Inferring strain-level mutational drivers of phage-bacteria interaction phenotypes

Adriana Lucia-Sanz[1,^], Shengyun Peng[2,#,^], Chung Yin (Joey) Leung[3,#], Animesh Gupta[4], Justin R. Meyer[5] and Joshua S. Weitz[6,7,8,#,*]

[1] School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

[2] Adobe Inc., Palo Alto, California, USA

[3] GSK, Stevenage, Herts, United Kingdom

[4] Department of Physics, University of California San Diego, La Jolla, California, USA

[5] Department of Ecology, Behavior and Evolution, University of California San Diego, La Jolla, California, USA

[6] Department of Biology, University of Maryland, College Park, MD, USA

[7] Department of Physics, University of Maryland, College Park, MD, USA

[8] Institut d'Biologie, École Normale Supérieure, Paris, France

^ Equal contributions

Former address: School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

[*] Corresponding author, E-mail: jsweitz@umd.edu (JSW)

Abstract

The enormous diversity of bacteriophages and their bacterial hosts presents a significant challenge to predict which phages infect a focal set of bacteria. Infection is largely determined by complementary -and largely uncharacterized- genetics of adsorption, injection, and cell take-over. Here we present a machine learning (ML) approach to predict phage-bacteria interactions trained on genome sequences of and phenotypic interactions amongst 51 *Escherichia coli* strains and 45 phage λ strains that coevolved in laboratory conditions for 37 days. Leveraging multiple inference strategies and without *a priori* knowledge of driver mutations, this framework predicts both who infects whom and the quantitative levels of infections across a suite of 2,295 potential interactions. The most effective ML approach inferred interaction phenotypes from independent contributions from phage and bacteria mutations, predicting phage host range with 86% mean classification accuracy while reducing the relative error in the estimated strength of the infection phenotype by 40%. Further, transparent feature selection in the predictive model revealed 18 of 176 phage λ and 6 of 18 *E. coli* mutations that have a significant influence on the outcome of phage-bacteria interactions, corroborating sites previously known to affect phage λ infections, as well as identifying mutations in genes of unknown function not previously shown to influence bacterial resistance. While the genetic variation studied was limited to a focal, coevolved phage-bacteria system, the method's success at recapitulating strain-level infection outcomes provides a path forward towards developing strategies for inferring interactions in non-model systems, including those of therapeutic significance.

Introduction

Next-generation sequencing technology has revealed widespread diversity in microbial and viral communities [(Aylward FO, 2017), (Munson-McGee JH, 2018), (Breitbart, 2018), (Guillermo Dominguez-Huerta, 2022), (Sunagawa S., 2015) , (Nayfach, 2021), (Sunagawa, 2020)]. In parallel, the development of analytical tools to characterize species interaction networks from co-occurrence and/or time series data has led to a better understanding of microbial community structure and function [(Faust K, 2012) (Flannick J, 2006), (Stein RR, 2013), (Berry D, 2014), (Liao C., 2020), (Jiliang Hu J., 2022), (Shaer Tamar E, 2022)]. In principle, it should be possible to infer microbial interaction networks directly from genotypes and the environmental context [(Manrubia S, 2021)]. Such inference is predicated on a simple principle: adsorption is required for a bacteriophage (phage) to infect a focal bacterial strain [(Neurath AR, 1986) (Wang J H. M., 2000) (Chatterjee S, 2012), (Gaborieau B, 2023)]; such adsorption requires expression of specific cell-surface receptors (e.g., protein, lipid, carbohydrate), although in many cases the specific receptor remains unknown or modulated by poorly characterized biosynthetic pathways [(Tetz, 2022)]]. However, even if a phage adsorbs to a bacteria, there are many intracellular resistance mechanisms that could assist or inactivate phage infection altogether [(Zborowsky S., 2019), (Koonin, 2020), (Gao Z., 2023)]. Categorizing effective, extracellular adsorption and intracellular replication remains challenging. Hence, despite significant progress in linking microbial genotype to phenotype, less progress has been made with understanding the genetics of traits that influence microbial species interactions (including virus and host pairs) given the additional complication that the phenotypic output of an association may depend on the joint effects of

two separate genomes [(Bajic D, 2018), (de Jonge PA, 2019) (Buckling A, 2002) (Elena SF, 2003) (Koskella B, 2014) (Poullain V, 2008) (Kaltz O, 2002) (Beckett SJ, 2013) (Weitz JS, 2013) (Gurney J, 2017)].

The problem of understanding the genetic basis of interactions requires the development of new computational approaches to construct genotype-to-phenotype maps. Conventional approaches try to correlate phenotypic differences with genetic variation (e.g., this is true for the broad scope of work in genome-wide associated studies [(Horton MW, 2014) (D, 2016) (Power RA, 2017)]). The challenge for inferring interaction-associated phenotypes is that such interactions arise due to the combination of multiple genotypes (e.g., phage and host genotypes) leading to new combinatorial challenges. Initial steps towards interaction inference have been made through mutation-based association approaches that have successfully uncovered combinations of virus and host mutations that correlate with successful virus-host interactions [(MacPherson A, 2018) (Jallow M, 2009) (Scanlan PD, 2011), (Shaer Tamar E, 2022), (Borin JM L. J.-S., 2023)]. Conceptually, the challenge of uncovering interaction phenotypes is similar to attempts to tackle the problem of studying complex traits where gene-by-gene (G x G) interactions or gene-by-environment (G x E) interactions shape phenotypes [(Wei WH, 2014), (An P, 2009), (G, 2015) (Gupta A Z. L., 2022)].

In the case of virus-microbe systems, efforts to predict interaction phenotypes require leveraging specific system features and may depend on taxonomic scales. For example, computational approaches are increasingly used to predict the host range of viruses in a broad taxonomic sense, e.g., leveraging tetranucleotide frequencies and other sequence-specific

63 information [(Edwards RA, 2016) (Dutilh BE, 2017)]). However, predicting strain-specific
64 interactions remains a difficult task, particularly because taxonomic markers are known to be a
65 poor proxy for infection profiles [(Sullivan NJ, 2003), (Kauffman KM, 2022)]. Recent studies have
66 shown some improvement in resolving strain-specific interaction phenotypes, e.g., by using
67 CRISPR spacers and metagenomic data to identify recent phage infection[(Simon Roux, 2021),
68 (Szabo RE., 2022), (George, 2023)] or by co-clustering phage and bacteria mutations, respectively,
69 amongst strains that tend to interact as a means to identify associated gene or sequence
70 differences [(Kauffman KM, 2022)].

71 Here, we link whole genome-wide changes in phage and bacteria with observed changes
72 in interaction phenotypes using a machine learning inference framework. We do so by leveraging
73 emergent genotype and phenotype changes in coevolving populations of *Escherichia coli* B strain
74 REL606 and bacteriophage λ strain cI26 during a 37-day experiment [(Gupta A P. S., 2022)]. The
75 key idea is to recapitulate infection phenotypes from an interaction network through a
76 hierarchical regression approach without *a priori* assumptions about driver mutations or the
77 nature of genetic interactions. In contrast, prior work on microevolutionary changes in infectivity
78 have focused on changes to genes or proteins with known functions in model organisms [(Meyer
79 JR D. D., 2012) (Lobo FP, 2009) (Modi SR, 2013), (Gaborieau B, 2023)]. Such approaches are
80 dependent on the existing annotation of genes or mutations, and thus are limited by both the
81 quality and quantity of annotations. Our regression framework predicts a substantial portion of
82 phage-host infection phenotypes, including: i) who infects whom and ii) with what efficiency. In
83 doing so, we identify prioritized phage and bacterial mutations underlying changes in infection

84 phenotypes and reveal that additive effects of phage and host mutations can be sufficient to
85 predict interaction phenotypes. As we explain, this finding suggests a route to generate testable
86 hypotheses for phage and genome sites underlying interactions that could also become priority
87 targets for modification in environmental inference and the development of phage therapeutics.

88 ## Results

89 ### The mutation and cross-infection matrices for phage and bacteria

90 From a previous study [(Gupta A P. S., 2022)], we analyzed genome sequences of 50 bacterial
91 host (descended from E. *coli* B strain REL606) and 44 phage (descended from λ strain cI26) strains
92 isolated at varying time points during a 37-day coevolution experiment. For the observed
93 genotypes, the mutation profiles of the host and phage revealed many changes in their genomes,
94 including 18 and 176 unique mutations for the host and phage, respectively (Table S1). The
95 interactions of all phage-bacterial pairs including the ancestors were measured, yielding a 51 by
96 45 cross-infection matrix. Interaction strength was estimated by the efficiency that a phage
97 infected a given host compared to its ability to infect the sensitive ancestor (referred to as the
98 efficiency of plating or EOP). Additional details of the EOP calculations are described in (Gupta A
99 P. S., 2022) and Methods section “Experimental setup and data collection”. At the beginning of
100 the experiment, the isogenic host strain was susceptible to all phage strains, and by the end of
101 the experiment on day 37, most of the host isolates had evolved resistance to all phage strains.
102 A summary of the mutation profiles and the EOP matrix showing the complexity of the observed
103 phenotypes is shown in Fig 1. Based on the measurement of 2295 phage-host pairwise

104 interactions, we found 913 successful (EOP > 0) and 1382 unsuccessful (EOP = 0) phage infections.
105 The distribution of EOP values was skewed, with 95% of values ranging from 0 to 1.5, and
106 presented a long tail with a significant variability in the observed phenotypes (S1 Fig). The co-
107 occurrence of mutations in different genomic contexts (S2 Fig) suggested it might be feasible to
108 infer host and phage mutations that disproportionately impact the interaction phenotype.
109
110 **Model for predicting the phage-bacteria interaction network**
111 Initially, we developed a framework for predicting the effect that mutational profiles have on the
112 host-phage cross-infection network irrespective of the interaction strengths (e.g. EOP>0,
113 presence of infection; EOP=0, absence of infection; illustrated in Fig 2a). The underlying
114 framework utilizes a logistic regression approach to predict the presence or absence of infection
115 phenotype (referred to here as POA) from mutational 'features' (see Materials and Methods
116 corresponding section). We evaluate different models based on distinct sets of mutations that
117 support infection predictions. These include models relying solely on a linear combination of
118 mutations, either from the host or phage mutational profiles (referred to as H and P individual
119 models), as well as a model that incorporates the additive effects of phage and host mutational
120 features in a linear combination (linear model). Additionally, we consider the possibility that
121 combinations of mutations in phage and host act in combination to impact the cross-infection
122 matrix. Therefore, we incorporate a set of mutational features that account for joint effects
123 between phage and host mutations (the nonlinear model) and a model that includes both 'first-
124 order' (additive phage and host mutations) and 'second-order' (nonlinear combination of phage

and host mutations) effects (the mixed model). A comprehensive description of how each feature is constructed is provided in the Methods section “Feature construction”.

By comparing the performance of the logistic regression models built based on the different sets of features, we find that all three models that contain both phage and bacteria mutations predict the original POA phenotypes significantly better than a null model. In addition, the linear model outperforms all other models in the validation step (P < 9.44e-5) with a mean classification accuracy of ~86% (Fig 2a). This suggests that the linear model in principle contains the best set of features for predicting the POA phenotype for a given phage-host pair in this dataset. We further compared predictions of POA, and the mutational features predicted to have the largest effects on the POA for the linear, nonlinear, and mixed models (Fig 3). The results show that a linear combination of phage and host mutations can recapitulate the POA matrix without explicit inclusion of interaction effects. Mutational features identified via this method with a positive coefficient increase the probability of infection, and the opposite is true for negative coefficients. Notably, we observe that bacterial mutations are more likely to have a negative effect due to the evolution of host resistance, whereas phage mutations tend to have a positive effect, indicating selection for counter-defense traits that expand host range (see (Gupta A P. S., 2022)). Feature importance analysis (detailed in the Methods section) reveals 5 host mutations and 32 phage mutations that have a positive effect on predicting phage-host interaction network, compared with 7 host mutations and 15 phage mutations that have a negative effect (Fig 5a, S2 Table).

146 **Model for predicting the efficiency of infection**

147 We extended the prediction framework described in the prior section to identify phage and host
148 mutations that have large impacts on the efficiency of phage infection (referred to as the EFF
149 model) in the existing cross-infection network (see Methods for a detailed explanation). We used
150 log-transformed EOP values of individual infection pairs (Shapiro-Wilk test P = 3.283e-8, S3 Fig)
151 as a proxy of EFF phenotypes, while keeping the cross-interaction network fixed (Fig 4a). We
152 performed a linear regression model to quantify the impact that different sets of mutation
153 features have on EFF phenotypes. Model performances were compared based on the validation
154 mean absolute error (MAE). As in the analysis of EOP, including both phage and host mutation
155 features led to the highest performing model predictions. The linear regression model with the
156 additive feature set gives the lowest validation MAE (P < 3.95e-14) with ~40% reduction of the
157 mean error compared to the null model (Fig 2b). Next, we built linear models based on all three
158 phage and host combinations of mutational features to predict EFF phenotypes to identify
159 corresponding mutational features that have the largest impact in the predictions (Fig 4). The EFF
160 phenotypes are best predicted by a linear combination of phage and host mutation profiles.
161 Mutational features predicted by this method impact the EOP profile of the phage-host
162 interaction network (principally affecting positively or negatively the efficiency of infection).
163 Feature importance analysis identified 8 host mutations and 25 phage mutations that promote
164 the efficiency of phage infection, compared with 6 host mutations and 28 phage mutations that
165 reduce the efficiency of phage infection (Fig 5b, S3 Table).
166

Molecular mechanism behind driver mutational features

Several putatively important mutations are revealed by the feature analysis using final predictive models of POA (Fig 5a, S2 Table) and EFF (Fig 5b, S3 Table) phenotypes. We found 3 phage mutations and 1 bacterial mutation that show a significant positive effect for the POA model. For phage, these mutations include 2 nonsynonymous mutations in genes *S* and *J* and a synonymous mutation in gene *J* and for the bacteria we identified a nonsynonymous mutation in the *ccmA* gene. We also found 3 mutations in the host and 1 in the phage that have a significant negative effect in the POA model. For the bacteria, these include a nonsynonymous mutation in *ompF* and two deletions Δ777bp in *insB* and Δ141bp in *malT*; whereas for phage we identified a nonsynonymous mutation in *J* (Fig 5a).

For the EFF model, 16 mutations are predicted to have a significant effect (7 positive and 9 negative) and the majority are in phage. Of the 7 positive predicted features, only 1 is bacterial, a nonsynonymous mutation in *uup* gene. For phage, we identify 2 insertions, 1 deletion, and 1 synonymous mutation in *J* gene that should increase infectivity, another synonymous mutation in *bor* gene and a nonsynonymous mutation in the *lom* gene that increase the efficiency of infection. Whereas synonymous mutations are not expected to influence phage's ability to infect, and insertions and deletions in the *J* coding region are anticipated to have detrimental effects overall, we identified these mutations as influential to increase EFF prediction accuracy, corroborating prior work that demonstrated the impact of these mutations arising through recombination on phage fitness [(Borin JM A. S., 2021)]. Of the 9 negative predicted features, 1 is in the bacteria and 8 are in phage. The only bacterial mutation that negatively affects the EFF

188 was already identified by the POA model: the Δ777bp deletion in *insB*. For the phage we identify
189 2 different intergenic mutations with significant negative effects downstream of *lambdap79* gene;
190 3 nonsynonymous, 1 synonymous (that was positive for POA and also reported in [(Borin JM A.
191 S., 2021)]) and Δ1bp deletion mutations in *J* gene and 1 intergenic mutation between *Rz* and *bor*
192 genes (Fig 5b).

193 Our inference framework was able to recapitulate known biology without *a priori*
194 knowledge of driver mutations. We find mutations in the bacterial *malT* gene, a trans positive
195 regulator of LamB [(Debarbouille M, 1978), (Blanche S, 2013), (Maynard ND, 2010), (Banzhaf,
196 2020)], and several mutations located in the phage *J* gene region that were important for both
197 POA and EFF phenotype predictions. The *J* gene encodes the tail fiber of phage λ which is critical
198 to the process of injecting phage DNA into the host via LamB [(Wang J H. M., 2000), (Werts C,
199 1994), (Wang J M. V., 1998) (Maddamsetti R, 2018)]. Therefore, mutations in both *malT* and *J*
200 gene region are expected to impact the phage-host interaction network and the quantitative
201 efficiency of infection – consistent with our model predicting the mutations to be important for
202 both POA and EFF. A nonsynonymous mutation in the outer membrane porin OmpF, is the most
203 important feature for predicting a decrease in POA, but was not found to be important for
204 predicting EFF. This mutation is shared by 10 host strains, 2 of which were sampled from day 28
205 and 8 were from day 37. These 10 host strains were super-resistant, that is, they were resistant
206 to the ancestral phage λ strain, and all the phage isolates from the coevolution experiment.
207 Previous studies on this bacterial population showed that phage λ evolves to use OmpF as a
208 second receptor after *E. coli* evolves to down-regulate LamB [(Meyer JR D. D., 2012)]. Therefore,

this OmpF mutation is expected to confer resistance to these evolved phage λ strains and so affects the POA (host-range), but not the EFF (efficiency of infection). Similar OmpF mutations have been described to provide resistance to a related phage, phi21, after it similarly evolved to use OmpF [(Borin JM L. J.-S., 2023)]. Each model also identified mutations in *manY* which is an inner membrane transporter that enables phage λ to inject its DNA into the cytoplasm. Mutations in this protein or others in the ManXYZ complex are known to confer resistance to λ [(Erni B, 1987), (Burmeister AR, 2021), (Borin JM L. J., 2023)] and all of them impacted negatively both POA and EFF phenotypes. Most interestingly, both models were able to uncover the importance of Δ777bp deletion in *insB* by an IS element from *E. coli* which affects genes not previously identified to interact with phage λ [(Maynard ND, 2010), (Blanche S, 2013)], but was recently identified to confer resistance through epistasis with other resistance mutation in *malT* through an unknown mechanism [(Gupta A P. S., 2022)]. This illustrates the capability of our machine learning approach to identify candidate, pivotal genes involved in phage-host interactions.

Discussion

In this study, we developed a machine learning framework leveraging hierarchical logistic regression to predict the network and efficiency of phage-bacteria interactions by linking infection phenotypes with genetic mutation profiles of both phage and bacterial host. The basis for our inference was an assumption that mutations can contribute directly or via gene-gene interactions to changes in the infection phenotype. Our comparative analysis revealed that a model that incorporates additive mutational effects of phage and host separately had the highest

predictive value in inferring phenotype from genotype. In doing so, the framework identified gene regions already recognized in mediating the efficiency of infection for bacteriophage λ and *E. coli* [(Meyer JR D. D., 2012) (Blanche S, 2013), (Burmeister AR, 2021), (Gupta A Z. L., 2022)] and predicted mutations that conferred a resistant phenotype in bacteria through epistasis with other mutations (Gupta et al., 2022). The model also identified features that were located in phage gene *J* region, including a number of synonymous mutations as well as insertions and deletions that in principle should be detrimental, but have been shown to modulate host-range expansion and counter-defense through recombination [(Borin JM A. S., 2021)]. Hence, the framework has the potential to identify novel genes and mutations that modulate both qualitative and quantitative features of virus-microbe interactions while being cognizant of the potential for the framework to erroneously also identify hitchhiking mutations as driver mutations when they are likely proxies for adjacent driver mutants linked via recombination.

Based on the feature importance analysis, we identified one mutation located in the phage *S* gene region that is found to be uniquely important for predicting the presence (or absence) of infection. This gene encodes the holin which is a small inner membrane protein required for phage-induced host lysis [(Chang CY, 1995)]. Notably, the phage-host interaction network observed in our experiment is based on the quantitative plaque assay, in which clearings (plaques) would appear where bacterial cells were infected and lysed by the phage [(Anderson B, 2011), (Sambrook J, 2006)]. Thus, we interpret the feature analysis to imply that a mutation in the S gene has a direct impact on the lysis of the host cells, which would then have an impact on the final observed phenotype. Similar mutations were uncovered via experimental evolution to

counteract a gene deletion in the host that helps facilitate phage DNA replication [(Gupta A S. A., 2020)]. This mutation may extend the infection process and allow the phage more time to initiate DNA replication in the debilitated host, increasing the chance of a successful infection. We hypothesize that this mutation may have a similar function to counteract host mutations that interfere with λ's lytic life cycle. Another mutation identified by our method in the phage *lom* gene region was exclusively important in positively modulating infection efficiency but not the interaction itself; we note that this site was previously reported to increase phage resistance through an unknown mechanism [(Borin JM A. S., 2021)].

The model selection procedure identified an additive model as the best predictor of interaction phenotype from phage and bacterial genotype. In the additive model, individual phage and bacterial mutations act independently, rather than synergistically (whether positively or negatively), to determine infection outcome. Hence complex interaction networks may be (partially) predictable based on direct effects rather than relying on direct inference of complex interactive effects that are more challenging to measure [(Shaer Tamar E, 2022)]. Nonetheless, it is important to note that this result may reflect the nature of our training and test sets, and might be limited by sampling, and does not exclude the possibility that higher order gene-gene interactions affect infection phenotypes. The number of phage-host mutation pairs scales as the product of the number of phage and host mutations in higher order models (nonlinear and mixed models), but most of these combinations were not observed in our strains. In essence, fitting higher order models leads to underdetermined systems even with the introduction of regularization terms meant to limit the number of weak contributions from mutations – whether

direct or in combination. Future work would have to significantly scale-up genotyped combinations of overlapping mutations in different contexts to robustly infer phage-bacteria interaction mutational pairs.

Our inference framework was able to detect the importance of previously identified adaptive mutations that modify phage-host interactions. Although false positives and false negatives are possible, we note that evolutionary effects including genetic hitchhiking and recombination may move adaptive mutations onto different backgrounds, improving detection of driver mutations of infection. We did not expect the identification of adaptive mutations to be comprehensive. Instead, by linking genotype to phenotypic changes as measured by a subset of phage and host isolates that arose via coevolution, we can identify mutations of potential relevance to infection (and fitness) in an ecologically relevant context even if significant regimes of mutational space are left unexplored.

In summary, we have developed a framework for predicting genotypic drivers of both the qualitative and quantitative nature of host-pathogen interactions. In doing so, we recapitulated the finding of mutations known to influence infection outcome as well as identified novel sites. Moving forward, this framework could help prioritize research on identifying novel drivers of infection, focusing efforts on mutations with highest absolute values and those most likely to alter the phenotype (primarily nonsynonymous mutations). Although we applied this framework in the context of experimental phage-bacteria coevolution and with relatively low genetic diversity, this data-driven approach does not require *a priori* knowledge of driver genes and mutations and could be applied to other, even poorly characterized, phage-bacteria systems. As

such, we expect this approach will be relevant in improving understanding of interactions in natural systems as well as for phages that target bacterial pathogens.

Materials and Methods

Experimental setup and data collection

We analyzed data from Gupta et al., 2022 where *E. coli* B strain REL606 and phage λ strain cI26 were cocultured for a 37-day period. Samples were taken on checkpoint days for pairwise quantitative plaque assays as described in (Gupta A P. S., 2022). The EOP value measures the efficiency of a phage infecting a derived host strain relative to that for infecting the ancestral strain. The EOP value for a phage, j, infecting a host, i, is computed as

$$e_{ij} = \frac{q_{(i,j)}}{q_{(anc,j)}} \times d^{s_{(i,j)} - s_{(anc,j)}}, \quad (1)$$

where $q_{(i,j)}$ is the number of plaques for phage j against host i, $q_{(anc,j)}$ is the number of plaques for phage j against the ancestral host strain, $s_{(i,j)}$ is the number of dilutions performed to observe distinguishable and countable clear plaques for phage j against host i, $s_{(anc,j)}$ is the number of dilutions performed to observe distinguishable and countable clear plaques for phage j against the ancestral host strain and d is the dilution ratio which is 5 in our experiment. A positive EOP value from the cross-infection plaque assay indicates a successful infection event for a given phage-host pair. In contrast, a zero EOP value indicates the phage has no capacity to infect. A larger EOP value from the cross-infection plaque assay indicates that the phage can infect a given host more efficiently than the ancestral host strain.

313 For each phage and host samples taken from each checkpoint, the DNA extraction, library
314 preparation and sequencing experiment was performed as described in (Gupta A P. S., 2022).
315 Mutation profiles based on the genome sequencing data were constructed using *breseq* as
316 described in (Gupta A P. S., 2022). In addition to the mutations revealed by *breseq* results, for
317 both host and phage we created an artificial mutation as the indicator for the ancestral strain to
318 add the ancestral strain into the mutation profile table. For this artificial mutation, only the
319 ancestral strain is indicated to have this mutation. All other strains were shown to not have this
320 mutation in the mutation profile table.
321
322 **Feature construction**
323 For a total number of U host samples and V phage samples, we denote the EOP value for the i-th
324 host against j-th phage as e_{ij} where $i \in [1, U]$ and $j \in [1, V]$. Let N be the total number of
325 unique mutations observed for the host and M be the total number of unique mutations
326 observed for the phage, the host mutation profile H is a matrix of dimension U by N, and the
327 phage mutation profile P is a matrix of dimension V by M. Let h_{il} be an element from H, then
328 $h_{il} = 1$ corresponds to the presence of the l-th mutation in the i-th host whereas $h_{il} = 0$
329 corresponds to the absence of the l-th mutation in the ith host. Similarly, let p_{jk} be an element
330 from P, then $p_{jk} = 1$ corresponds to the presence of the k-th mutation in j-th phage whereas
331 $p_{jk} = 0$ corresponds to the absence of the k-th mutation in the j-th phage.

Five sets of features were constructed based on the mutation profiles of the host and phage. The H-only model is constructed based on a linear combination of 'host only' mutation profiles. The H-only model, denoted as $\phi_{ij}^{(1)}$, can be represented as:

$$\phi_{ij}^{(1)} = \gamma_1 + \sum_{l=1}^{N} \alpha_l h_{il} \,, \qquad (2)$$

where γ_1 represents a scalar of the bias term and α_l is the coefficient for the l-th host mutation. γ_1 and α_l will be learned from the model. The H-only model can also be represented in matrix form as:

$$\Phi^{(1)} = \Gamma_1 + H \cdot R_\alpha, \qquad (3)$$

where Γ_1 is a U by V matrix by repeating γ_1, i.e. $\Gamma_1 = [\gamma_1]_{U\times V}$, R_α is a N by V matrix by stacking the same coefficient vector α horizontally, i.e. $[\alpha|\alpha|\cdots|\alpha|\alpha]_{N\times V}$.

The P-only model is constructed based on a linear combination of 'phage only' mutational profiles. The P-only model, denoted as $\phi_{ij}^{(2)}$, can be represented as:

$$\phi_{ij}^{(2)} = \gamma_2 + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk}, \qquad (4)$$

where γ_2 represents a scalar of the bias term and $\tilde{\alpha}_k$ is the coefficient for the k-th phage mutation. γ_2 and $\tilde{\alpha}_k$ will be learned from the model. This model can also be represented in matrix form as:

$$\Phi^{(2)} = \Gamma_2 + [P \cdot R_{\tilde{\alpha}}]^T \,, \qquad (5)$$

where Γ_2 is a U by V matrix by repeating γ_2 and $R_{\tilde{\alpha}}$ is a M by U matrix by stacking the same coefficient vector $\tilde{\alpha}$ horizontally, i.e. $[\tilde{\alpha}|\tilde{\alpha}|\cdots|\tilde{\alpha}|\tilde{\alpha}]_{M\times U}$.

The linear model, denoted as $\phi_{ij}^{(3)}$, utilizes a linear combination of phage and host mutational features and can be represented as:

$$\phi_{ij}^{(3)} = \gamma_3 + \sum_{l=1}^{N} \alpha_l h_{il} + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk}, \quad (6)$$

where γ_3 represents a scalar of the bias term, α_l is the coefficient for the l-th host mutation and $\tilde{\alpha}_k$ is the coefficient for the k-th phage mutation. γ_3, α_l and $\tilde{\alpha}_k$ will be learned from the model. The linear model can also be represented in matrix form as:

$$\Phi^{(3)} = \Gamma_3 + H \cdot R_\alpha + [P \cdot R_{\tilde{\alpha}}]^T, \quad (7)$$

where Γ_3 is a U by V matrix by repeating γ_3, i.e. $\Gamma_3 = [\gamma_3]_{U\times V}$, R_α is a N by V matrix by stacking the same coefficient vector α horizontally, i.e. $[\alpha|\alpha|\cdots|\alpha|\alpha]_{N\times V}$ and $R_{\tilde{\alpha}}$ is a M by U matrix by stacking the same coefficient vector $\tilde{\alpha}$ horizontally, i.e. $[\tilde{\alpha}|\tilde{\alpha}|\cdots|\tilde{\alpha}|\tilde{\alpha}]_{M\times U}$. The assumption for the linear model is that the impact of mutations from both the phage and host have additive effects on the observed outcome.

The nonlinear model, denoted as $\phi_{ij}^{(4)}$, utilizes nonlinear combination of phage and host mutational features as the input and can be represented as:

$$\phi_{ij}^{(4)} = \gamma_4 + \sum_{l=1}^{N} \sum_{k=1}^{M} \beta_{lk} h_{il} p_{jk}, \quad (8)$$

where γ_4 represents a scalar of the bias term, β_{lk} denotes the coefficient for the l-th host mutation and k-th phage mutation in the corresponding i-th host and j-th phage pair. γ_4 and β_{lk} will be learned from the model. This nonlinear model can also be represented in the matrix form as:

$$\Phi^{(4)} = \Gamma_4 + H \cdot B \cdot P^T, \quad (9)$$

where Γ_4 is a U by V matrix by repeating γ_4, i.e. $\Gamma_4 = [\gamma_4]_{U\times V}$, B is the N by M coefficient matrix. The assumption for the nonlinear model is that the impact of the genetic mutations on the observed outcome comes from the additive effects of co-occurring phage-host mutation pairs. In other words, $h_{il}p_{jk} = 1$ only when both the host i has mutation l and phage j has mutation k.

Based on the formulation of the linear and nonlinear models, it is natural to combine both effects to get a more sophisticated input feature, by adding up both effects. The mixed model, denoted as $\phi_{ij}^{(5)}$, utilizes a mixed combination of linear and nonlinear effects of host and phage mutation features and can be represented as:

$$\phi_{ij}^{(5)} = \gamma_5 + \sum_{l=1}^{N} \alpha_l h_{il} + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk} + \sum_{l=1}^{N}\sum_{k=1}^{M} \beta_{lk} h_{il} p_{jk} \,. \qquad (10)$$

The matrix form of the mixed model is:

$$\Phi^{(5)} = \Gamma_5 + H \cdot R_\alpha + [P \cdot R_{\tilde{\alpha}}]^T + H \cdot B \cdot P^T \,, \qquad (11)$$

where Γ_5 is a U by V matrix by repeating γ_5, i.e. $\Gamma_5 = [\gamma_5]_{U\times V}$.

Framework design

We designed a framework comprised of two types of predictions. First, we designed a framework that predicts the phage-host cross interaction network (i.e., the phage host range). This model tries to find the set of features that can best distinguish between successful (EOP > 0) and unsuccessful (EOP = 0) infections using classification models. Second, we built a framework to predict the strength of the interaction of the subset of phage-ho pairs where the host is susceptible to the phage (EOP > 0). This model of our framework is designed to evaluate the

potential impact of the genotype on this observed phenotype by modeling the efficiency of the phage in infecting a host.

Model for predicting phage host cross-infection network (POA)

In order to determine the presence or absence of a successful infection event for a phage-host pair, we binarized the EOP values e_{ij} into 0 and 1, i.e.

$$d_{ij} = 1_{\{e_{ij}>0\}}, \qquad (12)$$

where $d_{ij} = 0$ indicates a failure of the infection and $d_{ij} = 1$ indicates success. Here we used logistic regression to model the relationship between mutation profiles and the existence of successful infection in phage-host pairs, that is:

$$\phi_{ij}^{(\cdot)} = ln\left(\frac{d_{ij}}{1-d_{ij}}\right). \qquad (13)$$

Each of the five sets of features, namely H-only, P-only, linear, nonlinear and mixed, were used as the input features for the models $\phi_{ij}^{(1)}$, $\phi_{ij}^{(2)}$, $\phi_{ij}^{(3)}$, $\phi_{ij}^{(4)}$ and $\phi_{ij}^{(5)}$ respectively. In practice, we used LASSO for feature selection and regularization. The penalty term parameter for LASSO was determined by using 10-fold cross-validation on the training data. The prediction classification error, $\frac{FalsePositives+FasleNegatives}{TestSamples}$, was used to assess the performance for this model. The mean classification error was calculated by taking the mean of classification error from 200 runs.

Model for predicting infection efficiency (EFF)

We applied a log transformation on the positive EOP values to normalize the distribution. For a given phage-host pair where a successful infection event is present, that is $e_{ij} > 0$, we denote the natural log transformed EOP value as:

$$e'_{ij} = ln\left(e_{ij}\right). \quad (14)$$

Shapiro-Wilk test was performed to check the normality of the distribution of e'_{ij}.

Linear regression was used to model the relationship between mutation profiles and the intensity of successful infections in phage-host pairs, that is:

$$\phi_{ij}^{(\cdot)} = e'_{ij}. \quad (15)$$

Each of the five sets of features, namely H-only, P-only, linear, nonlinear and mixed, were used as the input features for the models $\phi_{ij}^{(1)}$, $\phi_{ij}^{(2)}$, $\phi_{ij}^{(3)}$, $\phi_{ij}^{(4)}$ and $\phi_{ij}^{(5)}$ respectively. For the linear model, we also used LASSO for feature selection and regularization. The penalty term parameter for LASSO was determined by using 10-fold cross-validation on the training data. Finally, the MAE was used to evaluate the performance of the model.

Train-validation split and feature evaluation

To assess the performance of different features for the logistic regression model, we performed 200 bootstrap runs to predict the existence of phage infection. Specifically, in each run the training set was generated by randomly select $U \times V$ samples from the entire dataset with replacement. The d_{ij} values that were not selected as training samples form the validation set. As a control, for each run, a null model was built to predict the outcomes by randomly sample d_{ij} values from a Bernoulli distribution $Bern(\hat{p})$ where $\hat{p}$ is the maximum likelihood estimator

(MLE) of the proportion of successful infection from the training set of that run. After the 200 runs, the training and validation prediction error were compared between pairs of the models including the null model and models based on phage and host mutations only and linear, nonlinear, and mixed combinations of phage and host mutational features.

Similarly, we also performed 200 bootstrap runs for the linear model to predict the infection efficiency. Specifically, in each run the training set was generated by randomly sample e'_{ij} with replacement. The size of e'_{ij} sampled as the training set in each run matches the total number of the e'_{ij}. The e'_{ij} that were not selected in the training set forms the validation set. As a control, for each run, a null model was built by always predicting the efficiency of infection as the mean e'_{ij} of the training set for that run. After the 200 runs, the training and validation MAEs were compared between pairs of the models including the null model and every feature model set.

Final predictions and feature important analysis

After comparing the training and validation performance of models based on the different mutational sets with 200 bootstrap runs, a final model, that integrates predictions of POA and EFF was constructed. The penalty term parameter for each of the prediction frameworks was chosen as the mean of the best penalty term parameter from each of the 200 bootstrap runs. After model fitting, the predicted outcome $\tilde{d}_{ij}$ for step 1 and $\tilde{e}'_{ij}$ for step 2. For each step of the final models, the importance of feature was measured by the absolute value of coefficients learned from each step.

Author Contributions

Conceptualization: JSW, CYL & JRM

Methodology: ALS, SP, CYL & JSW

Investigation: ALS, SP, CYL, AG

Visualization: ALS, SP

Writing – original draft: ALS, SP & JSW

Writing – review & editing: ALS, SP, CYL, AG, JRM, JSW

Software Availability

https://github.com/aluciasanz/genotype_to_phenotype_inference_model.

Acknowledgments

JSW - Army Research Office (W911NF1910384), NIH (R01-AI146592-01), Simons Foundation (930283), Chaires Blaise Pascal of the Île-de-France region.

JRM - Howard Hughes Medical Institute Emerging Pathogens Initiative grant 7012574.

References

An P, M. O. (2009). The challenge of detecting epistasis (G x G interactions): Genetic Analysis Workshop 16. *Genet Epidemiol. 2009;*, 33 Suppl 1:S58-67. doi: 10.1002/gepi.20474.

Anderson B, R. M. (2011). Enumeration of bacteriophage particles: Comparative analysis of the traditional plaque assay and real-time QPCR- and nanosight-based assays. *Bacteriophage*, 1(2):86-93.

Aylward FO, B. D.-C. (2017). Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proc Natl Acad Sci U S A*, 114(43):11446-51. Epub 2017/10/27. doi: 10.1073/pnas.1714821114. P.

Bajic D, V. J. (2018). On the deformability of an empirical fitness landscape by microbial evolution. *Proc Natl Acad Sci USA*, 115(44): 11286-11291. https://doi.org/10.1073/pnas.1808485115.

Banzhaf, W. C. (2020). Subtle Environmental Differences have Cascading Effects on the Ecology and Evolution of a Model Microbial Community. In W. C. Banzhaf, *Evolution in Action: Past, Present and Future: A Festschrift in Honor of Erik D. Goodman* (pp. 273 – 288). Springer International Publishing.

Beckett SJ, W. H. (2013). . Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks. *Interface Focus*, 3(6):20130033. doi: 10.1098/rsfs.2013.0033.

Berry D, W. S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front Microbiol*, ;5:219. doi: 10.3389/fmicb.2014.00219.

Blanche S, W. S. (2013). The Protein Interaction Network of Bacteriophage lambda with its Host Escherichia coli. *J Virol*, 87(23), https://doi.org/10.1128/JVI.02495-13.

Borin JM, A. S. (2021). Coevolutionary phage training leads to greater bacterial suppression and delays the evolution of phage resistance. *Proc. Natl. Acad. Sci*, 118 (23): e2104592118.

Borin JM, L. J. (2023). Comparison of bacterial suppression by phage cocktails, dual-receptor generalists, and coevolutionarily trained phages. *Evolutionary Applications*, 16, 152–162.

Borin JM, L. J.-S. (2023). Rapid bacteria-phage coevolution drives the emergence of multi-scale networks. *Science*, 382,674-678.

Breitbart, M. B. (2018). Phage puppet masters of the marine microbial realm. *Nature Microbiology 3*, 754–766.

Buckling A, R. P. (2002). Antagonistic coevolution between a bacterium and a bacteriophage. *Proc Biol Sci*, 269(1494):931-6. doi: 10.1098/rspb.2001.1945.

Burmeister AR, S. R. (2021). Sustained coevolution of phage lambda and Escherichia coli involves inner-as well as outer-membrane defences and counter-defences. *Microbiology*, 167:001063. doi: 10.1099/mic.0.001063.

Chang CY, N. K. (1995). S gene expression and the timing of lysis by bacteriophage lambda. . *J Bacteriol*, 177(11):3283-94.

Chatterjee S, R. E. (2012). Interaction of bacteriophage lambda with its E. coli receptor, LamB. *Viruses*, 4(11):3162-78. Epub 2012/12/04. doi: 10.3390/v4113162.

D, F. (2016). Bacterial genomics: Microbial GWAS coming of age. *Nat Microbiol*, 1:16059. doi: 10.1038/nmicrobiol.2016.59.

de Jonge PA, N. F. (2019). Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends Microbiol*, 27(1):51-63 doi: 10.1016/j.tim.2018.08.006.

Debarbouille M, S. H. (1978). Dominant constitutive mutations in malT, the positive regulator gene of the maltose regulon in Escherichia coli. *J Mol Biol*, 124: 359–371.

Dutilh BE, R. A. (2017). Virus Discovery by Metagenomics: The (Im)possibilities. *Front Microbiol*, 8:1710. doi: 10.3389/fmicb.2017.01710.

Edwards RA, M. K. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev*, ;40(2):258-72. doi: 10.1093/femsre/fuv048.

Elena SF, L. R. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation . *Nat Rev Genet*, 4(6):457-69. doi: 10.1038/nrg1088.

Erni B, Z. B. (1987). he mannose permease of Escherichia coli consists of three different proteins. Amino acid sequence and function in sugar transport, sugar phosphorylation, and penetration of phage lambda DNA. *Journal of Biological Chemistry*, 262 (11): 5238 - 5247.

Faust K, R. J. (2012). Microbial interactions: from networks to models. *Nat Rev Microbiol.*, 10(8):538-50. doi: 10.1038/nrmicro2832. PubMed PMID: 22796884.

Flannick J, N. A. (2006). Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res*, 16(9):1169-81. Epub 2006/08/11. doi: 10.1101/gr.5235706.

G, G. (2015). *A primer of human genetics.* Sinauer Associates is an imprint of Oxford Univesity Press.

Gaborieau B, V. H. (2023). Predicting phage-bacteria interactions at the strain level from genomes. *bioRxiv*, https://doi.org/10.1101/2023.11.22.567924 .

Gao Z., a. F. (2023). Bacteriophage strategies for overcoming host antiviral immunity . *Front. Microbiol. 14*, 1211793.

George, N. H. (2023). CRISPR-resolved virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts. *Sci Rep* , 13, 5611. https://doi.org/10.1038/s41598-023-32078-6.

Guillermo Dominguez-Huerta, A. A. (2022). Diversity and ecological footprint of Global Ocean RNA viruses. *Science 376,*, 1202-1208.

Gupta A, P. S. (2022). Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing. *Ecol Lett*, 25(4):876-888. doi: 10.1111/ele.1.

Gupta A, S. A. (2020). Bacteriophage lambda overcomes a perturbation in its host–viral genetic network through mutualism and evolution of life history traits. *Evolution*, 74 (4): 764–774.

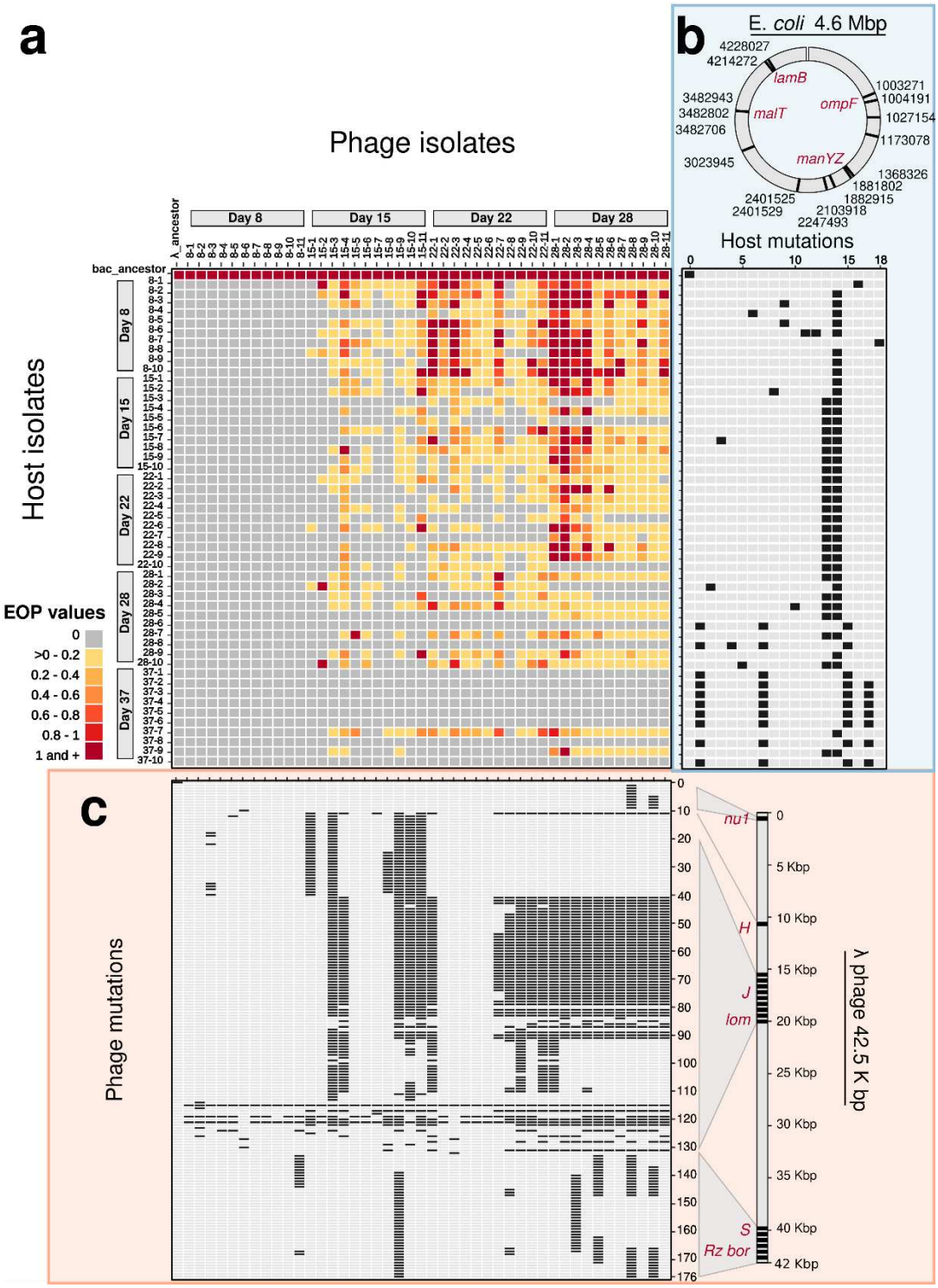
Gupta A, Z. L. (2022). Host-parasite coevolution promotes innovation through deformations in fitness landscapes. *eLife*, 11:e76162.

549 Gurney J, A. L.-B. (2017). Network structure and local adaptation in co-evolving bacteria-phage
550 interactions. *Mol Ecol*, 26(7):1764-77. doi: 10.1111/mec.14008.
551 Horton MW, B. N. (2014). Genome-wide association study of Arabidopsis thaliana leaf microbial
552 community. *Nat Commun*, 5:5320. doi: 10.1038/ncomms6320.
553 Jallow M, T. Y. (2009). Genome-wide and fine-resolution association analysis of malaria in West
554 Africa. *Nat Genet*, 41(6):657-65. doi: 10.1038/ng.388.
555 Jiliang Hu J., A. D. (2022). Emergent phases of ecological diversity and dynamics mapped in
556 microcosms. *Science 378*, 85-89.
557 Kaltz O, S. J. (2002). Within-and among-population variation in infectivity, latency and spore
558 production in a host–pathogen system. *Journal of Evolutionary Biology*, 15(5):850-60.
559 Kauffman KM, C. W. (2022). Resolving the structure of phage–bacteria interactions in the
560 context of natural diversity. *Nat Commun*, 13, 372. https://doi.org/10.1038/s41467-021-
561 27583-z.
562 Koonin, E. M. (2020). Evolutionary entanglement of mobile genetic elements and host defence
563 systems: guns for hire. *Nat Rev Genet 21*, 119–131.
564 Koskella B, B. M. (2014). Bacteria-phage coevolution as a driver of ecological and evolutionary
565 processes in microbial communities. *FEMS Microbiol Rev*, 38(5):916-31. doi:
566 10.1111/1574-6976.12072.
567 Liao C., W. T. (2020). Modeling microbial cross-feeding at intermediate scale portrays
568 community dynamics and species coexistence. *PLOS Computational Biology 16(8)*,
569 e1008135.
570 Lobo FP, M. B. (2009). Virus-host coevolution: common patterns of nucleotide motif usage in
571 Flaviviridae and their hosts. *PLoS One*, 4(7):e6282. doi: 10.1371/journal.pone.0006282.
572 MacPherson A, O. S. (2018). Keeping Pace with the Red Queen: Identifying the Genetic Basis of
573 Susceptibility to Infectious Disease. *Genetics*, 208(2):779-89. doi:
574 10.1534/genetics.117.300481.
575 Maddamsetti R, J. D. (2018). Gain-of-function experiments with bacteriophage lambda uncover
576 residues under diversifying selection in nature. *Evolution*, 72: 2234-2243.
577 Manrubia S, C. J.-U.-M. (2021). From genotypes to organisms: state-of-the-art and perspectives
578 of a cornerstone in evolutionary dynamics. *Phys Life Rev*, 38: 55-106. doi:
579 10.1016/j.plrev.2021.03.004.
580 Maynard ND, B. E. (2010). A Forward-Genetic Screen and Dynamic Analysis of Lambda Phage
581 Host-Dependencies Reveals an Extensive Interaction Network and a New Anti-Viral
582 Strategy. *PLOS Genetics* , 6(7): e1001017.
583 https://doi.org/10.1371/journal.pgen.1001017.
584 Meyer JR, D. D. (2012). Repeatability and contingency in the evolution of a key innovation in
585 phage lambda. *Science*, 335(6067):428-32. doi: 10.1126/science.1214449.
586 Meyer JR, D. D. (2012). Repeatability and contingency in the evolution of a key innovation in
587 phage lambda . *Science*, 335 (6067): 428-432.
588 Modi SR, L. H. (2013). Antibiotic treatment expands the resistance reservoir and ecological
589 network of the phage metagenome. *Nature*, ;499(7457):219-22. doi:
590 10.1038/nature12212.

591 Munson-McGee JH, P. S. (2018). A virus or more in (nearly) every cell: ubiquitous networks of
592 virus-host interactions in extreme environments. *ISME J*, 12(7):1706-14. Epub
593 2018/02/23. doi: 10.1038/s41.
594 Nayfach, S. P.-E. (2021). Metagenomic compendium of 189,680 DNA viruses from the human
595 gut microbiome. *Nat Microbiol 6*, 960-970.
596 Neurath AR, K. S. (1986). Identification and chemical synthesis of a host cell receptor binding
597 site on hepatitis B virus. *Cell*, 46(3):429-36. Epub 1986/08/01. PubMed PMID: 3015414.
598 Poullain V, G. S. (2008). The evolution of specificity in evolving and coevolving antagonistic
599 interactions between a bacteria and its phage. *Evolution*, ;62(1):1-11. doi:
600 10.1111/j.1558-5646.2007.00.
601 Power RA, P. J. (2017). Microbial genome-wide association studies: lessons from human GWAS.
602 *Nat Rev Genet*, 18(1):41-50. doi: 10.1038/nrg.2016.132.
603 Roux S., P.-E. D. (2021). IMG/VR v3: an integrated ecological and evolutionary framework for
604 interrogating genomes of uncultivated viruses. *Nucleic Acids Research*, 49(1):764–775.
605 Sambrook J, R. D. (2006). *The Condensed Protocols from Molecular Cloning: A Laboratory*
606 *Manual.* N.Y.: Cold Spring Harbor Laboratory Press.
607 Scanlan PD, H. A.-P. (2011). Genetic basis of infectivity evolution in a bacteriophage. *Mol Ecol*,
608 20(5):981-9. Epub 2010/11/16. doi: 10.1111/j.1365-294X.2010.04903.x.
609 Shaer Tamar E, K. R. (2022). Multistep diversification in spatiotemporal bacterial-phage
610 coevolution. *Nat Commun* , 13, 7971.
611 Simon Roux, D. P.-E.-M. (2021). IMG/VR v3: an integrated ecological and evolutionary
612 framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research*,
613 49(1): 764-775.
614 Stein RR, B. V. (2013). Ecological Modeling from Time-Series Inference: Insight into Dynamics
615 and Stability of Intestinal Microbiota. *PLOS Computational Biology 9 (12)*, e1003388.
616 Sullivan NJ, G. T. (2003). Accelerated vaccination for Ebola virus haemorrhagic fever in non-
617 human primates. *Nature*, 424(6949):681-4. doi: 10.1038/nature01876.
618 Sunagawa S., C. L. (2015). Structure and function of the global ocean microbiome. *Science, 348*,
619 1261359.
620 Sunagawa, S. A. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nat Rev*
621 *Microbiol 18*, 428–445.
622 Szabo RE., P. S. (2022). Historical contingencies and phage induction diversify bacterioplankton
623 communities at the microscale. *Proc. Natl. Acad. Sci 119 (30)*, e211774811.
624 Tetz, V. T. (2022). Novel prokaryotic system employing previously unknown nucleic acids-based
625 receptors. *Microb Cell Fact 21*, 202.
626 Vica Pacheco S, G. G. (1997). The lom gene of bacteriophage lambda is involved in Escherichia
627 coli K12 adhesion to human buccal epithelial cells. *FEMS Microbiol Lett.* , 156(1):129-32.
628 Wang J, H. M. (2000). The C-terminal portion of the tail fiber protein of bacteriophage lambda is
629 responsible for binding to LamB, its receptor at the surface of Escherichia coli K-12. *J*
630 *Bacteriol*, 182(2):508-12.

631 Wang J, M. V. (1998). Cloning of the J gene of bacteriophage lambda, expression and
632 solubilization of the J protein: first in vitro studies on the interactions between J and
633 LamB, its cell surface receptor. *Res Microbiol*, 149(9):611.
634 Wei WH, H. G. (2014). Detecting epistasis in human complex traits. *Nat Rev Genet*, 15(11):722-
635 33. doi: 10.1038/nrg3747.
636 Weitz JS, P. T. (2013). Phage-bacteria infection networks. *Trends Microbiol*, 21(2):82-91. doi:
637 10.1016/j.tim.2012.11.00.
638 Werts C, M. V. (1994). Adsorption of bacteriophage lambda on the LamB protein of Escherichia
639 coli K-12: point mutations in gene J of lambda responsible for extended host range. *J*
640 *Bacteriol*, 176(4):941-7.
641 Zborowsky S., a. D. (2019). Resistance in marine cyanobacteria differs against specialist and
642 generalist cyanophages . *Proc. Natl. Acad. Sci. 116 (34)* , 16899-16908.
643

Figures



646 **Fig 1. Phage-bacteria cross-infection matrix and mutation profiles.** (a) Cross-infection matrix,
647 including host and phage ancestor strains, and 50 bacteria (rows) and 44 phage (columns) strains
648 isolated during 37-day coevolution experiment (day of isolation indicated). Names correspond to
649 "day of isolation – number of isolate". Colored cells are EOP values of infection as in legend, grey
650 cells indicate no infection. (b-c) Mutation profiles for each isolate (positions mutated are in black
651 and in grey otherwise) for 18 (host) and 127 (phage) found mutations numbered in sequential
652 order of appearance in the corresponding genome. (b, in blue) Host isolates (rows) and mutation
653 profiles (columns) for 1 to 18 unique mutations found in nt position 1,003,271 to 4,228,027 of
654 the E. *coli* genome (c, in orange) Phage isolates (columns) and mutation profiles (rows) for 1 to
655 127 unique mutations found in nt position 175 to 42,491 of the λ phage genome. For the
656 complete list of host and phage mutations see S1 Table. Important genes for phage-host
657 interaction are highlighted in red and discussed in the main text.
658

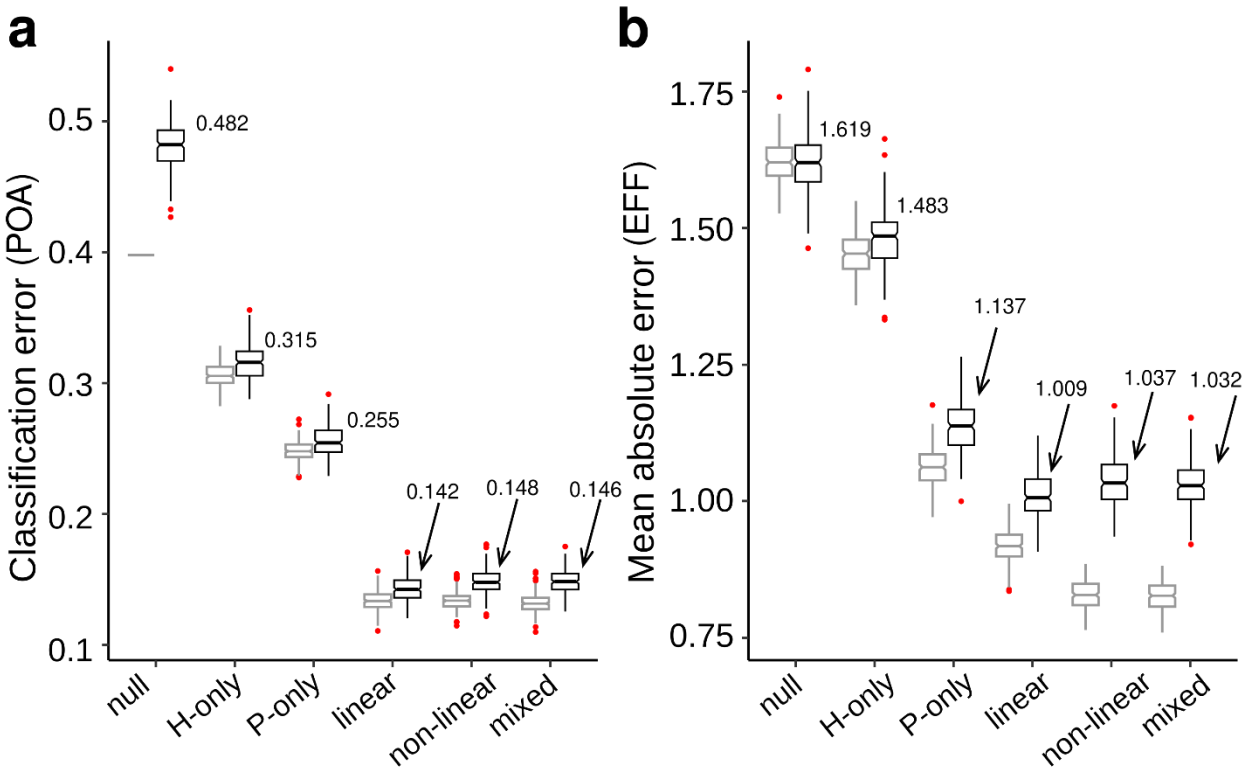


Fig 2. Model performances for different feature sets. The lowest mean value in the validation set for POA and EFF models corresponds to the linear model. (a) Classification error distributions in the training (grey) and validation (black) sets for the predictions of the phage-host interaction network (POA) (ANOVA post hoc Tukey $p<0.01$). The lowest mean value in the validation set corresponds to the linear model (b) Mean absolute error distributions in the training (grey) and validation (black) sets for the predictions of efficiency of infection (EFF) (ANOVA post hoc Tukey $p<0.001$, comparing different mutation feature models and a null model. Boxplots contain 25^{th}-75^{th} percentiles, whiskers indicate minimum and maximum values, middle lines are the median (value indicated) of 200 bootstrap runs. Red dots are outliers.

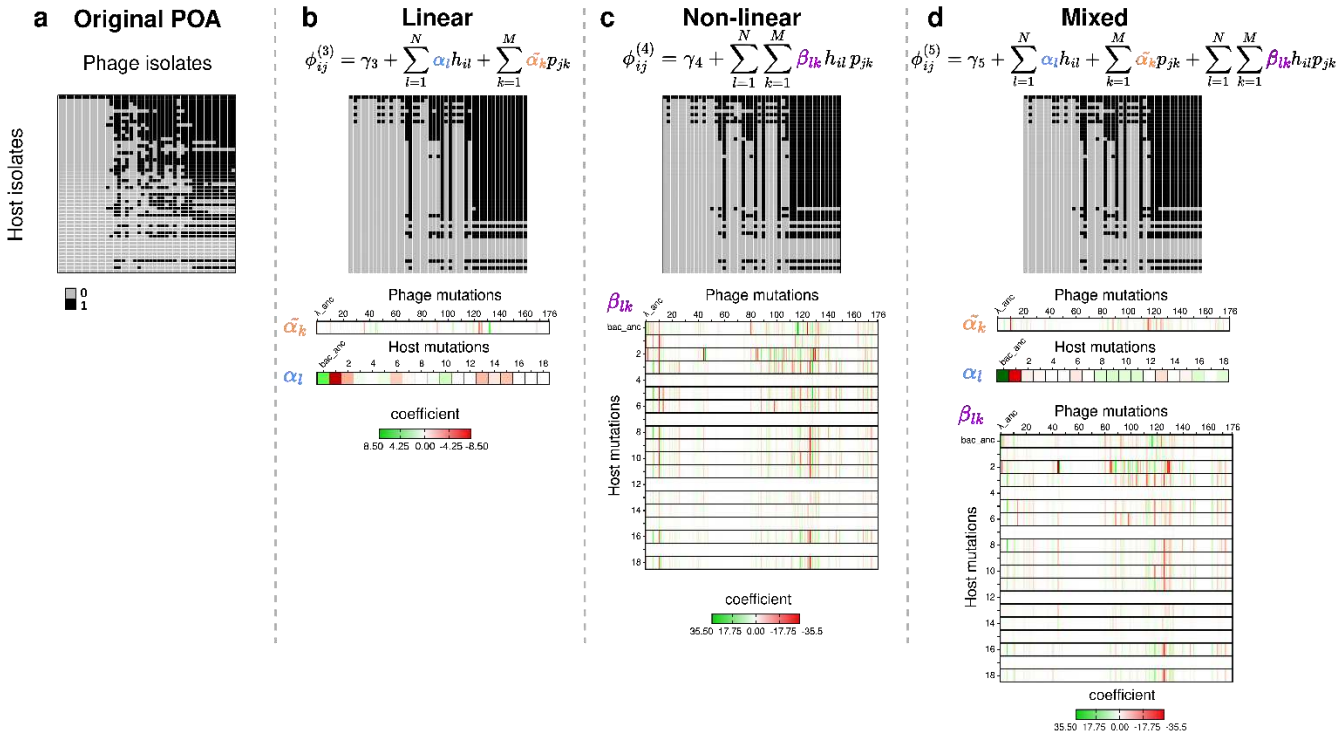


Fig 3. Model for predicting phage-host interaction network. (a) Original POA matrix showing presence (black) and absence (gray) of successful infection between phage (columns) and host (rows) isolated pairs. (b-d) Results of the different model predictions as of the POA matrices, and coefficient values for 176 phage and 18 host mutations plus the ancestor trait using (b) a linear mutation set (equation [6]), (c) nonlinear mutation set (equation [8]) and (d) mixed combination of phage and host mutation set. The color of the coefficient indicates positive (green) to negative (red) effects of each mutation (phage: $\tilde{\alpha}_k$, host: α_l)combination of mutations, β_{lk}.

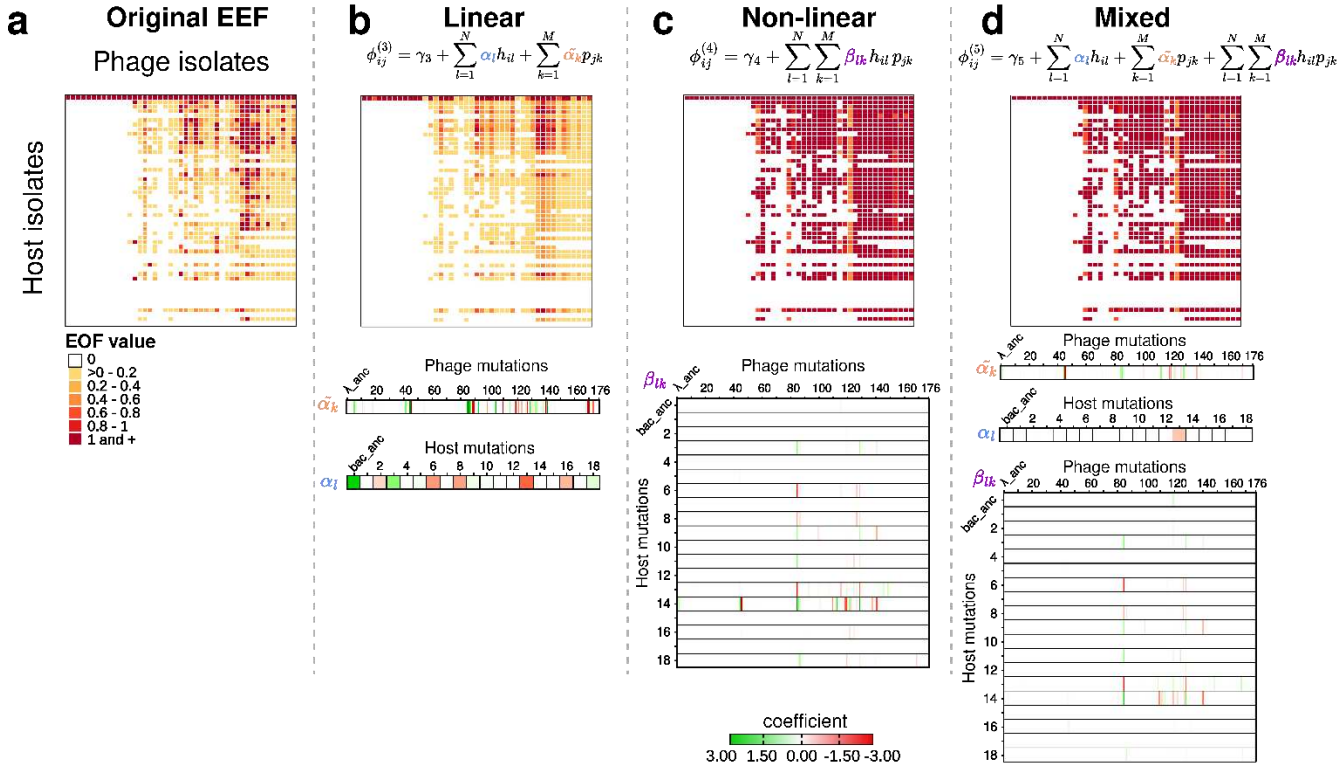
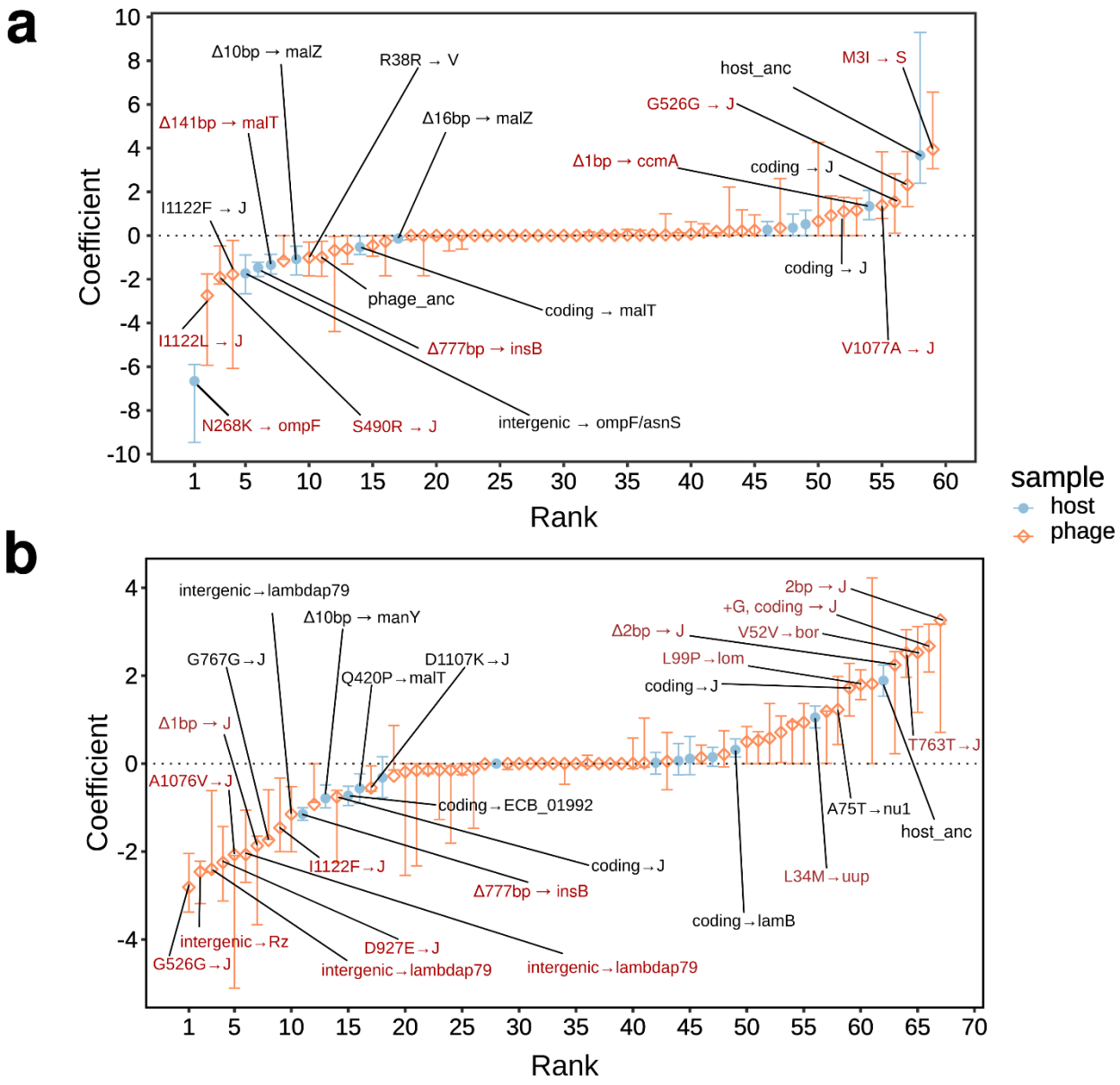


Fig 4. Model for predicting the efficiency of infection. (a) Original cross-infection matrix where colors are EOP values of infection between phage (columns) and host (rows) isolate pairs, white cells indicate no infection. (b-d) Results of the different model predictions as of the EFF matrices, and coefficient values for 176 phage and 18 host mutations plus the ancestor trait using (b) a linear mutation set (equation [6]), (c) nonlinear mutation set (equation [8]) and (d) mixed combination of phage and host mutation set (equation [10]). The color of the coefficient indicates positive (green) to negative (red) effects of each mutation (phage: $\tilde{\alpha}_k$, host: α_l) combination of mutations, β_{lk}.



686

687 **Fig 5. Rank ordered (most negative-most positive) coefficients for important mutations in the**

688 **final model.** The importance of features was measured by the absolute value of the coefficients

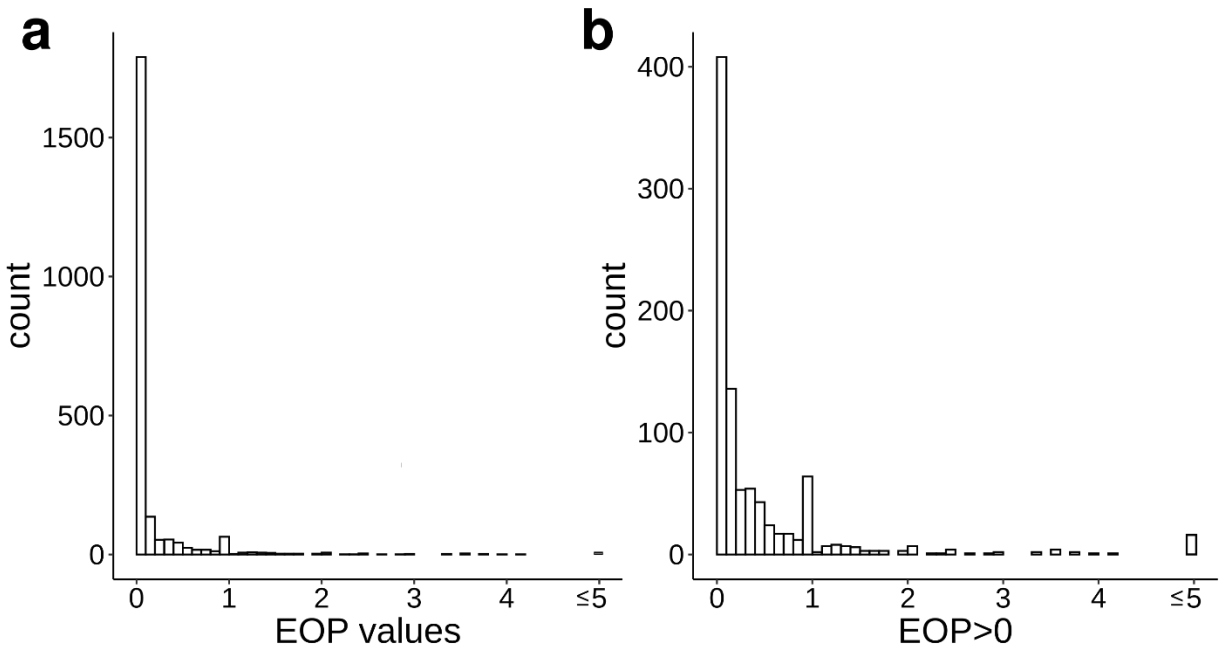
689 learned from each model. Error bars indicate 0.9 quantile. Labels indicate "mutation → gene"

690 when the 90th quantile excludes 0. Mutations in red have the highest positive (negative)

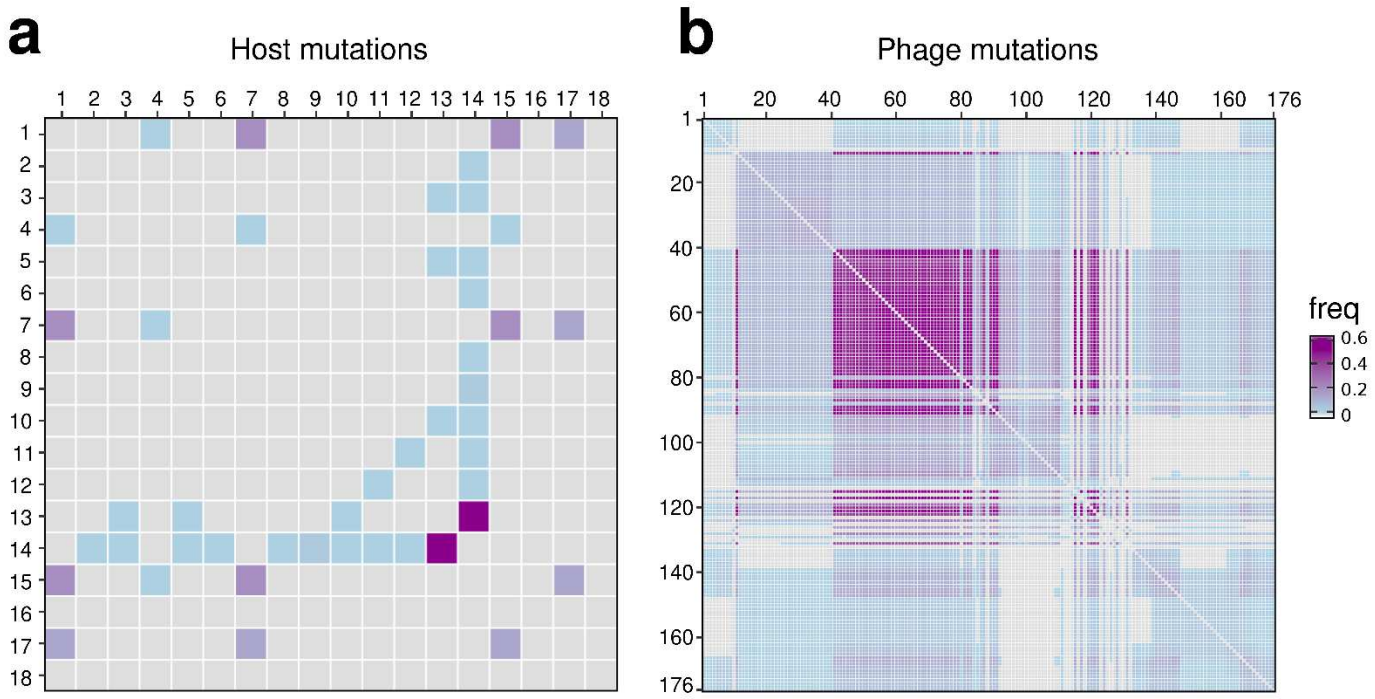
691 coefficients which lowest (highest) value is larger (smaller) or equal to 0 (from 200 bootstrap runs)

692 and are discussed in the main text. Important features for (a) the final model predicting POA
693 include a total of 59 non-zero coefficients, and (b) 67 non-zero coefficient values for the final
694 model predicting EFF. The complete lists of mean, maximum and minimum values of the
695 coefficients associated to mutations predicting POA and EFF are shown in S2 Table and S3 Table
696 respectively.

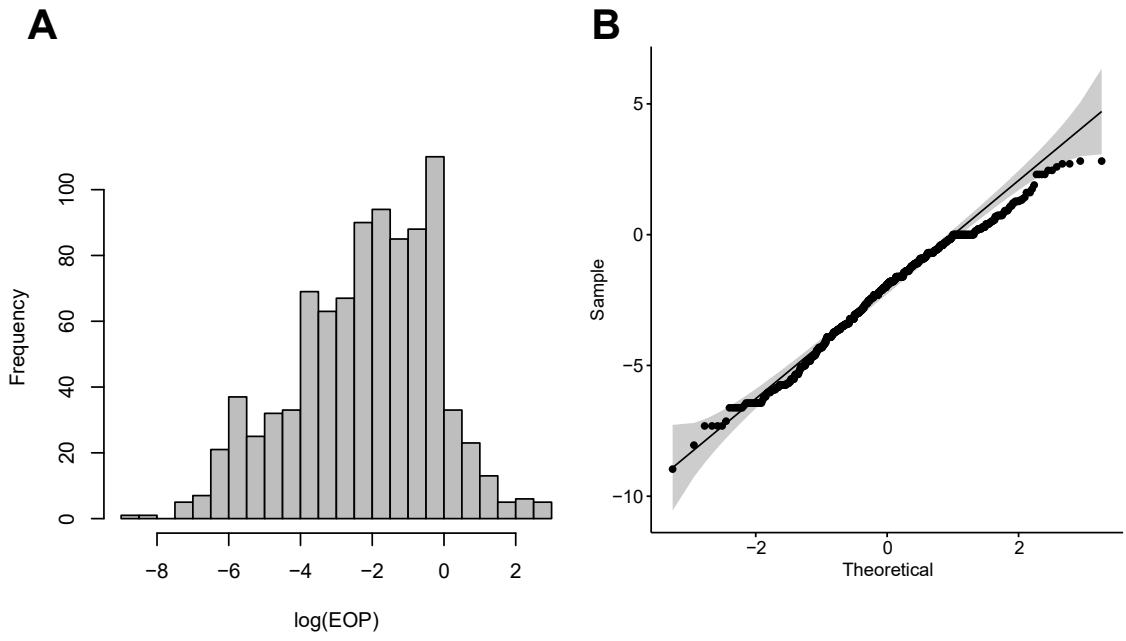
Supporting information



S1 Fig. Distribution of the experimentally obtained EOP values. (A) Original distribution of the EOP values for 2295 phage-host infection pairs. (B) Distribution of 913 positive EOP values. Bin width=0.1.



S2 Fig. Correlations of mutational appearances in host and phage. (a) 18x18 host and (b) 176x176 phage mutation matrices representing the frequency with which pairs of mutations simultaneously appear within the same genetic background.



705 **S3 Fig. Log transformed positive EOP value distribution.** (A) Distribution of the log positive EOP

706 values (B) Q-Q plot for log positive EOP values against normal quantiles (Shapiro-Wilk test *P* value

707 = 3.283e-8)

708 **S1 Table. Mutation profile tables for host and phage.**

709 **S2 Table. Ordered features with non-zero coefficients from final model for predicting POA**

710 **based on a linear combination of phage and host mutation profiles.**

711 **S3 Table. Ordered features with non-zero coefficients from final model for predicting EFF based**

712 **on a linear combination of phage and host mutation profiles.**