing, and the particle VMR (pVMR), based on VLP-approaches.

In our calculations of pVMR and of phage absolute abundance from gVMR, we require an estimate of the microbial density of the gut microbiome. Note here that we use the term “microbe” to denote all microorganisms (including archaea, bacteria, and unicellular eukaryotes); in practice, the vast majority of gut microbes are bacteria (32) and this is reflected in our taxonomic estimations from Phanta. For all such calculations, we used a standardized value of $0.92\times10^{11}$ microbes/g stool obtained from a comprehensive meta-analysis of stool microbe abundance quantifications from humans >1 month of age (32). Using a single standardized value is justified by the minimal variation of total gut microbial density across human populations >1 month of age (32). Doing so also eliminates the confounding effect of inter-study variability gut microbial density. Indeed, a few gut virome studies reported bacterial density estimates of $\sim10^{9}$ or $\sim10^{10}$ microbes/g stool (23,26), orders of magnitude below well-established values, which led to inflated values of pVMR. We do not know the origin of these discrepancies, but we assume based on the weight of evidence that the microbial density is closer to $0.92\times10^{11}$ microbes/g stool.

For the direct VLP-stool comparisons in **Fig. 1C,D**, we used only subjects for which VLP metagenomes, stool metagenomes, and VLP absolute quantification were available. For the dataset in (25), matching metagenomes were available for the subjects only at the

643 8-month timepoint, and VLP quantification was performed only during months 9-12.
644 Thus, we used the month 8 metagenomes in combination with the average of month 9-
645 12 measurements of each subject. We found that in both infant and adults these VLP
646 metagenomic samples had a high gVMR, median ~$10^3$ compared a median gVMR of
647 ~3-4 found in the corresponding stool samples (as measured by Phanta). This large
648 gVMR difference persists even if *Microviridae* species or all species found only in VLP
649 sample are removed, indicating that the VLP-stool overlap in **Fig. 1C,D** is likely not due
650 to bacterial contamination of the VLP pool.
651

652 **Analysis of prophage copy number data**

653 For our estimation of induction rate from prophage copy number, we use results from
654 Kieft et al. (48), which developed a computational tool, PropagAtE, for estimating
655 whether an integrated prophage is active. They applied their tool to several
656 metagenomic sequencing studies and we use the values of $R$ estimated by their tool
657 (available in Table S3B of their manuscript). We use only prophage-sample
658 combinations detected as present by their tool and perform additional quality filtering
659 requiring minimum median host and prophage coverage >1, and prophage coverage
660 breadth >0.5. We show the resulting summary statistics across cohorts in **Table 2**.
661

| Sample set | *n* samples | *n* pairs | Mean $R$ | Median $R$ | Clipped mean $R$ |
|---|---|---|---|---|---|
| All adults | 123 | 3,459 | 1.04 | 1.01 | 1.09 |
| CRC | 15 | 474 | 1.02 | 0.99 | 1.07 |
| HeQ | 96 | 2842 | 1.05 | 1.02 | 1.09 |
| IjazUZ | 12 | 143 | 1.04 | 1.04 | 1.08 |
| All infants | 79 | 702 | 1.27 | 1.1 | 1.40 |
| Infant (non-abx) | 22 | 254 | 1.33 | 1.22 | 1.43 |
| Infant (abx) | 57 | 448 | 1.24 | 1.02 | 1.39 |

**Table 2: Summary statistics of prophage copy number $R$ across different metagenomic sequencing cohorts, computed based on results from Kieft et al. (48).** "*n* samples" denotes the number of metagenomic sequencing samples in the cohort, "*n* pairs" denotes the total number of prophage-bacterial host pairs identified as present and passing the coverage/breadth requirements in those samples. The clipped mean is the mean computed with values of $R < 1$ set to $R = 1$. The CRC dataset is composed of adults with colorectal adenoma and healthy adults, HeQ is composed of adults with Crohn's disease and healthy adults, IjazUZ is composed of adults with Crohn's disease, Infant (non-abx) is composed infants that were not exposed to antibiotics, Infant (abx) is composed of infants that were exposed to antibiotics.

**Processing and analysis of metagenomic datasets**

All metagenomic datasets analyzed in this manuscript were first subjected to quality control/filtering and adapter removal using the BBDuk decontamination tool in BBTools (81). Settings used were kmer length “k = 23”, “hdist = 1”, trim direction “qtrim = rl” (trim both ends), minimum entropy “entropy = 0.5”, sliding window for entropy calculation “entropywindow = 50”, kmer length for entropy calculation “entropyk = 5”, minimum quality “trimq = 25”, and minimum read length “minlen = 50”. Samples were then deduplicated using the clumplify tool in BBTools. The maximum number of substitutions between duplicate reads was zero (“subs = 0”). We found that deduplication minimally influenced the estimated community compositions.

For taxonomic quantification of samples, we used Phanta, a kmer-based method that simultaneously profiles phage and bacteria (24). Phanta was run using default settings: confidence threshold “confidence_threshold 0.1”, viral genome coverage requirement “cov_thresh_viral 0.1”, viral unique minimizer threshold “minimizer_thresh_viral 0”, bacterial genome coverage requirement “cov_thresh_bacterial 0.01”, and bacterial unique minimizer threshold “minimizer_thresh_bacterial 0”. The “uhggv2_uhgv_mqplus_v1” database was used, which is based on the prokaryotic UHGG database and viral UHGV database. For taxonomy-based analyses, the provided UHGV taxonomy was used, except for quantification of *Microviridae* abundance, for which we used the provided ICTV taxonomy.

**hCom2 metagenome reconstruction from bacterial genomes**

696 To generate mock versions of hCom2-colonized mouse fecal metagenomes, we
697 generated synthetic short-read sequencing datasets using the set of isolate genomes
698 (53). To determine the relative abundances of genome reads within each sample, we
699 used the bacterial compositions estimated by NinjaMap (53). NinjaMap is designed to
700 quantify the composition of synthetic communities in which sequenced genomes are
701 available for all member strains. For each sample, we specified the relative fraction of
702 reads from each genome based on that strain's relative abundance and normalized by
703 its genome length. For genomes that are not assembled into a single contig, the read
704 abundance was split among the contigs weighted by the length of each contig. To
705 generate synthetic shotgun samples, we used Grinder (82) with the following settings:
706 quality levels "-qual levels 33 31", insert distance "-insert_dist 800", read length "-
707 read_dist 140", forward-reverse mate orientation "-mate_orientation FR", characters
708 deleted from reference sequences "-delete_chars '-~*NX'", and distribution of mutations
709 "-mutation_dist uniform 0". The total number of reads generated for each sample was
710 equal to the post-QC read number of the corresponding original mouse fecal sample.
711 Fecal samples with $<10^5$ reads were excluded from the analysis. The resulting samples
712 were subjected to the standard pre-processing pipeline applied to all other
713 metagenomic sequencing data in the manuscript. For the comparison of hCom2 stool to
714 human stool, we exclude human stool samples with $<10^5$ reads.
715

716 **Computation of community summary statistics in metagenomic samples**

717 For all analyses except the hCom-hMock comparison in **Fig. 3** and **Fig. S3**, we used
718 relative taxonomic abundances (which are computed in Phanta by normalizing relative

read abundances to the phage or bacterial genome size). For the hCom-hMock comparison, relative read abundances were used to better compare reconstruction fidelity between phage and bacterial communities. To obtain the genomic virus to microbe ratio (gVMR) of a sample, we calculated the ratio of the total taxonomic abundance of members within the viral superkingdom to the total taxonomic abundance of members of the archaeal, bacterial, and non-human eukaryotic superkingdoms. In practice, the denominator of the gVMR is vastly dominated by the bacterial taxonomic abundance. To obtain the virulent to temperate ratio (VTR), we calculated the ratio of total taxonomic abundance of phage classified as virulent to the total taxonomic abundance of phage classified as temperate. In the main figures, we used the virulence predictions from the Phanta UHGV database (24), which utilizes a combination of scores from BACPHLIP (56) along with information from the PHROG database (83) and geNomad (84). An alternative VTR estimate was performed with scores from PhaTYP (57) (**Fig. S5**). PhaTYP was run on the UHGV genomes using default settings.

Shannon diversity was computed at the species level as $H = -\sum_i x_i \log_2(x_i)$, where $x_i$ is the relative taxonomic abundance of species $i$ within the bacterial or phage community. Weighted richness was computed such that the richness contribution of each species is weighted by $1 - \exp(-x_i/x_0)$, where $x_0 = 10^{-3}$. For the hCom2 reconstruction analysis, the Jensen-Shannon similarity was computed as $\mathrm{JSS} = 1 - \left(\frac{1}{2}\right)\sum_i p_i \log_2\left(\frac{p_i}{m_i}\right) - \left(\frac{1}{2}\right)\sum_i q_i \log_2\left(\frac{q_i}{m_i}\right)$, where $p_i$ and $q_i$ are the relative read abundances of the communities being compared, normalized to sum to 1 within a given taxonomic grouping (e.g., phage at the species level), and $m_i = \frac{p_i+q_i}{2}$.

742

743 **Overview of phage mathematical model**

744 We begin with a mathematical model of a single phage and bacterial species in a well-
745 mixed environment, similar to that of (38). We show here how this model can be
746 reduced to the model presented in the main text. This model involves the concentration
747 of a nutrient ($C$), susceptible cells ($S$), cells containing quiescent prophage ($P$), cells in
748 which the prophage has been activated ($P_a$), viral particles ($V$), and dead cells/viruses of
749 various kinds ($D_i$). All populations are diluted at rate $\delta$. All populations also experience
750 non-dilution mortality/degradation at rate $\omega_i$. All bacterial cells experience the same
751 non-dilution mortality rate $\omega_B$. Susceptible cells and cells containing quiescent prophage
752 grow by consuming the resource. Resource consumption occurs with uptake rate $\mu(C)$
753 for susceptible cells and $\mu(C)(1+s)$ for prophage-containing cells, where $\mu(\cdot)$ is the
754 growth function and $s$ is the fitness benefit/cost of carrying a quiescent prophage.
755 Resources are supplied at a constant rate $\Gamma$. Susceptible cells are exposed to viral
756 particle infection by mass-action kinetics at rate $\kappa$, with a fraction $f_L$ becoming quiescent
757 prophage-containing cells and a fraction $1-f_L$ shifting to the activated cell class ($f_L$
758 models the lysis-lysogeny decision upon initial infection). Prophage-containing cells are
759 induced at rate $\xi$, shifting to the activated class. Cells in the activated class are
760 assumed not to grow and lyse at rate $\gamma$, producing a burst of $B$ viral particles. Viral
761 particles are lost by infecting susceptible cells, failed infection of prophage-containing
762 cells (e.g., to superinfection inhibition mechanisms (41)), and non-dilution mortality.
763 Failed infection occurs at rate $r_e\kappa$, where $r_e$ is the ratio of infection coefficients of
764 prophage-containing and susceptible cells. Dead susceptible, prophage-containing, and

activated cells, with concentrations $D_S$, $D_P$, and $D_a$, respectively, and dead viruses with concentration $D_V$, are produced by non-dilution mortality. Cells that die by phage lysis are tracked separately with concentration $D_L$. Non-lysed dead cells are degraded at rate $\omega_D$, while lysed cells are degraded at rate $\omega_{DL}$, and dead viruses are degraded at rate $\omega_{VD}$. The dynamics governing this model are thus:

$$\frac{dC}{dt} = \Gamma - \mu(C)S - (1+s)\mu(C)P - \delta C, \tag{S1$a$}$$

$$\frac{dS}{dt} = \mu(C)S - \kappa SV - (\delta + \omega_B)S, \tag{S1$b$}$$

$$\frac{dP}{dt} = f_L \kappa SV + (1+s)\mu(C)P - \xi P - (\delta + \omega_B)P, \tag{S1$c$}$$

$$\frac{dP_a}{dt} = (1-f_L)\kappa SV + \xi P - \gamma P_a - (\delta + \omega_B)P_a, \tag{S1$d$}$$

$$\frac{dV}{dt} = \gamma B P_a - \kappa SV - r_e \kappa PV - (\delta + \omega_V)V, \tag{S1$e$}$$

$$\frac{dD_S}{dt} = \omega_B S + (\delta + \omega_D)D_S, \tag{S1$f$}$$

$$\frac{dD_P}{dt} = \omega_B P + (\delta + \omega_D)D_P, \tag{S1$g$}$$

$$\frac{dD_a}{dt} = \omega_B P_a + (\delta + \omega_D)D_a, \tag{S1$h$}$$

$$\frac{dD_L}{dt} = \gamma P_a + (\delta + \omega_{DL})D_L, \tag{S1$i$}$$

$$\frac{dD_V}{dt} = \omega_V V + (\delta + \omega_{DV})D_V, \tag{S1$j$}$$

To recover the model discussed in the main text (**Eq. 1-3**), we make a separation of timescales assumption to reduce the number of state variables in the model, assuming that the nutrient, activated cells, and dead cells are in pseudo-steady-state with the

remaining state variables (i.e., $\frac{dC}{dt} = \frac{dP_a}{dt} = \frac{dD_i}{dt} = 0$). This assumption yields the following expressions:

$$P_a^* = \frac{(1 - f_L)\kappa SV + \xi P}{\gamma + \delta + \omega_B}, \tag{S2a}$$

$$D_S^* = \frac{\omega_B S}{\delta + \omega_D}, \tag{S2b}$$

$$D_P^* = \frac{\omega_B P}{\delta + \omega_D}, \tag{S2c}$$

$$D_a^* = \frac{\omega_B P_a^*}{\delta + \omega_D}, \tag{S2d}$$

$$D_L^* = \frac{\gamma P_a^*}{\delta + \omega_{DL}}, \tag{S2e}$$

$$D_V^* = \frac{\omega_V V}{\delta + \omega_{DV}}, \tag{S2f}$$

These equations can be used to define a simplified set of equations with only the sensitive, prophage, and phage particle abundances

$$\frac{dS}{dt} = \mu(C^*)S - \kappa SV - (\delta + \omega_B)S, \tag{S3a}$$

$$\frac{dP}{dt} = f_L \kappa SV + (1 + s)\mu(C^*)P - \xi P - (\delta + \omega_B)P, \tag{S3b}$$

$$\frac{dV}{dt} = \tilde{B}\xi P + \tilde{B}(1 - f_L)\kappa SV - \kappa SV - r_e \kappa PV - (\delta + \omega_V)V, \tag{S3c}$$

where $C^*(S, P)$ is defined implicitly by $0 = \Gamma - \mu(C^*)S - (1 + s)\mu(C^*)P - \delta C^*$ and $\tilde{B} = Bf_\gamma = B\left(\frac{\gamma}{\gamma + \delta + \omega_B}\right)$. $f_\gamma$ can be interpreted as the fraction of activated cells that are not diluted or die before lysis occurs and thus $\tilde{B}$ can be interpreted as an effective burst size.

**Conditions for robust phage invasion**

In the following two sections, we will assess invasion and stability of phage populations in the model defined by **Eq. S3**. We assume a linear growth function $\mu(C) = \alpha C$ for these derivations, leading to $C^* = \frac{\Gamma}{\alpha S + (1+s)\alpha P + \delta}$. In the absence of bacteria, the resource concentration will saturate at a steady-state value of $C_0^* = \Gamma/\delta$. Bacteria will be able to invade this ecosystem when their initial growth rate exceeds the death and dilution rate: $\frac{\alpha\Gamma}{\delta} > \delta + \omega_B$. Given the stable bacterial colonization seen in the human gut, we assume this condition to be satisfied. More strongly, given that $\Gamma$ and $\delta$ likely vary substantially over time even within a single host (corresponding to variation in food intake and passage time), robust colonization requires $\frac{\alpha\Gamma}{\delta} \gg \delta + \omega_B$.

In the absence of virus, susceptible bacteria will saturate at an equilibrium abundance

$$S_0^* = \frac{\Gamma}{\delta + \omega_B}\left(1 - \frac{\delta(\delta + \omega_B)}{\Gamma\alpha}\right) \approx \frac{\Gamma}{\delta + \omega_B}, \tag{S4}$$

where the approximation follows from the robust bacterial colonization assumption ($\alpha\Gamma \gg \delta(\delta + \omega_B)$). In the absence of lysogeny ($f_L = 0$), viruses will be able to invade this susceptible population if the initial phage replication is greater than death: $\left(\tilde{B} - 1\right)\kappa S_0^* - (\delta + \omega_V) > 0$. Equivalently, the (lytic) basic reproductive number of the virus must be greater than one:

$$R_0 \equiv \frac{\left(\tilde{B} - 1\right)\kappa S_0^*}{\delta + \omega_V} > 1. \tag{S5}$$

As above, since $\Gamma$ and $\delta$ will vary (and thus $S_0^*$ will vary), robust phage invasion will require that $R_0 \gg 1$, and thus this is the regime we are primarily interested in.

**Stability of the prophage-dominated steady state**

Given the estimated abundance of prophage in the gut (**Fig. 1B**), we are particularly interested in the properties of the prophage-only steady state of the model. We will show that under reasonable assumptions this steady state is likely stable in the gut and thus can be invoked in interpretating our induction rate estimates. The prophage-only steady state has $S^* = 0$, $P^* \approx \left(\frac{\delta+\omega_B}{\delta+\omega_B+\xi}\right) S_0^*$, and $V^* = \frac{\tilde{B}\xi P^*}{r_e \kappa P^* + \delta + \omega_V}$, and $C^* \approx \frac{\delta + \omega_B + \xi}{\alpha(1+s)}$. As we are in the robust bacterial colonization regime, we neglect the contribution of dilution to nutrient elimination. This steady state is robust to small perturbations of $P$, $V$, and $C$. From an invasion analysis, it will be robust to small invasion of susceptible bacteria if the net growth rate of these susceptible bacteria is negative:

$$\mu(C^*(0, P^*)) - \kappa V^* - (\delta + \omega_B) < 0. \tag{S6}$$

Substituting in the definition of $C^*$, dividing by $\delta + \omega_B$, and rearranging yields

$$\frac{\frac{\xi}{\delta + \omega_B} - s}{1 + s} < \frac{\kappa V^*}{\delta + \omega_B}. \tag{S7}$$

We can express $V^*$ in terms of $R_0$ as

$$V^* = \frac{\frac{\tilde{B} R_0}{(\tilde{B} - 1)\kappa}(\delta + \omega_B)\xi}{\left(\frac{r_e R_0}{\tilde{B} - 1}\right)(\delta + \omega_B) + \delta + \omega_B + \xi}, \tag{S8}$$

and substituting this equation into the invasion condition yields

$$\frac{\frac{\xi}{\delta + \omega_B} - s}{1 + s} < \frac{\frac{\tilde{B} R_0}{(\tilde{B} - 1)}\left(\frac{\xi}{\delta + \omega_B}\right)}{\frac{r_e R_0}{\tilde{B} - 1} + 1 + \frac{\xi}{\delta + \omega_B}}. \tag{S9}$$

We are particularly interested in the regime where the direct cost (if negative) of the prophage $s$ is small relative to one, but still potentially large relative to other small parameters in the system. This limit is consistent with the modest energetic cost of replicating a phage genome (52) and that non-lytic mobile genetic elements have been observed to rapidly undergo compensatory adaptation to reach very low fitness costs (85). Expanding to lowest order in $s$ leads to

$$\frac{\xi}{\delta+\omega_B} - s < \frac{\frac{\tilde{B}R_0}{(\tilde{B}-1)}\left(\frac{\xi}{\delta+\ \omega_B}\right)}{\frac{r_e R_0}{\tilde{B}-1} + 1 + \frac{\xi}{\delta+\ \omega_B}}. \tag{S10}$$

This condition is violated at very high induction rate ($\xi \gtrsim R_0(\delta+\omega_B)$) and at low induction rate when

$$\frac{\xi}{\delta+\omega_B} < (-s)\left(\frac{r_e}{\tilde{B}} + \frac{1-\tilde{B}^{-1}}{R_0}\right). \tag{S11}$$

The term on the right-hand-side is much smaller than $-s$ in the empirically relevant regime where $\tilde{B} \gg 1$ and $R_0 \gg 1$. Thus, as long as the dominant cost of gut prophage is induction, i.e., $\xi > -s$, as has been experimentally observed for some phage (86), then the gut ecosystem likely exists within a regime where the prophage-only state is stable.

**Overview of induction rate estimation approach**

In the following sections, we show detailed derivations of the induction rate estimates presented in the main text, starting from the single phage-bacteria model in **Eq. S3**. In addition to the estimations based on phage particle to prophage ratio and prophage copy number, we also show an estimation based on cell viability. With currently available data, this estimator is poorly constrained and thus not included in the main

text, but some results from this derivation are used in the derivation of the estimate based on prophage copy number.

In deriving these estimates, we begin with a general form of the calculation that makes no assumptions about the relative abundance of prophage. This approach leads to estimates of the total lysis rate, which includes both induction of prophage and lysis of sensitive cells via non-lysogenic infection. We then simplify these estimates by assuming that the gut is prophage-dominated, leading to the expressions for the average induction rate in the main text. This simplification only affects the interpretation of the resulting estimate: if the prophage-dominated simplification is incorrect and a substantial amount of phage particle production occurs from sensitive cells, then the estimates are still valid as total lysis rate estimates. In the final model section, we show how our framework can be extended to communities with multiple species of bacteria and phage with explicitly time-varying parameters.

**Total lysis rate and induction rate estimate from phage particle to prophage ratio**

Here, we estimate the average lysis rate (including both induction of lysogens and lysis after non-lysogenic infection) from the phage particle to prophage ratio. We begin with the viral dynamics from the timescale-separated prophage model (**Eq. S3**). We define the population-weighted total lysis rate $\eta$ such that $\eta(S+P) = \xi P + (1-f_L)\kappa SV$. We can also rewrite this as $\eta = \xi x_P + (1-f_L)\kappa V x_S$ where $x_i$ are the population relative abundance within the $S+P$ pool of cells. By rearranging **Eq. S3c** when $\frac{dV}{dt}$ is on average zero (i.e. $\frac{1}{\Delta t}\int_0^{\Delta t}\frac{dV}{dt}\,dt \approx 0$), we can obtain an expression for $\eta^*$:

$$\eta^* = \frac{1}{\tilde{B}}\left(\frac{V^*}{S^* + P^*}\right)(\kappa S^* + r_e \kappa P^* + \omega_V + \delta), \tag{S12}$$

where asterisks denote the time-averaged value $x^* = \frac{1}{\Delta t}\int_0^{\Delta t} x(t)\ dt$. This approximation assumes that the microbiome is in a statistical steady state (no net trend in $V$) and that $\Delta t$ is long enough that time averages have converged to their ensemble-averaged values. At a minimum, this assumption requires that $\Delta t \gtrsim 1/\delta$. The assumption of a statistical steady state is supported by the results of our absolute abundance meta-analysis (**Fig. 1**). Given the limited knowledge of $\kappa$, $r_e$, and $\omega_V$ in the gut, we use **Eq. S12** to construct a lower bound on $\eta^*$:

$$\eta^* \geq \frac{1}{\tilde{B}}\left(\frac{V^*}{S^* + P^*}\right)\delta. \tag{S13}$$

In the prophage-dominated regime (i.e., $S^* = 0$) we recover $\eta^* = \xi^* \geq \left(1/\tilde{B}\right)(V^*/P^*)\delta$, a bound on the induction rate.

In practice, measured pVMR may not be $\frac{V^*}{S^*+P^*}$ due to the contribution of dead cells and dead viruses. If all populations are represented in the measurement, the pVMR will instead be

$$\text{pVMR} = \frac{V^* + D_V^*}{S^* + P^* + P_a^* + D_S^* + D_P^* + D_a^* + D_L^*}. \tag{S14}$$

This is related to $\frac{V^*}{S^*+P^*}$ by:

$$\frac{V^*}{S^* + P^*} = \left(\frac{\nu_V}{\nu_B}\right)\left(\frac{S^* + P^* + P_a^*}{S^* + P^*}\right) \cdot \text{pVMR}, \tag{S15}$$

where $\nu_i$ is the cell or viral viability fraction (the fraction of cells or viruses that are viable). We now substitute this expression into **Eq. S12** and use the fact that $\frac{\nu_V}{1-\nu_V} = \frac{V^*}{D_V^*} = \frac{\delta+\omega_{DV}}{\omega_V}$ to yield

$$\eta^* = \frac{\mathrm{pVMR}}{\tilde{B}}\left(\frac{S^* + P^* + P_a^*}{S^* + P^*}\right)\left(\frac{1}{\nu_B}\right)(\nu_V \kappa S^* + \nu_V r_e \kappa P^* + (1-\nu_V)\omega_{DV} + \delta). \tag{S16}$$

Thus, utilizing a pVMR including the dead populations still functions as a lower bound estimate of $\eta^*$:

$$\eta^* \geq \left(\frac{\mathrm{pVMR}}{\tilde{B}}\right)\delta. \tag{S17}$$

**Cell death and total lysis rate estimates from live cell fraction**

Here, we derive estimates of both the non-lysis cell mortality rate $\omega_B$ and the total lysis rate $\eta$ based on the fraction of cells that are living/viable within the microbiome, defined in our model as $\nu_B^* = \frac{S^*+P^*+P_a^*}{S^*+P^*+P_a^*+D_S^*+D_P^*+D_a^*+D_L^*}$. We begin by substituting in the steady-state population abundances to the expression $\frac{\nu_B^*}{1-\nu_B^*} = \frac{S^*+P^*+P_a^*}{D_S^*+D_P^*+D_a^*+D_L^*}$,

$$\frac{\nu_B^*}{1-\nu_B^*} = \frac{S^* + P^* + P_a^*}{(S^* + P^* + P_a^*)\left(\frac{\omega_B}{\delta + \omega_D}\right) + \frac{\gamma P_a^*}{\delta + \omega_{DL}}}, \tag{S18}$$

which when solved for $\omega_B$ yields

$$\omega_B = (\delta + \omega_D)\left(\frac{1-\nu_B^*}{\nu_B^*} - \frac{\gamma P_a^*}{(\delta + \omega_{DL})(S^* + P^* + P_a^*)}\right). \tag{S19}$$

This equation provides an upper bound estimate for $\omega_B$:

$$\omega_B \leq (\delta + \omega_D)\left(\frac{1-\nu_B^*}{\nu_B^*}\right), \tag{S20}$$

which will be utilized later in deriving the estimate of total lysis rate from prophage copy number. The value of $\nu_B^*$ in stool has been estimated at $\sim 0.5 - 0.8$ based on cell permeability measurements, leading to $\frac{1-\nu_B^*}{\nu_B^*} \approx 1$ (87,88). Thus, the bacterial death rate is at most similar in magnitude to the sum of dilution and cell degradation rate.

From **Eq. S18**, we can also derive an estimate for the total lysis rate using the steady-state expression $P_a^* = \frac{\eta(S^*+P^*)}{(\gamma+\delta+\omega_D)}$. Solving **Eq. S18** for $\eta$ yields:

$$\frac{\eta}{\gamma+\delta+\omega_D} = \frac{\left(\frac{1-\nu_B^*}{\nu_B^*} - \frac{\omega_B}{\delta+\omega_D}\right)}{\frac{\gamma}{\delta+\omega_{DL}} - \left(\frac{1-\nu_B^*}{\nu_B^*} - \frac{\omega_B}{\delta+\omega_D}\right)}. \tag{S21}$$

This equation is potentially usable to provide another lysis rate estimate, and in the prophage-dominated regime becomes an induction rate estimate. However, the values of $\omega_D$, $\omega_B$, and $\omega_{DL}$ are currently poorly constrained. For example, one cell viability measurement method is based on comparing the fraction of 16S rDNA found inside and outside of intact cells (87) and it is not known how rapidly extracellular DNA is degraded inside the gut. There are also potential technical issues in the measurement of $\nu_B$, as it is unclear to what extent cells lysed by phage are detected by current cell viability measurements. If the lysis process degrades the host genome or leads to total destruction of the cellular structure, the population of cells dying due to lysis would be underestimated by methods relying on extracellular genomic DNA or permeable cell remains.

**Total lysis rate and induction rate estimates from integrated prophage copy number**

Here, we estimate the average total lysis rate using lysogeny copy number $R$. We assume that $R$ includes the contribution of viral particles, dead viral particles, and all dead cells, and we show that the inclusion of these classes does not substantially alter our induction rate estimation. Each activated cell contributes $B_a$ prophage copies, each lysogen cell contributes one prophage copy, and lysed cells contribute no prophage copies. All cells contribute a single bacteria genome copy. We first define $R$ in terms of our steady-state model populations:

$$R^* \equiv \mathrm{gVMR} = \frac{V^* + D_V^* + P^* + B_a P_a^* + D_P^* + B_a D_a^*}{S^* + P^* + P_a^* + D_S^* + D_P^* + D_a^* + D_L^*}. \tag{S22}$$

As all cells have the same non-lysis mortality rate, **Eq. S22** can be rearranged to

$$R^* = \mathrm{pVMR} + \frac{P^* + B_a P_a^*}{S^* + P^* + P_a^*}\left(1 + \frac{D_P^*}{P^*}\right)\nu_B^*, \tag{S23}$$

$$R^* - \mathrm{pVMR} = \frac{P^* + B_a P_a^*}{S^* + P^* + P_a^*}\left(1 - \frac{D_L^*}{S^* + P^* + P_a^* + D_S^* + D_P^* + D_a^* + D_L^*}\right). \tag{S24}$$

We can then substitute in the steady-state population values to express all dead cell populations in terms of living cells populations:

$$R^* - \mathrm{pVMR} = \frac{P^* + B_a P_a^*}{S^* + P^* + P_a^*}\left(1 - \frac{\frac{\gamma P_a^*}{\delta + \omega_{DL}}}{(S^* + P^* + P_a^*)\left(1 + \frac{\omega_B}{\delta + \omega_D}\right) + \frac{\gamma P_a^*}{\delta + \omega_{DL}}}\right). \tag{S25}$$

Rearranging and using the fact that $\gamma = \left(\frac{f_\gamma}{1-f_\gamma}\right)(\delta + \omega_B)$ yields

$$R^* - \mathrm{pVMR} = \left(\frac{P^* + B_a P_a^*}{S^* + P^* + P_a^*\left[1 + \left(\frac{\delta + \omega_D}{\delta + \omega_{DL}}\right)\left(\frac{f_\gamma}{1 - f_\gamma}\right)\left(\frac{\delta + \omega_B}{\delta + \omega_D + \omega_B}\right)\right]}\right), \tag{S26}$$

which when solved for $P_a^*$ yields

$$P_a^* = \frac{(S^* + P^*)(R^* - \mathrm{pVMR}) - P^*}{B_a - (R^* - \mathrm{pVMR})\mathrm{Q}}, \tag{S27}$$

where $Q = 1 + \left(\frac{\delta+\omega_D}{\delta+\omega_{DL}}\right)\left(\frac{f_\gamma}{1-f_\gamma}\right)\left(\frac{\delta+\omega_B}{\delta+\omega_D+\omega_B}\right)$. From our steady-state solution for $P_a^*$ we have that $\eta^* = P_a^*\left(\frac{\gamma+\omega_B+\delta}{S^*+P^*}\right)$, providing an estimate of $\eta^*$:

$$\eta^* = (\gamma + \omega_B + \delta)\frac{(R^* - \mathrm{pVMR}) - x_P^*}{B_a - (R^* - \mathrm{pVMR})\mathrm{Q}}. \tag{S28}$$

The effect of dead cells and viruses enters the expression via the factor $\mathrm{Q}$, which will inflate the lysis rate. However, this term cannot be greater than $O(1)$, and thus if $B_a$ is large, the impact of dead material is minimal. Empirically, the value of $\omega_B$ is poorly constrained, but we can use results from the cell viability derivation above (**Eq. S20**) to relate this rate to the cell viability fraction $\nu_B$ and the degradation rate of dead cells $\omega_D$:

$$\eta^* \leq \left[\gamma + (\delta + \omega_D)\left(\frac{1-\nu_B^*}{\nu_B^*}\right) + \delta\right]\frac{(R^* - \mathrm{pVMR}) - x_P^*}{B_a - (R^* - \mathrm{pVMR})\mathrm{Q}}. \tag{S29}$$

To reach the order of magnitude bound shown in the main text, we assume prophage dominance $x_P^* \approx 1$, that the number of prophage copies in activated cells is similar to the burst size $B_a \approx \tilde{B}$, and that dead cells are primarily removed by dilution $\omega_D \ll \delta$. Based on empirical measurements, we also assume that $R^* \gg \mathrm{pVMR}$, $\tilde{B} \gg 1$, and $\frac{1-\nu_B^*}{\nu_B^*} \approx 1$, yielding

$$\xi^* \lesssim \frac{(\gamma + \delta)(R^* - 1)}{(\tilde{B} - 1) - (R^* - 1)}. \tag{S30}$$

We now briefly discuss potential bioinformatic/sequencing technical artifacts that could influence the measurement of $R$. One potential factor that could systematically skew the

above induction rate estimate is sequencing bias between prophage and host (e.g., due to GC content differences between host bacteria and prophage (89)). However, based on the negative control analyses performed by (48), these biases do not appear significant, as non-induced prophage had $R \approx 1$. If large biases existed, $R$ in non-induced phage would differ significantly from $1$. Another possible confounding factor in estimating $R$ is the presence of the prophage within only a subpopulation of the bacterial host, leading to a lower value of $R$. However, this is unlikely to affect our current analyses, as the $R$ values we analyze were computed based on metagenomically assembled contigs containing both prophage and bacterial host sequence. Assembly of such mixed contigs from a mixed lysogen/sensitive population is highly unlikely due to degeneracies in the possible assembly paths. In the case of both possible biases, our framework can readily accommodate improved estimates of $R$ as sequencing and bioinformatic approaches improve.

**Extension of the model to multiple phage and bacterial species in time-varying environments**

We now generalize our model to complex communities in time-varying environments. For simplicity, we begin with the timescale-separated version of the model and focus on the prophage-dominated case in which most lysis is due to induction, as for the single bacteria-phage regime studied above.

We now track the dynamics of multiple types of phage (indexed by $i$) and multiple types of bacteria (indexed by $j$), such that the total number of phage particles is $V(t) \equiv$

$\sum_i V_i(t)$ and the total number of bacteria is $N(t) \equiv \sum_j N_j(t)$. The bacterial "type" $j$ encompasses both the taxonomic identity of a bacteria and its infection status (i.e., the $N_j$ also include bacteria infected by a prophage). To keep track of infection status and the phage-bacteria interaction network, we introduce bookkeeping parameters $I_{ij}$ and $A_{ijk}$, respectively. The parameter $I_{ij}$ is 1 if bacteria $j$ is infected with a prophage of phage $i$ and zero otherwise. Thus, the total number of prophage in this system is $P(t) \equiv \sum_i P_i(t) = \sum_{ij} I_{ij} N_j(t)$. Note that generally $P(t) \neq N(t)$, as multiple phage can infect a single bacteria. The second parameter, $A_{ijk}$, is 1 if an infection of bacteria of type $j$ by phage $i$ produces an infected bacterium of type $k$, and 0 otherwise. Using this notation, we now define the multispecies generalization of **Eq. S3**:

$$\frac{dN_j}{dt} = \mu_j(t)N_j - \sum_i \kappa_{ij}(t)N_jV_i + \sum_{i,k} A_{ikj}\kappa_{ij}(t)N_kV_i - \sum_i I_{ij}\xi_{ij}(t)N_j - \delta(t)N_j - \omega_{B,j}(t)N_j, \quad \text{(S31}a\text{)}$$

$$\frac{dV_i}{dt} = \sum_j I_{ij}B_{ij}(t)\xi_{ij}(t)N_j - \sum_j \kappa_{ij}(t)N_jV_i - \sum_j r_{ij}(t)\kappa_{ij}(t)N_jV_i - \delta(t)V_i - \omega_i(t)V_{V,i}, \quad \text{(S31b)}$$

where we have also allowed the rate parameters to explicitly depend on time. We have also approximated $f_L = 1$ for simplicity.

To relate these dynamics to the total pVMR, we now sum **Eq. S13b** over the viral index $i$ and substitute $\frac{dx}{dt}\left(\frac{1}{x}\right) = \frac{d\log(x)}{dt}$, yielding

$$\frac{d\log V}{dt} = \frac{\sum_{ij} I_{ij}B_{ij}(t)\xi_{ij}(t)N_j}{\sum_{ij} I_{ij}N_j} \cdot \left(\frac{P}{V}\right) - \frac{\sum_{ij}\kappa_{ij}(t)N_jV_i}{\sum_i V_i} - \frac{\sum_{ij} r_{ij}(t)\kappa_{ij}(t)N_jV_i}{\sum_i V_i} - \delta(t) - \frac{\sum_{ij}\omega_{V,i}(t)V_i}{\sum_i V_i}. \quad \text{(S32)}$$

**Eq. S14** can be rewritten in a more compact form as

$$\frac{d\log V}{dt} = \overline{B}(t)\overline{\xi}(t)\left(\frac{P}{V}\right) - \overline{\psi}_I(t) - \overline{\psi}_N(t) - \delta(t) - \overline{\omega}_V(t), \quad \text{(S33)}$$

where we have defined the microbiome averages

$$\overline{B}(t) \equiv \frac{\sum_{ij} I_{ij} B_{ij}(t) \xi_{ij}(t) N_j}{\sum_{ij} I_{ij} \xi_{ij}(t) N_j}, \tag{S34a}$$

$$\bar{\xi}(t) \equiv \frac{\sum_{i,j} I_{ij} \xi_{ij}(t) N_j}{\sum_{ij} I_{ij} N_j}, \tag{S34b}$$

$$\overline{\psi}_I(t) \equiv \frac{\sum_{ij} \kappa_{ij}(t) N_j V_i}{\sum_i V_i}, \tag{S34c}$$

$$\overline{\psi}_N(t) \equiv \frac{\sum_{ij} r_{ij}(t) \kappa_{ij}(t) N_j V_i}{\sum_i V_i}. \tag{S34d}$$

$$\overline{\omega}_V(t) \equiv \frac{\sum_i \omega_{V,i}(t) V_i}{\sum_i V_i} \tag{S34e}$$

Integrating **Eq. S15** over long times yields

$$0 \approx \bar{B}^* \bar{\xi}^* \left(\frac{P}{V}\right)^* - \overline{\psi}_I^* - \overline{\psi}_N^* - \delta^* - \overline{\omega}_V^*, \tag{S35}$$

where the asterisks again denote the time-averaged value $x^* = \frac{1}{\Delta t} \int_0^{\Delta t} x(t)\ dt$. Since the microbiome is in a statistical steady state over long times (**Fig. 1B**), we can estimate the averages over time by taking an average over independent hosts. This procedure yields a connection between the rate parameters and the VLP-to-prophage ratio from **Fig. 1** and thus a lower bound similar to the one estimated from the single phage-bacteria model:

$$\bar{\xi}^* \gtrsim \frac{1}{\bar{B}^*} \left(\frac{V^*}{P^*}\right) \delta^*. \tag{S36}$$

This bound assumes that the burst size, induction rate, and particle-to-prophage ratio are largely uncorrelated in time. If this assumption is violated, the estimate represents a particular weighted average of the induction rate bound:

$$\frac{1}{\bar{B}^*}\left(\frac{V^*}{P^*}\right)\delta^* = \left(\frac{1}{\Delta t}\right)\frac{\int_0^{\Delta t}\bar{\xi}\,\bar{B}\left(\frac{P}{V}\right)dt}{\bar{B}^*\left(\frac{V^*}{P^*}\right)}. \tag{S37}$$

To generalize this bound to the case of pVMR including dead material, we begin with the multispecies version of the dead virus dynamics:

$$\frac{dD_{V,i}}{dt} = \omega_{V,i}(t)V_i - \delta(t)D_{V,i} - \omega_{DV,i}(t)D_{V,i}, \tag{S38}$$

which then yields an expression for the dynamics of the total dead virus population:

$$\frac{d\log D_{V,i}}{dt} = \bar{\omega}_V(t)\left(\frac{V}{D_V}\right) - \delta(t) - \bar{\omega}_{DV,i}(t), \tag{S39}$$

where $\bar{\omega}_{DV,i}(t) \equiv \frac{\sum_i \omega_{DV,i}(t)D_{V,i}}{\sum_i D_{V,i}}$. At statistical steady state, this leads to $\frac{\nu_V^*}{1-\nu_V^*} = \frac{V^*}{D_V^*} \approx \frac{\delta^*+\bar{\omega}_{DV}^*}{\bar{\omega}_V^*}$, which when combined with **Eq. S35** shows the lower bound is preserved when the pVMR accounts for dead populations, as in the earlier single species derivation.

Note that in this section we have only analyzed a simple case of this community model, and further analysis, such as exploring the role of temporal correlations and the relative contribution of induction and direct lysis, is a promising direction for future theoretical phage ecology work.

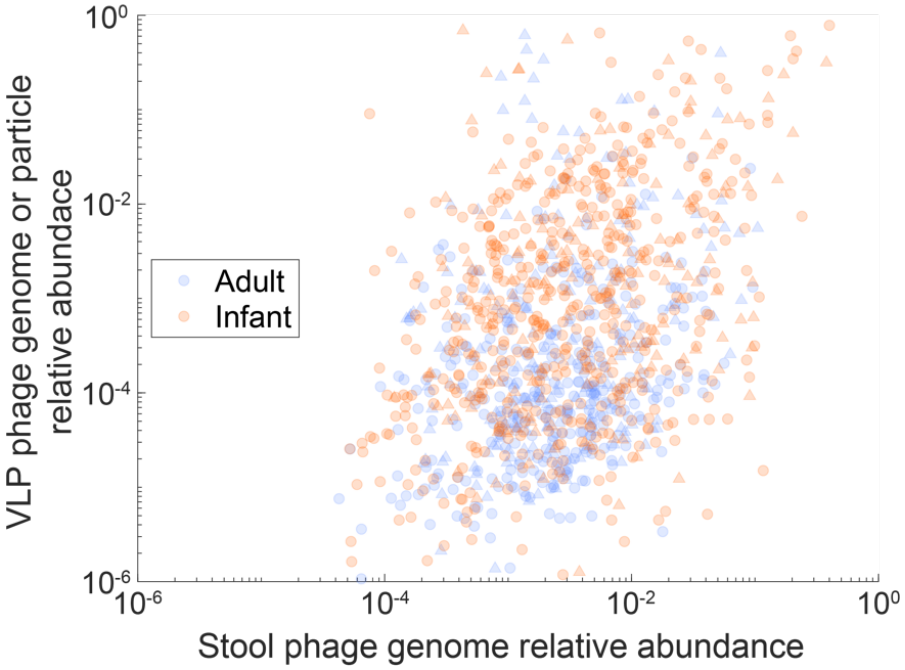
Here, we have shown the multispecies generalization of the induction rate estimate from VLP-to-prophage ratio. The rate estimates computed from cell viability will similarly extend to the multispecies context, as we model all sources of death in aggregate, independent of which phage causes lysis. The induction rate estimate from the lysogen copy number is performed on a prophage-by-prophage basis, hence it is not affected in a multispecies context. However, this context will lead to a difference in the kind of

1060 average used in the estimate: unlike the average computed from the VLP-to-prophage
1061 ratio, the average from copy number average is not abundance weighted and includes
1062 only lysogens captured with their host contig.
1063

## Data Availability

All code used in this manuscript is available at https://github.com/jaimegelopez/gut_phage_quantification. All data analyzed in this manuscript is publicly available. Processed final versions of the datasets (e.g. estimated taxonomic compositions) are available in the GitHub repository.

1069 **Supplementary Figures**




1070

1071 **Figure S1: Species-level relative abundance analysis of the overlap between**

1072 **phage community quantifications using VLP- and stool-based approaches.** Data

1073 and plotting methods are the same as **Fig. 1C**, except that relative abundance was

1074 used instead of absolute genome/particle density. This figure includes only phage

1075 shared between the VLP fraction and stool, corresponding to the central scatter plot of

1076 **Fig. 1C**. Relative abundance was defined relative to total taxonomic abundance of

1077 phage within VLPs/stool. Note that without the absolute abundance normalization

1078 employed in **Fig. 1C**, adult VLP abundances appear to be substantially lower than infant

1079 abundances.

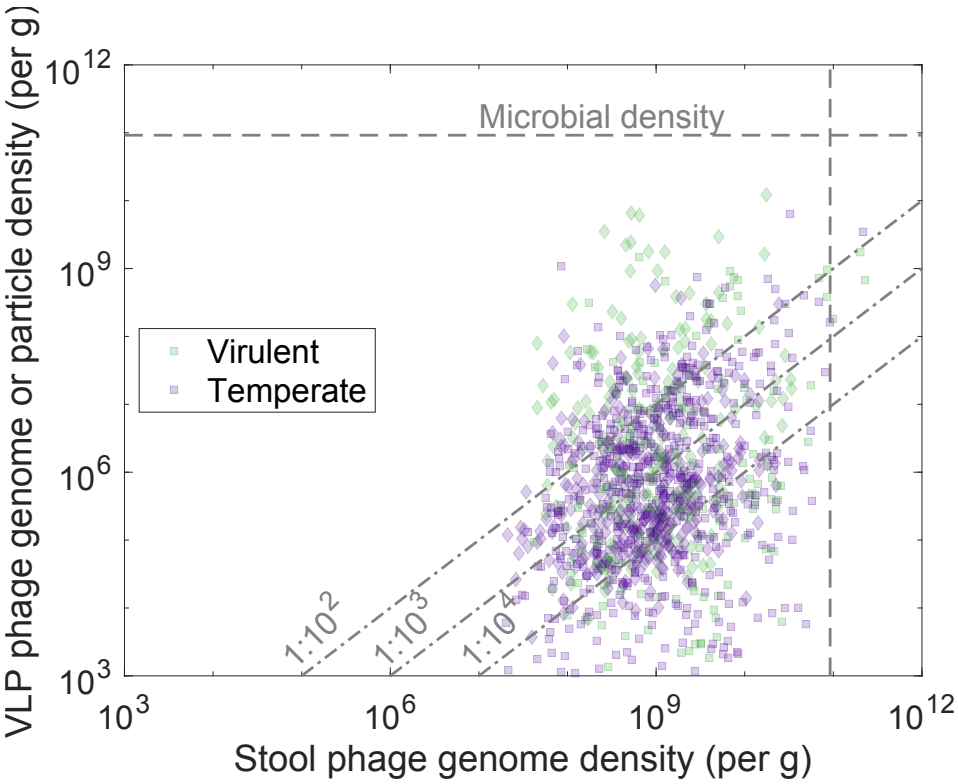



**Figure S2: Species-level absolute abundance analysis of the overlap between VLP- and stool-based phage community quantifications, colored by predicted virulence.** This figure is the same as **Fig. 1C**, except points are colored by predicted virulence rather than their dataset of origin. Diamond markers represent points from adult samples, while squares represent points from infant samples.

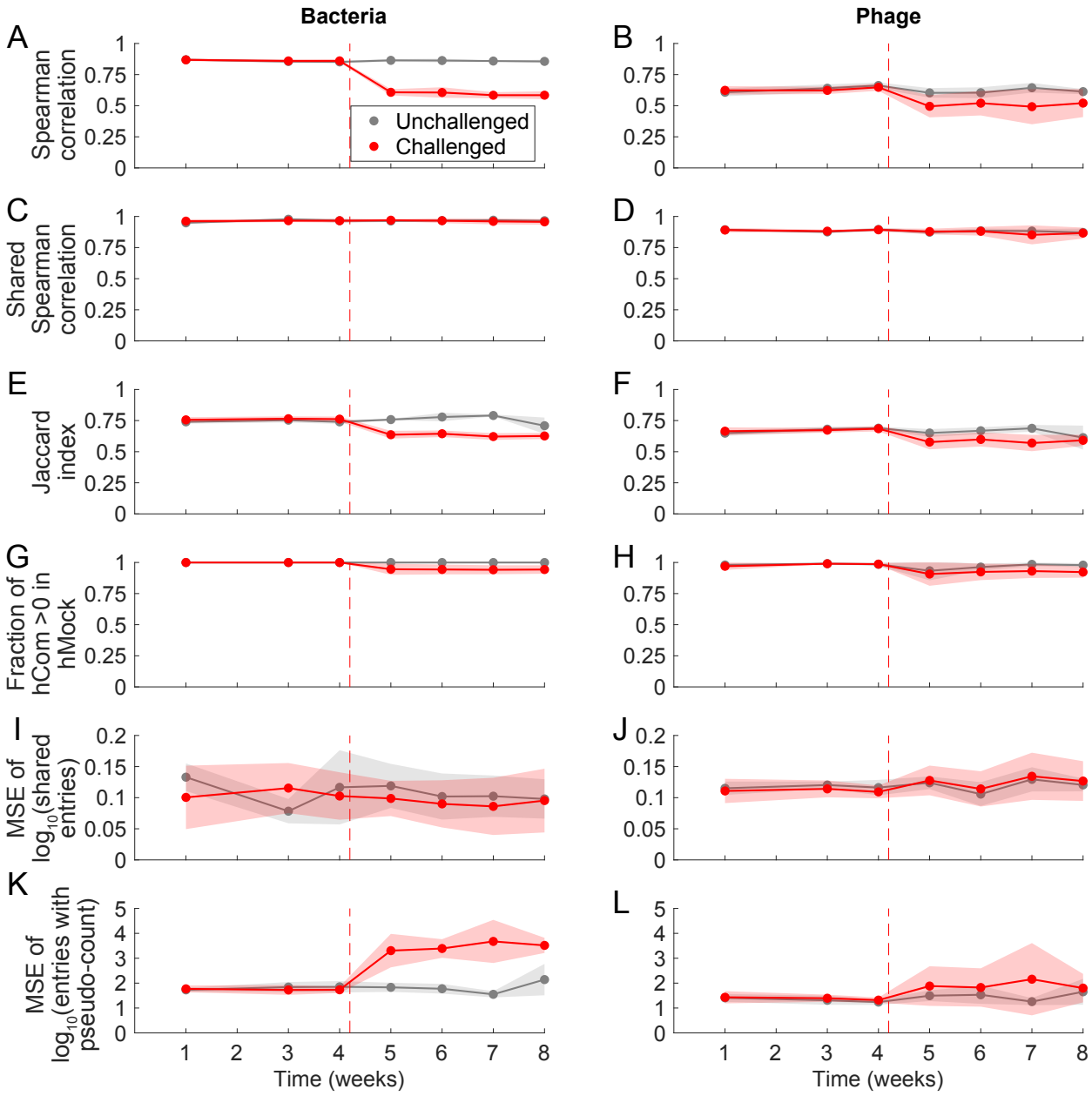



**Figure S3: Time series of various reconstruction evaluation metrics in hCom2-colonized mice.** Each row is a version of **Fig. 3C,D** but instead showing Spearman correlation of taxa read abundances (A,B), Spearman correlation of read abundances of species shared between hCom2/hMock sample pairs (C,D), Jaccard index (E,F), fraction of read abundances in hCom2-colonized mouse feces that are nonzero in hMock (G,H), mean squared error (MSE) of $\log_{10}$(read abundance) of species shared

1093 between hCom2/hMock sample pairs (I,J), and MSE of $\log_{10}$(read abundance)
1094 computed with a relative abundance pseudocount of $10^{-7}$ (K,L). All metrics were
1095 computed using species-level relative read abundances. Jaccard index is the number of
1096 shared species between an hCom2/hMock sample pair divided by the total number of
1097 species with nonzero abundance in at least one of the two samples. For (G,H), shared
1098 read abundances were normalized to the total bacterial or phage abundance in the
1099 hCom2-colonized mouse fecal sample.

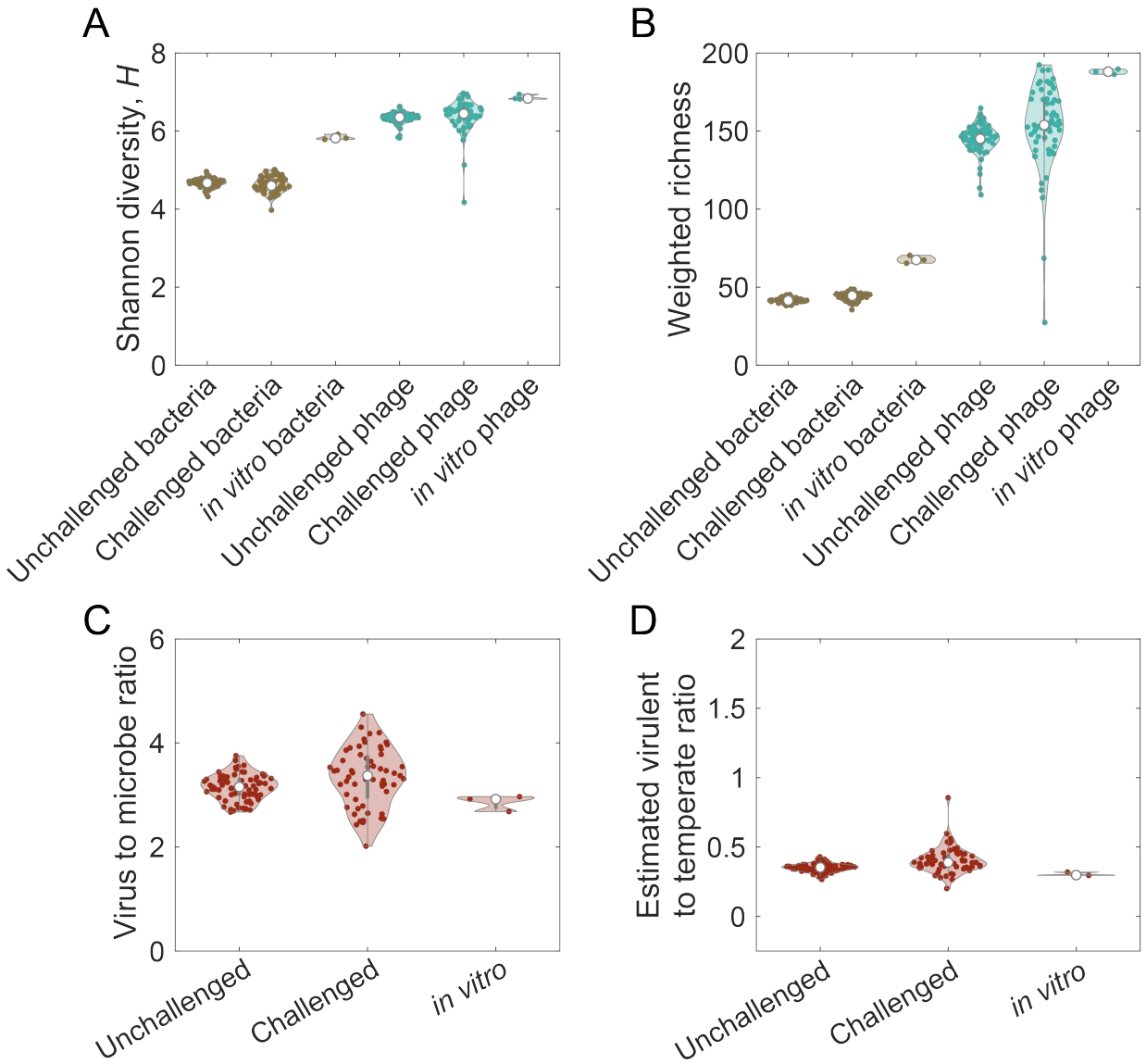



**Figure S4: Comparison of phage and bacterial community properties in the feces of hCom2-colonized mice across conditions.** Community properties and plotting methods are the same as in **Fig. 4.** 'Unchallenged' corresponds to *in vivo* samples from hCom2-colonized gnotobiotic mice that have not been exposed to a human stool sample, while 'Challenged' denotes samples from hCom2-colonized mice that have been exposed to a human stool challenge. '*in vitro*' corresponds to hCom2 communities grown *in vitro*.

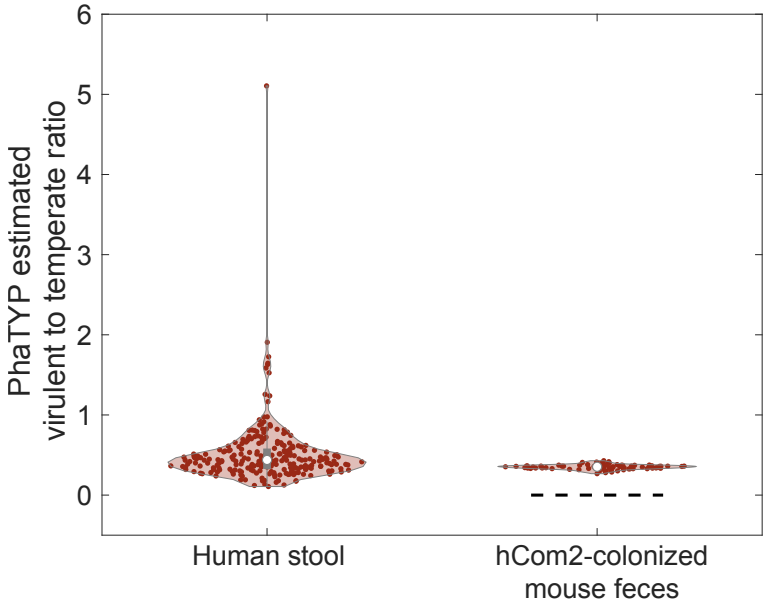



**Figure S5: Comparison of the virulent-to-temperate ratio (VTR) between stool from humans and feces from hCom2-colonized mice, as estimated by PhaTYP.** Equivalent to **Fig. 4D**, except the virulence prediction of phage genomes was performed using PhaTYP (57) instead of using the Phanta UHGV database (24).

## References


1. Routy B, Le Chatelier E, Derosa L, Duong CP, Alou MT, Daillère R, et al. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science. 2018;359(6371):91–7.

2. Thiele Orberg E, Meedt E, Hiergeist A, Xue J, Heinrich P, Ru J, et al. Bacteria and bacteriophage consortia are associated with protective intestinal metabolites in patients receiving stem cell transplantation. Nat Cancer. 2024 Jan 3;5(1):187–208.

3. Zhang F, Aschenbrenner D, Yoo JY, Zuo T. The gut mycobiome in health, disease, and clinical applications in association with the gut bacterial microbiome assembly. Lancet Microbe. 2022 Dec;3(12):e969–83.

4. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6(7):960–70.

5. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184(4):1098–109.

6. Guerin E, Shkoporov AN, Stockdale SR, Comas JC, Khokhlova EV, Clooney AG, et al. Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. Microbiome. 2021;9(1):1–21.

7. Shen J, Zhang J, Mo L, Li Y, Li Y, Li C, et al. Large-scale phage cultivation for commensal human gut bacteria. Cell Host Microbe. 2023;31(4):665–77.

8. Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI. Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut. Proc Natl Acad Sci. 2013;110(50):20236–41.

9. Sutcliffe SG, Shamash M, Hynes AP, Maurice CF. Common oral medications lead to prophage induction in bacterial isolates from the human gut. Viruses. 2021;13(3):455.

10. Borodovich T, Shkoporov AN, Ross RP, Hill C. Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. Gastroenterol Rep. 2022;10:goac012.

11. Frazão N, Sousa A, Lässig M, Gordo I. Horizontal gene transfer overrides mutation in *Escherichia coli* colonizing the mammalian gut. Proc Natl Acad Sci. 2019 Sep 3;116(36):17906–15.

12. Johansen J, Atarashi K, Arai Y, Hirose N, Sørensen SJ, Vatanen T, et al. Centenarians have a diverse gut virome with the potential to modulate metabolism and promote healthy lifespan. Nat Microbiol. 2023 May 15;8(6):1064–78.

13. Brunse A, Deng L, Pan X, Hui Y, Castro-Mejía JL, Kot W, et al. Fecal filtrate transplantation protects against necrotizing enterocolitis. ISME J. 2022;16(3):686–94.

14. Ott SJ, Waetzig GH, Rehman A, Moltzau-Anderson J, Bharti R, Grasis JA, et al. Efficacy of sterile fecal filtrate transfer for treating patients with Clostridium difficile infection. Gastroenterology. 2017;152(4):799–811.

15. Sinha A, Li Y, Mirzaei MK, Shamash M, Samadfam R, King IL, et al. Transplantation of bacteriophages from ulcerative colitis patients shifts the gut bacteriome and exacerbates the severity of DSS colitis. Microbiome. 2022;10(1):1–23.

16. Adiliaghdam F, Amatullah H, Digumarthi S, Saunders TL, Rahman RU, Wong LP, et al. Human enteric viruses autonomously shape inflammatory bowel disease phenotype through divergent innate immunomodulation. Sci Immunol. 2022;7(70):eabn6660.

17. Fuhrman JA. Marine viruses and their biogeochemical and ecological effects. Nature. 1999 Jun;399(6736):541–8.

18. Suttle CA. Marine viruses — major players in the global ecosystem. Nat Rev Microbiol. 2007 Oct;5(10):801–12.

19. Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. Science. 2021 Oct 22;374(6566):488–92.

20. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018 Jun 4;3(7):754–66.

21. Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol Oceanogr. 2000 Sep;45(6):1320–8.

22. Maslov S, Sneppen K. Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems. Sci Rep. 2017 Jan 4;7(1):39642.

23. Liang G, Zhao C, Zhang H, Mattei L, Sherrill-Mix S, Bittinger K, et al. The stepwise assembly of the neonatal virome is modulated by breastfeeding. Nature. 2020 May 28;581(7809):470–4.

24. Pinto Y, Chakraborty M, Jain N, Bhatt AS. Phage-inclusive profiling of human gut microbiomes with Phanta. Nat Biotechnol. 2023;1–12.

25. Shkoporov AN, Clooney AG, Sutton TD, Ryan FJ, Daly KM, Nolan JA, et al. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe. 2019;26(4):527–41.

26. Kim MS, Park EJ, Roh SW, Bae JW. Diversity and abundance of single-stranded DNA viruses in human feces. Appl Environ Microbiol. 2011;77(22):8062–70.

27. Dahlman S, Avellaneda-Franco L, Kett C, Subedi D, Young RB, Gould JA, et al. Temperate gut phages are prevalent, diverse, and predominantly inactive. bioRxiv [Internet]. 2023; Available from: https://www.biorxiv.org/content/early/2023/08/18/2023.08.17.553642

28. Shkoporov AN, Khokhlova EV, Stephens N, Hueston C, Seymour S, Hryckowian AJ, et al. Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. BMC Biol. 2021 Dec;19(1):163.

29. Hoyles L, McCartney AL, Neve H, Gibson GR, Sanderson JD, Heller KJ, et al. Characterization of virus-like particles associated with the human faecal and caecal microbiota. Res Microbiol. 2014 Dec;165(10):803–12.

30. Parras-Moltó M, Rodríguez-Galet A, Suárez-Rodríguez P, López-Bueno A. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. Microbiome. 2018;6(1):1–18.

31. Forterre P, Soler N, Krupovic M, Marguet E, Ackermann HW. Fake virus particles generated by fluorescence microscopy. Trends Microbiol. 2013 Jan;21(1):1–5.

32. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol. 2016;14(8):e1002533.

33. Kaletta J, Pickl C, Griebler C, Klingl A, Kurmayer R, Deng L. A rigorous assessment and comparison of enumeration methods for environmental viruses. Sci Rep. 2020 Oct 29;10(1):18625.

34. Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR, Daly KM, et al. Reproducible protocols for metagenomic analysis of human faecal phageomes. Microbiome. 2018 Dec;6(1):68.

35. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021 Jan;39(1):105–14.

36. Schmidtke DT, Hickey AS, Liachko I, Sherlock GJ, Bhatt AS. Analysis and culturing of the prototypic crAssphage reveals a phage-plasmid lifestyle. bioRxiv. 2024;2024–03.

37. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, et al. Cryptic prophages help bacteria cope with adverse environments. Nat Commun. 2010 Dec 21;1(1):147.

38. Li G, Cortez MH, Dushoff J, Weitz JS. When to be temperate: on the fitness benefits of lysis vs. lysogeny. Virus Evol. 2020 Jul 1;6(2):veaa042.

1218 39. Stewart FM, Levin BR. The population biology of bacterial viruses: why be
1219 temperate. Theor Popul Biol. 1984;26(1):93–117.

1220 40. Geng Y, Nguyen TVP, Homaee E, Golding I. Using bacterial population dynamics to
1221 count phages and their lysogens. Nat Commun. 2024 Sep 6;15(1):7814.

1222 41. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al.
1223 Prophages mediate defense against phage infection through diverse mechanisms.
1224 ISME J. 2016 Dec 1;10(12):2854–66.

1225 42. Wang Z, Goldenfeld N. Fixed points and limit cycles in the population dynamics of
1226 lysogenic viruses and their hosts. Phys Rev E. 2010 Jul 22;82(1):011918.

1227 43. Lee YY, Erdogan A, Rao SS. How to assess regional and whole gut transit time with
1228 wireless motility capsule. J Neurogastroenterol Motil. 2014;20(2):265.

1229 44. Shao Y, Wang IN. Effect of late promoter activity on bacteriophage λ fitness.
1230 Genetics. 2009;181(4):1467–75.

1231 45. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et
1232 al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut
1233 and infects Bacteroides intestinalis. Nat Commun. 2018;9(1):1–8.

1234 46. Ramos-Barbero MD, Gómez-Gómez C, Sala-Comorera L, Rodríguez-Rubio L,
1235 Morales-Cortes S, Mendoza-Barberá E, et al. Characterization of crAss-like phage
1236 isolates highlights Crassvirales genetic heterogeneity and worldwide distribution.
1237 Nat Commun. 2023;14(1):4295.

1238 47. Parsons RJ, Breitbart M, Lomas MW, Carlson CA. Ocean time-series reveals
1239 recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso
1240 Sea. ISME J. 2012;6(2):273–84.

1241 48. Kieft K, Anantharaman K. Deciphering active prophages from metagenomes.
1242 mSystems. 2022;7(2):e00084-22.

1243 49. Ellis EL, Delbruck M. The growth of bacteriophage. J Gen Physiol. 1939;22(3):365–
1244 84.

1245 50. Hadas H, Einav M, Fishov I, Zaritsky A. Bacteriophage T4 development depends on
1246 the physiology of its host Escherichia coli. Microbiology. 1997;143(1):179–85.

1247 51. Dennehy JJ, Wang IN. Factors influencing lysis time stochasticity in bacteriophage
1248 λ. BMC Microbiol. 2011;11:1–12.

1249 52. Mahmoudabadi G, Milo R, Phillips R. Energetic cost of building a virus. Proc Natl
1250 Acad Sci. 2017;114(22):E4324–33.

1251 53. Cheng AG, Ho PY, Aranda-Díaz A, Jain S, Feiqiao BY, Meng X, et al. Design,
1252 construction, and in vivo augmentation of a complex gut microbiome. Cell.
1253 2022;185(19):3617–36.

1254 54. Wilde J, Boyes R, Robinson AV, Daisley BA, Botschner AJ, Brettingham DJL, et al.
1255 Assessing phage-host population dynamics by reintroducing virulent viruses to
1256 synthetic microbiomes. Cell Host Microbe. 2024 May;32(5):768-778.e9.

1257 55. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al.
1258 Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes
1259 of the gut microbiota in colorectal cancer. Nat Med. 2019 Jun;25(6):968–76.

1260 56. Hockenberry AJ, Wilke CO. BACPHLIP: predicting bacteriophage lifestyle from
1261 conserved protein domains. PeerJ. 2021 May 6;9:e11396.

1262 57. Shang J, Tang X, Sun Y. PhaTYP: predicting the lifestyle for bacteriophages using
1263 BERT. Brief Bioinform. 2023 Jan 19;24(1):bbac487.

1264 58. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the
1265 faecal microbiota of monozygotic twins and their mothers. Nature.
1266 2010;466(7304):334–8.

1267 59. Sutcliffe SG, Reyes A, Maurice CF. Bacteriophages playing nice: Lysogenic
1268 bacteriophage replication stable in the human gut microbiota. iScience. 2023
1269 Feb;26(2):106007.

1270 60. Silveira CB, Rohwer FL. Piggyback-the-Winner in host-associated microbial
1271 communities. Npj Biofilms Microbiomes. 2016 Jul 6;2(1):16010.

1272 61. Berngruber TW, Lion S, Gandon S. Spatial structure, transmission modes and the
1273 evolution of viral exploitation strategies. PLoS Pathog. 2015;11(4):e1004810.

1274 62. Messinger SM, Ostling A. The consequences of spatial structure for the evolution of
1275 pathogen transmission rate and virulence. Am Nat. 2009;174(4):441–54.

1276 63. Knowles B, Silveira C, Bailey B, Barott K, Cantu V, Cobián-Güemes A, et al. Lytic to
1277 temperate switching of viral communities. Nature. 2016;531(7595):466–70.

1278 64. Weitz JS, Beckett SJ, Brum JR, Cael BB, Dushoff J. Lysis, lysogeny and virus–
1279 microbe ratios. Nature. 2017 Sep 21;549(7672):E1–3.

1280 65. Oh JH, Lin XB, Zhang S, Tollenaar SL, Özçam M, Dunphy C, et al. Prophages in
1281 Lactobacillus reuteri are associated with fitness trade-offs but can increase
1282 competitiveness in the gut ecosystem. Appl Environ Microbiol. 2019;86(1):e01922-
1283 19.

1284 66. Yang K, Niu J, Zuo T, Sun Y, Xu Z, Tang W, et al. Alterations in the Gut Virome in
1285 Obesity and Type 2 Diabetes Mellitus. Gastroenterology. 2021 Oct;161(4):1257-
1286 1269.e13.

1287 67. de Jonge PA, Wortelboer K, Scheithauer TP, van den Born BJH, Zwinderman AH,
1288 Nobrega FL, et al. Gut virome profiling identifies a widespread bacteriophage family
1289 associated with metabolic syndrome. Nat Commun. 2022;13(1):3594.

1290 68. Jiang L, Lang S, Duan Y, Zhang X, Gao B, Chopyk J, et al. Intestinal Virome in
1291 Patients With Alcoholic Hepatitis. Hepatology. 2020 Dec;72(6):2182–96.

1292 69. Grafen A. The phylogenetic regression. Philos Trans R Soc B. 1989;326(1233):119–
1293 57.

1294 70. Diard M, Bakkeren E, Cornuault JK, Moor K, Hausmann A, Sellin ME, et al.
1295 Inflammation boosts bacteriophage transfer between *Salmonella* spp. Science. 2017
1296 Mar 17;355(6330):1211–5.

1297 71. Boling L, Cuevas DA, Grasis JA, Kang HS, Knowles B, Levi K, et al. Dietary
1298 prophage inducers and antimicrobials: toward landscaping the human gut
1299 microbiome. Gut Microbes. 2020;11(4):721–34.

1300 72. Hu J, Ye H, Wang S, Wang J, Han D. Prophage activation in the intestine: insights
1301 into functions and possible applications. Front Microbiol. 2021;3930.

1302 73. Shkoporov AN, Stockdale SR, Lavelle A, Kondova I, Heuston C, Upadrasta A, et al.
1303 Viral biogeography of the mammalian gut and parenchymal organs. Nat Microbiol.
1304 2022 Aug 2;7(8):1301–11.

1305 74. Wu J, Fu K, Hou C, Wang Y, Ji C, Xue F, et al. Bacteriophage defends murine gut
1306 from Escherichia coli invasion via mucosal adherence. Nat Commun. 2024 Jun
1307 4;15(1):4764.

1308 75. Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J, et al. Bacteriophage
1309 adhering to mucus provide a non–host-derived immunity. Proc Natl Acad Sci.
1310 2013;110(26):10771–6.

1311 76. Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial
1312 microbiota. Nat Rev Microbiol. 2016 Jan;14(1):20–32.

1313 77. Shalon D, Culver RN, Grembi JA, Folz J, Treit PV, Shi H, et al. Profiling the human
1314 intestinal environment under physiological conditions. Nature. 2023 May
1315 18;617(7961):581–91.

1316 78. Garmaeva S, Sinha T, Gulyaeva A, Kuzub N, Spreckels JE, Andreu-Sánchez S, et
1317 al. Transmission and dynamics of mother-infant gut viruses during pregnancy and
1318 early life. Nat Commun. 2024 Mar 2;15(1):1945.

79. McNulty R, Sritharan D, Pahng SH, Meisch JP, Liu S, Brennan MA, et al. Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. Nat Microbiol. 2023 Apr 3;8(5):934–45.

80. Bikel S, López-Leal G, Cornejo-Granados F, Gallardo-Becerra L, García-López R, Sánchez F, et al. Gut dsDNA virome shows diversity and richness alterations associated with childhood obesity and metabolic syndrome. iScience. 2021;24(8):102900.

81. Bushnell B. BBTools [Internet]. Available from: sourceforge.net/projects/bbmap/

82. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. 2012 Jul 1;40(12):e94–e94.

83. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genomics Bioinforma. 2021 Jun 23;3(3):lqab067.

84. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2023 Sep 21;

85. Millan AS, Peña-Miller R, Toll-Riera M, Halbert ZV, McLean AR, Cooper BS, et al. Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat Commun. 2014 Oct 10;5(1):5208.

86. De Paepe M, Tournier L, Moncaut E, Son O, Langella P, Petit MA. Carriage of λ Latent Virus Is Costly for Its Bacterial Host due to Frequent Reactivation in Monoxenic Mouse Intestine. Hughes D, editor. PLOS Genet. 2016 Feb 12;12(2):e1005861.

87. Acosta EM, Little KA, Bratton BP, Lopez JG, Mao X, Payne AS, et al. Bacterial DNA on the skin surface overrepresents the viable skin microbiome. eLife. 2023;12:RP87192.

88. Maurice CF, Haiser HJ, Turnbaugh PJ. Xenobiotics Shape the Physiology and Gene Expression of the Active Human Gut Microbiome. Cell. 2013 Jan;152(1–2):39–50.

89. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience. 2020 Feb 1;9(2):giaa008.