


Gain-of-function experiments with bacteriophage lambda uncover residues under diversifying selection in nature

Rohan Maddamsetti,^{1,2}  Daniel T. Johnson,³ Stephanie J. Spielman,⁴ Katherine L. Petrie,^{3,5} Debora S. Marks,⁶ and Justin R. Meyer^{3,7}

¹Department of Biological Sciences, Old Dominion University, Norfolk, Virginia

²E-mail: rmaddams@odu.edu

³Division of Biological Sciences, University of California San Diego, La Jolla, California

⁴Department of Biological Sciences, Rowan University, Glassboro, New Jersey

⁵Earth-Life Science Institute, Tokyo Institute of Technology, Japan

⁶Department of Systems Biology, Harvard Medical School, Boston, Massachusetts

⁷E-mail: jrmeyer@ucsd.edu

Received May 25, 2018

Accepted August 13, 2018

Viral gain-of-function mutations frequently evolve during laboratory experiments. Whether the specific mutations that evolve in the lab also evolve in nature and whether they have the same impact on evolution in the real world is unknown. We studied a model virus, bacteriophage λ , that repeatedly evolves to exploit a new host receptor under typical laboratory conditions. Here, we demonstrate that two residues of λ 's J protein are required for the new function. In natural λ variants, these amino acid sites are highly diverse and evolve at high rates. Insertions and deletions at these locations are associated with phylogenetic patterns indicative of ecological diversification. Our results show that viral evolution in the laboratory mirrors that in nature and that laboratory experiments can be coupled with protein sequence analyses to identify the causes of viral evolution in the real world. Furthermore, our results provide evidence for widespread host-shift evolution in lambdoid viruses.

KEY WORDS: Experimental evolution, gain-of-function, host shift, genomic epidemiology, natural variation, synthetic biology.

Many viruses can expand their host range with a few mutations (Imai et al. 2012; Meyer et al. 2012; Longdon et al. 2014) that enable the exploitation of new receptors (Imai et al. 2012; Meyer et al. 2012). Such mutations may be the first steps toward an epidemic outbreak: this observation has motivated theoretical (Antia et al. 2003), experimental, and surveillance studies of host-range shifts in emergent pathogens, including avian influenza (Koel et al. 2013; Linster et al. 2014; Shi et al. 2014; Song et al. 2017), coronaviruses (Lu et al. 2013; de Wit et al. 2016), HIV (Rambaut et al. 2004), and ebolavirus (Holmes et al. 2016).

Ideally, evolution experiments could accelerate our understanding of host-range shifts; however, it is not clear whether host-range shifts observed in the laboratory can faithfully inform us about host-range shifts in nature, for at least two reasons. First,

evolutionary trajectories might be sensitive to differences in environmental conditions between the laboratory and nature. Second, the number of evolutionary paths sampled in laboratory experiments is very small compared to natural virus diversity due to the enormous size of viral populations. Indeed, some researchers have called for the suspension of gain-of-function experiments on the grounds that they would tell us little about real-world evolution at the risk of constructing pandemic strains (Casadevall and Imperiale 2014).

Here, we use a harmless virus, bacteriophage λ , to demonstrate how gain-of-function experiments can identify mutations and evolutionary processes that mirror those that occur in nature. Typical laboratory strains of λ infect *Escherichia coli* by binding to the outer membrane protein LamB (Chatterjee and Rothenberg

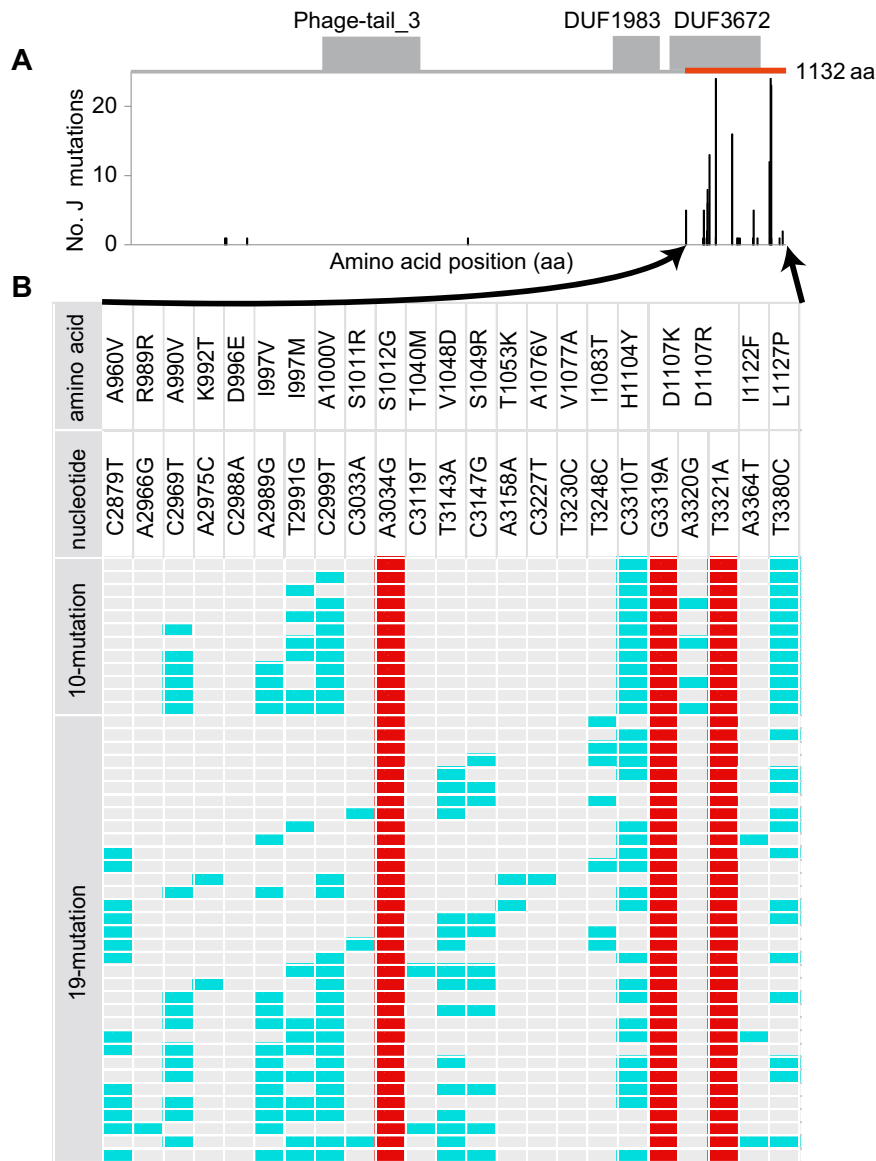


Figure 1. (A) Distribution of J mutations evolved en route to *OmpF*⁺ and (B) synthetic phage genotypes capable of using *OmpF*. (A) Mutations summed across 24 genomes independently evolved to exploit *OmpF* in Meyer et al. (2012). Protein domains as annotated in the Pfam database are shown in gray. The majority of mutations either occur in the DUF3672 domain, or in the C-terminus past the annotated boundary of DUF3672, in a region that we call the “specificity region.” The specificity region spans residues 960–1132, and is marked in red. (B) Synthetic *OmpF*⁺ genotypes indicated by colored cells (red marks the critical three) observed after combinatorial engineering of 10 common mutations or 19 mutations when the critical three were fixed. Amino acid and nucleotide changes are indicated by positions bookended by the wild-type state and then the evolved state. Amino acids in position 997 and 1107 have multiple derived states.

2012), but the phage rapidly evolves in the laboratory to exploit a different membrane protein, *OmpF* (Meyer et al. 2012; Meyer et al. 2016; Petrie et al. 2018). These experiments are a proxy for the ability of the phage to switch hosts. The evolved gain-of-function phenotype in λ , *OmpF*⁺, involves multiple nonsynonymous mutations in the host-recognition gene *J*. Each *OmpF*⁺ isolate in Meyer et al. (2012) had between 4 and 10 single nucleotide substitutions in *J*, and none had insertions or deletions (indels). These nucleotide substitutions evolve in parallel across

replicate experiments, indicating the action of strong positive selection rather than hypermutability. The 97% of the substitutions in 24 independently evolved *OmpF*⁺ λ phage occurred in the *J* protein between residues 960–1132, which we call the “specificity region” of *J* (Fig. 1A). Not much is known about the *J* protein structure and how exactly these mutations modify *J*. Homology based analyses of the protein reveal that they fall outside two structural domains of the *J* protein (a Phage-tail.3 domain spanning residues 330–500 and a DUF1983 domain spanning residues

845–925) and overlap with a third: a DUF3672 domain spanning residues 949–1090.

By comparing *J* among OmpF⁺ and OmpF⁻ λ , Meyer et al. (2012) suggested that the OmpF⁺ innovation required four mutations: one at residue 1012, two in the codon for residue 1107, and a fourth mutation somewhere between residues 990 to 1000, but these criteria were not directly tested. Here, our first step was to test this prediction by determining which combinations of *J* mutations facilitate the gain-of-function. Our second step was to determine whether the identified sites also evolve similarly in nature. For our final step, we tested whether the identified sites affect the evolutionary trajectory of the virus by examining whether they are associated with diversification. This last step was motivated by an observation in our alignment of *J* homologs: indels at residues 1012 and 1107 change the amino acid sequence of *J* and seem to define distinct *J* lineages. Since a signature of adaptive diversification is the birth of new clades (Herron and Doebeli 2013), we reasoned that if indels at residues 1012 and 1107 cause adaptive diversification through host-shifts, then they should trigger the birth of the clades observed in the *J* alignment. If true, then the indels should occur on long, diverging branches on the *J* phylogeny (Rozen et al. 2005). Indeed, this prediction was borne out in a phylogenetic analysis of the *J* homologs.

Material and Methods

COMBINATORIAL GENETICS EXPERIMENTS TO IDENTIFY WHICH MUTATIONS CAUSE OmpF⁺

Genetic edits were made by Multiplexed Automated Genome Engineering (MAGE) (Wang et al. 2009; Wang and Church 2011; Meyer et al. 2016) in λ strain cI857 (provided by Ing-Nang Wang, SUNY Albany) integrated into the genome of HWEC106 (provided by Harris Wang, Columbia University). MAGE uses the λ -red recombineering system, which uses oligonucleotides (oligos) to introduce genetic edits. The λ -red recombineering system was provided on the pKD46 plasmid (Datsenko and Wanner 2000). For a description of the oligos used, see Table S1.

We constructed two separate libraries. For the first, 18 rounds of MAGE were used to create a combinatorial library of 10 commonly evolved *J* mutations. We screened for OmpF⁺ isolates by plating on a lawn of *lamB*-deleted *E. coli* (JW3996 from the KEIO collection (Baba et al. 2006)), and then sequencing 33 randomly chosen plaques. We only sequenced the C-terminus of the *J* gene with the Sanger method. Unpurified PCR products (Forward primer: 5' CGCATCGTTACCTCTCACT; Reverse primer: 5' CCTGCGGGCGGTTTGTCATT) were submitted to the Genewiz La Jolla, CA facility for sequencing. One isolate had all 10 mutations, indicating that 18 rounds was sufficient to generate even the most unlikely genotype. Twelve unique genotypes were uncovered among the 33 isolates sequenced.

For the second library, we created a genomic backbone for further editing by using the oligos “a3034g” and “g3319a t3321a” to introduce the three most commonly observed mutations in OmpF⁺ λ genotypes. Next, we performed a number of different MAGE trials to maximize the number of unique alleles we observed. See Table S2 for our MAGE strategy. We screened for λ genotypes able to infect through OmpF by isolating strains that produced plaques on lawns of *E. coli* with *lamB* deleted. We sequenced 88 isolates from the lawns and uncovered 34 unique genotypes.

ANALYSIS OF NATURAL *J* VARIATION

We used the *evcouplings* pipeline (Hopf et al. 2017) to generate jackhammer (Eddy 2011) alignments of 1207 full-length *J* protein homologs (parameter settings: bitscore = 0.2; theta = 0.999; seqid_filter = 95 and 99 (thresholds for filtering highly similar sequences); minimum_sequence_coverage = 99 (require sequences to align to 99% of wild-type *J* protein); and minimum_sequence_coverage = 50). By default, this alignment only includes residues found in the wild-type *J* protein and excludes insertions relative to wild-type *J*. We included these insertions by querying full-length sequences in a larger Stockholm-formatted alignment produced by the *evcouplings* pipeline. When analyzing the specificity region (say, constructing a phylogeny), we used residues 960–1132 of this full-length alignment.

We used FastTree (Price et al. 2010) with an LG substitution matrix (Le and Gascuel 2008) to generate approximately maximum-likelihood phylogenies. To account for recombination when estimating site-specific evolutionary rates in the specificity region, we first ran a modified SBP algorithm on the specificity region (Kosakovsky Pond et al. 2006). We found a putative recombination breakpoint at alignment position 49, supported by Akaike's information criterion but not the Bayesian information criterion. We therefore partitioned the specificity region at position 49 and recalculated trees for each partition using FastTree with an LG substitution matrix. We next calculated site-specific evolutionary rates with LEISR (Spielman and Kosakovsky Pond 2018), a scalable implementation of Rate4Site (Pupko et al. 2002) that accounts for recombination breakpoints. We ran LEISR using an LG substitution matrix.

We carried out a complete phylogenetic analysis of recombination in the λ genome in order to compare to *J*. We made alignments of all 66 proteins in the reference proteome for phage λ in UniProt by using phmmer on the HMMER webserver (Finn et al. 2015) accessible through the EMBL-EBI portal (Chojnacki et al. 2017) to search for homologs with conserved protein domain architectures in nine closely related lambdoid viruses (Rohwer and Edwards 2002) in the UniProt Reference Proteomes database: Escherichia virus Lambda (taxid: 10710), Enterobacteria phage BP-4795 (taxid: 196242), Enterobacteria phage HK630

(taxid: 1147146), Enterobacteria phage HK629 (taxid: 1147148), Enterobacteria phage DE3 (taxid: 482822), Enterobacteria phage mEp460 (1147152), Enterobacteria phage *cdtI* (taxid: 414970), Escherichia phage Ayreon (taxid: 2040288), and Stx2-converting phage Stx2a_WGPS2 (taxid: 1226260). We used MAFFT (Katoch et al. 2017) to align the homologs, and FastTree to make phylogenies. We calculated normalized geodesic and Robinson-Foulds distances between the phylogenies for J, its specificity region, and the rest of the proteins in the λ genome using TreeCl (Gori et al. 2016). We visualized the distances between trees using metric multidimensional scaling in TreeCl. We visualized individual phylogenies using the ETE Toolkit (Huerta-Cepas et al. 2016).

We calculated gap entropy at each position as: $-[p \cdot \log_2(p) + (1 - p) \cdot \log_2(1 - p)]$, where p is the frequency of gap characters at that position, and $1 - p$ is the frequency of all amino acids at that position. To maximize independence between gap entropy and amino acid entropy, we excluded gap characters when calculating amino acid entropy.

For all nonparametric bootstrap calculations, we used 100,000 bootstraps, and chose groups of sites without replacement; that is, if calculating a statistic for a group of 17 residues, we would choose 17 different residues for one sample, and re-sample 100,000 times.

We tested whether indels at residues 1012, 1048, 1077, and 1107 occur on longer branches of the specificity region phylogeny as follows. We selected all branches on which indels were either gained or lost. We compared the lengths of these branches to the length of all other branches in the tree using a Kruskal–Wallis rank-sum test, excluding branches of length zero.

One possible problem with this analysis would occur if indels themselves disproportionately contributed to branch length, thereby causing an artificial correlation between branch length and these mutations. This possibility can be ruled out: when FastTree estimates branch lengths, positions with gaps are either ignored (when comparing two sequences) or weighted by their proportions of nongaps (when comparing two sequence profiles).

Results

IDENTIFICATION OF THE *OmpF*⁺ GAIN-OF-FUNCTION MUTATIONS

We determined which mutations were required for *OmpF* use by constructing a combinatorial library of 10 commonly evolved *J* mutations identified in Meyer et al. (2012), and then screening the library for *OmpF*⁺ transformants. All engineered *OmpF*⁺ strains possessed the mutation at residue 1012 as well as G3319A and T3321A mutations at residue 1107 (both are required to change an aspartic acid into a lysine). In contrast to previous observations,

some strains lacked a fourth mutation between residues 990 and 1000 (Fig. 1). Hence, we call the mutations at residues 1012 and 1107 the “critical mutations.”

To test whether the three critical mutations were sufficient to confer the *OmpF*⁺ phenotype, we constructed a synthetic phage with just these mutations. Even though the synthetic phage was viable, it proved unable to exploit cells without the ancestral LamB receptor, demonstrating that at least four *J* mutations were necessary for *OmpF*⁺. To find what further mutations might be needed to confer the *OmpF*⁺ phenotype, we constructed a second combinatorial library using the phage with the three critical *J* mutations as the baseline, and random combinations of 19 other *J* mutations found in the *OmpF*⁺ λ evolved by Meyer et al. (2012). These 19 included six of the original 10 mutations. Four mutations were excluded because they occur within the two codons we already modified. Eighty-eight *OmpF*⁺ isolates were sequenced from this much larger library. One *OmpF*⁺ isolate had just a single extra mutation at residue 1083 in addition to the three critical mutations (Fig. 1). However, the majority of the engineered *OmpF*⁺ phage did not possess this specific mutation, signifying that its function could be substituted by other *J* mutations. In total, this experiment revealed that four mutations are sufficient to evolve the innovation, but only two specific amino acid changes (at residues 1012 and 1107) are universally required to access *OmpF* in the context of these laboratory experiments.

ELEVATED VARIATION AND EVOLUTIONARY RATES AT KEY RESIDUES OF NATURAL *J* HOMOLOGS

To test whether our experiments reflected natural evolution, we collected and aligned full-length homologous *J* protein sequences from UniRef100 (Suzek et al. 2015) (1207 highly similar sequences). Most sequences were prophage uncovered in the genomes of their *Enterobacteriaceae* hosts, including bacterial genera *Escherichia*, *Salmonella*, *Citrobacter*, *Edwardsiella*, and even *Cronobacter*. The diversity of hosts suggests that the prophages are adapted to a wide range of different receptors and host species.

Recall that 97% of substitutions in *OmpF*⁺ gain-of-function experiments occur in the specificity region. Likewise, the natural *J* homologs had disproportionate variation here: 29% of the total amino acid variation occurred in the specificity region, despite it only being 16% of the total length of *J* (Fig. 2A). As we will discuss later, peaks in amino acid variation correspond to peaks in indel variation (Fig. 2B). This nonrandom clustering of variation in the specificity region strongly suggests that *J* has experienced substantial diversifying selection on host attachment. Furthermore, the 17 residues with substitutions in the screened *OmpF*⁺ isolates (excluding a synonymous R989R substitution) were significantly more variable than randomly chosen groups of 17 residues from *J* (nonparametric bootstrap: $P < 10^{-5}$) and

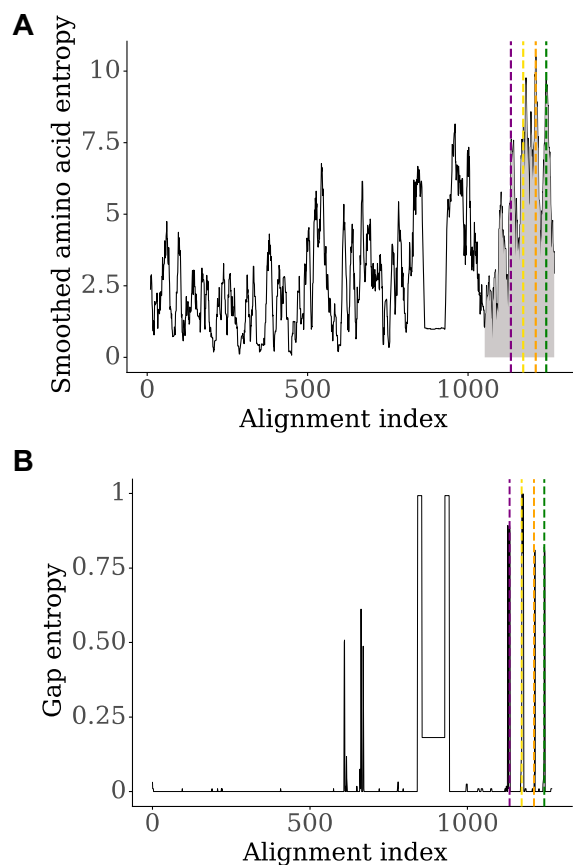


Figure 2. (A) Amino acid entropy averaged over a 10-residue window in the full J alignment, including insertions relative to wild-type J. Dashed purple and green lines indicate residues 1012 and 1107, and dashed yellow and orange lines are drawn at residues 1049 and 1077, respectively. The specificity region (residues 960–1132) is shaded in gray. (B) Gap entropy over the full J alignment, including insertions relative to wild-type J. Dashed lines drawn as in (A).

from the specificity region 960–1132 (nonparametric bootstrap: $P = 0.00075$) showing the experiments had identified evolutionary hotspots. However, the specific 19 substitutions that evolved at these 17 residues in the laboratory were not common in natural homologs, suggesting that our experiments had the resolution to predict where changes would evolve, but not the exact change (Fig. S1). Focusing in on the two residues critical for the OmpF⁺ gain of function, 1012 and 1107, we find that they are more variable than random pairs of sites in J (nonparametric bootstrap: $P = 0.00055$) and the specificity region (nonparametric bootstrap: $P = 0.020$). Finally, we calculated evolutionary rates for each site in the specificity region, controlling for recombination (Fig. S3). We found that the 17 residues studied in the gain-of-function experiments evolve faster than equally sized random samples taken from the specificity region (nonparametric bootstrap, $P = 0.012$) and residues 1012 and 1107 evolve faster than random pairs of

sites sampled from the specificity region (nonparametric bootstrap, $P = 0.0086$).

The most variable and rapidly evolving regions in the J homologs, including the critical residues 1012 and 1107, are hotspots for in-frame indels. We measured protein indel variation using gap entropy and found that most amino acid sites have none (mode and median gap entropy = 0), with a few notable exceptions. In the specificity region, we see four peaks in gap entropy (Fig. 2B) that correspond to peaks in sequence entropy (Fig. 2A). The purple and green peaks overlap residues 1012 and 1107, which are critical for the OmpF⁺ gain of function. The yellow and orange peaks occur at residues 1048 and 1077 that were not essential for OmpF use; however, λ evolved many mutations during the experiment near these positions (residues 1048, 1049, 1053, 1076, 1077, and 1083). Indels can have large beneficial effects on proteins, including altering specificity by changing surface loops (Chatterjee and Rothenberg 2012; Porcek and Parent 2015), or causing structural rearrangements that improve function (Arpino et al. 2014). We hypothesize that the four peaks represent distinct surface loops that affect host specificity, much as influenza's hemagglutinin contains variable-length surface loops that affect binding to avian and human host receptors (Peacock et al. 2017; Tzarum et al. 2017). Four additional peaks in gap entropy were observed in a region of J that did not evolve in the laboratory. Two particular peaks are notable because they are separated by a plateau caused by one indel nested within another. We are unsure of what consequence this variation has for λ function and evolution.

PHYLOGENETIC ANALYSIS OF J AND THE λ GENOME

To study J evolution, we constructed two separate phylogenies, one for the entire J protein (Fig. 3A) and one for the specificity region (Fig. 3B). In line with the SBP recombination breakpoint analysis previously discussed (Methods), these analyses lead to vastly different phylogenies (normalized Robinson-Foulds distance = 0.93 out of 1.00, normalized geodesic distance = 0.73 out of 1.00). This result suggests that there is frequent recombination and gene transfer between these two gene regions, causing each to inherit different evolutionary histories.

To verify that the J and specificity region phylogenies are significantly different from one another, we compared their phylogenies to phylogenies constructed from all λ proteins. Our hypothesis is supported if the J specificity region phylogeny is more similar to other λ proteins than the phylogeny constructed from the adjacent J sequence. We computed phylogenies for all proteins in the λ genome, and compared them using Robinson-Foulds distance (which measures the topological distance between phylogenies) and geodesic distance (which takes both topology and branch lengths into account). When comparing phylogenies by Robinson-Foulds distance, many distances could not be computed because some proteins were missing in many of the genomes.

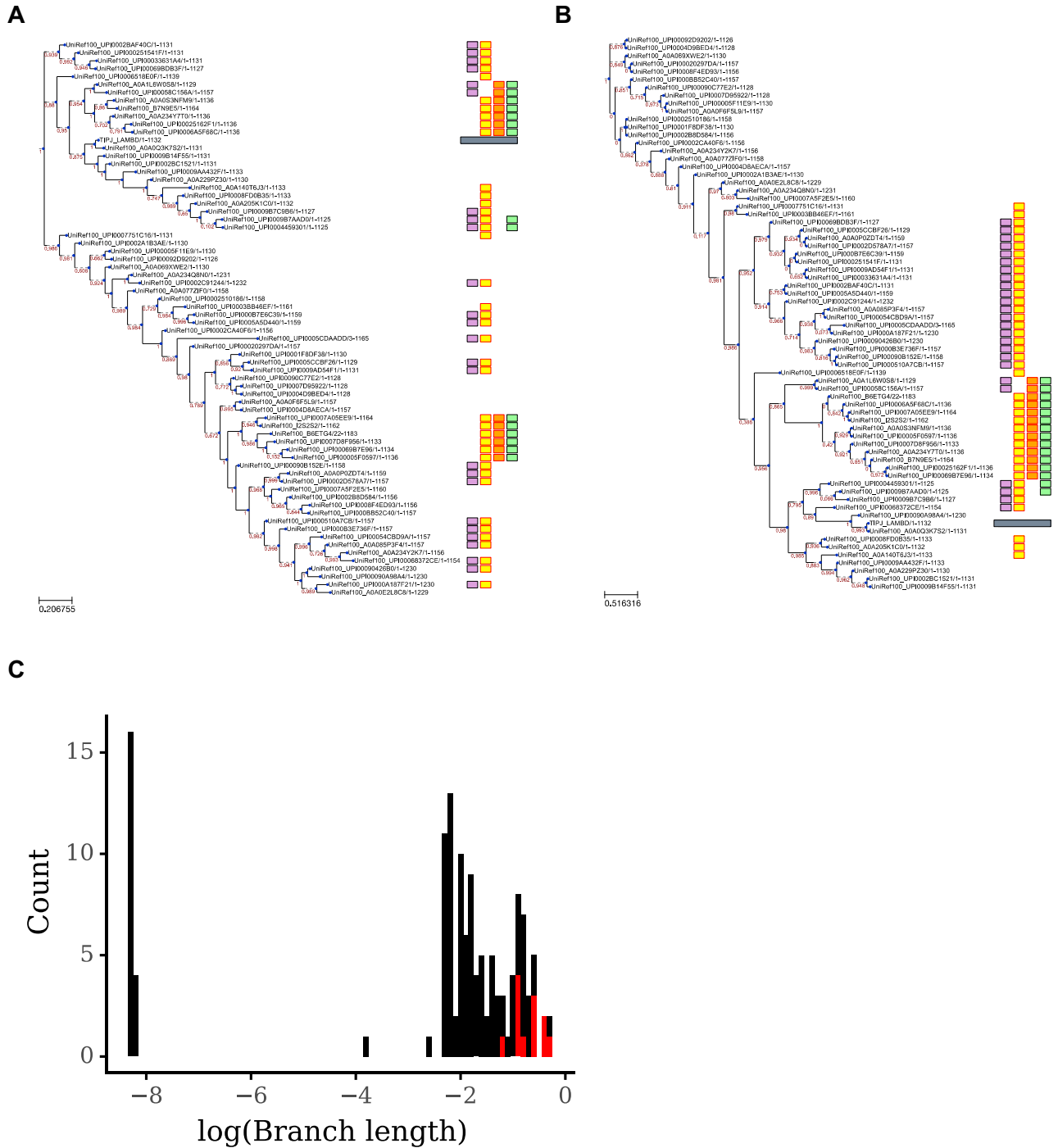


Figure 3. (A) Phylogeny for J protein. Only a subset of the operational taxonomic units (OTUs) is displayed: just OTUs with sequences $\geq 5\%$ divergence from each other. OTUs are labeled with their sequence identifier in the RefSeq100 database. The wild-type J protein sequence is labeled with a gray block. Sequences with indels at residue 1012 are labeled with a purple block. Sequences with indels between residues 1048–1049 are labeled with a yellow block. Sequences with indels between residues 1077–1078 are labeled with an orange block. Sequences with indels at residue 1107 are labeled in light green. Local branch support values, calculated with the Shimodaira–Hasegawa test, are in red. (B) Phylogeny for residues 960–1132 of J protein. OTUs have at least 5% divergence from each other, and are labeled as in (A). (C) Distribution of \log_{10} nonzero branch lengths in the phylogeny shown in (B). Branches containing indels at residues 1012 or 1107 or between residues 1048–1049 or 1077–1078 are in red; all others are in black.

In fact, the failure of this calculation demonstrates a high rate of gene loss, duplication, and horizontal transfer in lambdaoid viruses. When comparing phylogenies by geodesic distance, four distinct clusters of phylogenies are apparent, showing which proteins tend to be coinherited. Importantly, the J specificity region does not cluster with the phylogeny estimated from the remaining J sequence that encompasses structural domains (Fig. S5). From these analyses, as well as our visual observation of recombination in the J protein alignment (Data S1), we conclude that the specificity region is an evolutionary module that forms an independent functional unit distinct from the rest of J.

This discovery suggests that the specificity region circulates through the phage population and evolves as an independent segment. This modularity helps explain why this region's evolution was found to be distinct in both experiments and natural sequence entropy. The modularity may allow the specificity region to freely evolve and diversify without being constrained by requiring coordinated changes in conserved structural regions of the protein.

Within the specificity region the opposite pattern occurs: multiple sites show correlated evolution. The most striking example is the four indels discussed previously. The frame insertions at 1048–1049 and 1077–1078 (indels not associated with the critical residues) coincide with the deletions at the critical residues (1012 and 1107) (Fig. 3C and Fig. S2). Their correlated evolution suggests that there is functional dependence (epistasis) between the indels and that sites 1048–1049 and 1077–1078 play a role in host-range evolution.

INDEL VARIATION CORRELATES WITH DIVERSIFICATION OF SPECIFICITY REGION SUBTREES

Are the indels responsible for receptor-use evolution in nature? If true, then the four indel-rich regions would have a nonrandom distribution within the phylogeny. We reasoned that if the indels affect receptor tropism, then they would cause ecological differentiation and facilitate the long-term maintenance of distinct evolutionary lineages (Rozen et al. 2005; Herron and Doebeli 2013). To test this, we compared the length of branches on which indels occur to the length of all other branches. Indeed, the branches on which indels in those four regions occur are significantly longer than other branches of the specificity-region phylogeny (Kruskal–Wallis tests excluding zero-length branches: $P = 8.7 \cdot 10^{-7}$ in Fig. 3D; $P = 5.6 \cdot 10^{-10}$ in Fig. S4). This pattern suggests that the indels play a key role in shaping λ evolution by contributing to cladogenesis (Fig. S2).

Discussion

As repositories of gene and protein sequences grow and biologists begin to gain access to the overwhelming genetic diversity of

life, researchers are faced with the challenge of uncovering the relatively few mutations that are functionally important and evolutionarily significant. By combining laboratory evolution experiments with genome editing, we were able to identify protein residues with important functional consequences that have remarkable evolutionary properties, including disproportionate levels of heterogeneity and elevated rates of evolution. Given the concordance between evolutionary rates at critical residues in laboratory and natural populations, we posit that these properties of the natural variation are due to positive selection for receptor usage and host-range expansion.

A major question in evolutionary biology is what traits cause diversification. It is thought that key innovations allow species to unlock new ecological opportunities, which in turn drives enhanced diversification rates (Mayr 1970). This should be recorded in phylogenies as a sudden increase in tree bushiness within monophyletic groups. Analyses that test correlations between subtree diversification rates and trait states have recently been found to be prone to false positives (Rabosky and Goldberg 2015). Given this finding, we took an alternative approach to test whether variation identified in our study leads to cladogenesis. Rather than focusing on bushiness, we evaluated whether indels tended to occur on longer branches. We reasoned that if the indels cause host shifts, then the lineages they occur in will uncover new ecological opportunities. This will have two important consequences, (1) the host-shift may cause elevated rates of molecular evolution as the virus adapts to its new niche, and (2) lineages with different hosts will no longer compete with each other and will be more likely to coexist for deep evolutionary time. Each process would be recorded in the phylogeny as a correlation between innovative traits and the length of the branch they occur on. Indeed, there is overwhelming evidence that lineages that evolved indels at residues associated with an experimental host-shift take longer to coalesce than other lineages. This demonstrates that changes associated with receptor usage are associated with ecological diversification.

One possible alternative explanation for the patterns we observed here, specifically the heightened rates of evolution and elevated levels of diversity, may be that the critical residues in specificity region are prone to mutation. This explanation is unlikely given the evidence from gain-of-function experiments. If the critical residues were prone to mutation, then a diversity of substitutions at those residues should have been observed in the laboratory. Instead, those residues evolved in specific and parallel ways across replicate experiments, indicating the action of strong positive selection. Furthermore, if this variation was due to random mutations and not selection for functional changes, then we would not expect to observe an association between indel evolution and branch length.

The congruence we find between laboratory and natural evolution in λ contrasts with work showing that adaptation

in Lenski's long-term experiment anti-correlates with natural protein variation in *E. coli* (Maddamsetti et al. 2017). Presumably, the simplicity of Lenski's experiment leads the bacteria down evolutionary paths not taken in nature. By contrast, the dominant selection pressure in evolution experiments with bacteriophage—attachment to bacterial host cells—is probably also a dominant selection pressure on wild phage. Notably, another study on bacteriophage has observed residues under positive selection in both the lab and nature (Wichman et al. 2000), as have studies on canine parvovirus (Allison et al. 2014) and poliovirus (Stern et al. 2017). Apparently, the reverse strategy—identifying candidate residues for changes in phenotypes in natural populations and then validation in the laboratory—does not work as well, at least in one notable case (Liu et al. 2017). One reason put forth in Liu et al. (2017) for the failed nature-lab connection are overlooked effects of higher order epistasis (Weinreich et al. 2013). Indeed, our methodology revealed that a four-way interaction among the *J* mutations was required to endow new function on OmpF, and so we would not have been able to uncover them by typical methods that only measure the effects of single or pairs of mutations.

Based on the patterns of *J* variation we observe, we suggest that host-range evolution is common in this group of viruses, and perhaps others too. While the frequency of host-range evolution may be unsettling, our work also demonstrates potential methods to predict host shifts in natural populations. Such an approach is in the same spirit as pioneering work by Barlow and Hall, who had success in predicting the evolution of antibiotic resistance by combining evolution experiments with analyses of natural and clinical isolates (Hall 2002; Barlow and Hall 2003; Hall 2004; Salverda et al. 2010). In particular, worrisome mutations can be identified with experiments as described here or with other laboratory techniques such as deep mutational scanning (Bloom 2017). This information can be combined with genomic surveillance efforts (Gire et al. 2014; Grubaugh et al. 2017) to devise better strategies to eradicate potential pandemics.

AUTHOR CONTRIBUTIONS

J.R.M. and R.M. designed and conceived the study. D.T.J. conducted MAGE experiments. K.L.P. analyzed data. R.M. and S.S. performed statistical and phylogenetic analyses. R.M., D.S.M., and J.R.M. wrote the article and everyone edited it.

ACKNOWLEDGMENTS

We thank Anna Green, John Ingraham, Adam Riesselman, Kelly Brock, David Ding, and Alita Burmeister for helpful discussions. K.L.P. was supported by the ELSI Origins Network (EON), which is funded by the John Templeton Foundation. The ideas expressed in this publication are those of the authors and not necessarily those of the funding sources.

DATA ARCHIVING

Data for this project has been deposited in the Dryad Digital Repository at <https://doi.org/10.5061/dryad.cm86089>

CONFLICT OF INTERESTS

The authors declare no competing financial interests.

LITERATURE CITED

- Allison, A. B., D. J. Kohler, A. Ortega, E. A. Hoover, D. M. Grove, E. C. Holmes, and C. R. Parrish. 2014. Host-specific parvovirus evolution in nature is recapitulated by in vitro adaptation to different carnivore species. *PLoS Pathogens* 10:e1004475.
- Antia, R., R. R. Regoes, J. C. Koella, and C. T. Bergstrom. 2003. The role of evolution in the emergence of infectious diseases. *Nature* 426:658–661.
- Arpino, J. A., S. C. Reddington, L. M. Halliwell, P. J. Rizkallah, and D. D. Jones. 2014. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* 22:889–898.
- Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* 2:2006.0008.
- Barlow, M., and B. G. Hall. 2003. Experimental prediction of the natural evolution of antibiotic resistance. *Genetics* 163:1237–1241.
- Bloom, J. D. 2017. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol. Direct* 12:1.
- Casadevall, A., and M. J. Imperiale. 2014. Risks and benefits of gain-of-function experiments with pathogens of pandemic potential, such as influenza virus: a call for a science-based discussion. *MBio* 5:e01730–01714.
- Chatterjee, S., and E. Rothenberg. 2012. Interaction of bacteriophage λ with its *E. coli* receptor, LamB. *Viruses* 4:3162–3178.
- Chojnacki, S., A. Cowley, J. Lee, A. Foix, and R. Lopez. 2017. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res.* 45:W550–W553.
- Datsenko, K. A., and B. L. Wanner. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. USA* 97:6640–6645.
- de Wit, E., N. van Doremalen, D. Falzarano, and V. J. Munster. 2016. SARS and MERS: recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* 14:523–534.
- Eddy, S. R. 2011. Accelerated profile HMM searches. *PLoS Comp. Biol.* 7:e1002195.
- Finn, R. D., J. Clements, W. Arndt, B. L. Miller, T. J. Wheeler, F. Schreiber, A. Bateman, and S. R. Eddy. 2015. HMMER web server: 2015 update. *Nucleic Acids Res.* 43:W30–W38.
- Gire, S. K., A. Goba, K. G. Andersen, R. S. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345:1369–1372.
- Gori, K., T. Suchan, N. Alvarez, N. Goldman, and C. Dessimoz. 2016. Clustering genes of common evolutionary history. *Mol. Biol. Evol.* 33:1590–1605.
- Grubaugh, N. D., J. T. Ladner, M. U. G. Kraemer, G. Dudas, A. L. Tan, K. Gangavarapu, M. R. Wiley, S. White, J. Theze, D. M. Magnani, et al. 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 546:401–405.
- Hall, B. G. 2002. Predicting evolution by in vitro evolution requires determining evolutionary pathways. *Antimicrobial Agents Chemotherapy* 46:3035–3038.
- . 2004. Predicting the evolution of antibiotic resistance genes. *Nat. Rev. Microbiol.* 2:430.

- Herron, M. D., and M. Doebeli. 2013. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol.* 11:e1001490.
- Holmes, E. C., G. Dudas, A. Rambaut, and K. G. Andersen. 2016. The evolution of Ebola virus: insights from the 2013–2016 epidemic. *Nature* 538:193–200.
- Hopf, T. A., J. B. Ingraham, F. J. Poelwijk, C. P. Scharfe, M. Springer, C. Sander, and D. S. Marks. 2017. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35:128–135.
- Huerta-Cepas, J., F. Serra, and P. Bork. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33:1635–1638.
- Imai, M., T. Watanabe, M. Hatta, S. C. Das, M. Ozawa, K. Shinya, G. Zhong, A. Hanson, H. Katsura, S. Watanabe et al. 2012. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* 486:420–428.
- Katoh, K., J. Rozewicki, and K. D. Yamada. 2017. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* <https://doi.org/10.1093/bib/bbx108>.
- Koel, B. F., D. F. Burke, T. M. Bestebroer, S. van der Vliet, G. C. Zondag, G. Vervaet, E. Skepner, N. S. Lewis, M. I. Spronken, C. A. Russell et al. 2013. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* 342:976–979.
- Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* 23:1891–1901.
- Le, S. Q., and O. Gascuel. 2008. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25:1307–1320.
- Linster, M., S. van Boheemen, M. de Graaf, E. J. A. Schrauwen, P. Lexmond, B. Manz, T. M. Bestebroer, J. Baumann, D. van Riel, G. F. Rimmelzwaan et al. 2014. Identification, characterization, and natural selection of mutations driving airborne transmission of A/H5N1 virus. *Cell* 157:329–339.
- Liu, J., I. M. Cattadori, D. G. Sim, J.-S. Eden, E. C. Holmes, A. F. Read, and P. J. Kerr. 2017. Reverse engineering field isolates of myxoma virus demonstrates that some gene disruptions or losses of function do not explain virulence changes observed in the field. *J. Virol.* 91:e01289–01217.
- Longdon, B., M. A. Brockhurst, C. A. Russell, J. J. Welch, and F. M. Jiggins. 2014. The evolution and genetics of virus host shifts. *PLoS Pathog.* 10:e1004395.
- Lu, G., Y. Hu, Q. Wang, J. Qi, F. Gao, Y. Li, Y. Zhang, W. Zhang, Y. Yuan, J. Bao et al. 2013. Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. *Nature* 500:227–231.
- Maddamsetti, R., P. J. Hatcher, A. G. Green, B. L. Williams, D. S. Marks, and R. E. Lenski. 2017. Core genes evolve rapidly in the long-term evolution experiment with *Escherichia coli*. *Genome Biol. Evol.* <https://doi.org/10.1093/gbe/evx064>.
- Mayr, E. 1970. Populations, species, and evolution: an abridgment of animal species and evolution. Harvard Univ. Press, Harvard.
- Meyer, J. R., D. T. Dobias, S. J. Medina, L. Servilio, A. Gupta, and R. E. Lenski. 2016. Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science* 354:1301–1304.
- Meyer, J. R., D. T. Dobias, J. S. Weitz, J. E. Barrick, R. T. Quick, and R. E. Lenski. 2012. Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science* 335:428–432.
- Peacock, T. P., D. J. Benton, J. James, J.-R. Sadeyen, P. Chang, J. E. Sealy, J. E. Bryant, S. R. Martin, H. Shelton, and W. S. Barclay. 2017. Immune escape variants of H9N2 influenza viruses containing deletions at the hemagglutinin receptor binding site retain fitness in vivo and display enhanced zoonotic characteristics. *J. Virol.* 91:e00218–e00217.
- Petrie, K. L., N. D. Palmer, D. T. Johnson, S. J. Medina, S. J. Yan, V. Li, A. R. Burmeister, and J. R. Meyer. 2018. Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. *Science* 359:1542–1545.
- Porcek, N. B., and K. N. Parent. 2015. Key residues of *S. flexneri* OmpA mediate infection by bacteriophage Sf6. *J. Mol. Biol.* 427:1964–1976.
- Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.
- Pupko, T., R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl 1):S71–S77.
- Rabosky, D. L., and E. E. Goldberg. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. Biol.* 64:340–355.
- Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes. 2004. The causes and consequences of HIV evolution. *Nat. Rev. Genet.* 5:52.
- Rohwer, F., and R. Edwards. 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.* 184:4529–4535.
- Rozen, D. E., D. Schneider, and R. E. Lenski. 2005. Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. *J. Mol. Evol.* 61:171–180.
- Salverda, M. L., J. A. G. De Visser, and M. Barlow. 2010. Natural evolution of TEM-1 β -lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiol. Rev.* 34:1015–1036.
- Shi, Y., Y. Wu, W. Zhang, J. Qi, and G. F. Gao. 2014. Enabling the “host jump”: structural determinants of receptor-binding specificity in influenza A viruses. *Nat. Rev. Microbiol.* 12:822–831.
- Song, H., J. Qi, H. Xiao, Y. Bi, W. Zhang, Y. Xu, F. Wang, Y. Shi, and G. F. Gao. 2017. Avian-to-human receptor-binding adaptation by influenza a virus hemagglutinin H4. *Cell Rep.* 20:1201–1214.
- Spielman, S. J., and S. L. Kosakovsky Pond. 2018. Relative evolutionary rate inference in HyPhy with LEISR. *PeerJ* 6:e4339.
- Stern, A., M. Te Yeh, T. Zinger, M. Smith, C. Wright, G. Ling, R. Nielsen, A. Macadam, and R. Andino. 2017. The evolutionary pathway to virulence of an RNA virus. *Cell* 169:35–46. e19.
- Suzek, B. E., Y. Wang, H. Huang, P. B. McGarvey, and C. H. Wu. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31:926–932.
- Tzarum, N., R. P. de Vries, W. Peng, A. J. Thompson, K. M. Bouwman, R. McBride, W. Yu, X. Zhu, M. H. Verheije, J. C. Paulson, and I. A. Wilson. 2017. The 150-loop restricts the host specificity of human H10N8 influenza virus. *Cell Rep.* 19:235–245.
- Wang, H. H., and G. M. Church. 2011. Multiplexed genome engineering and genotyping methods applications for synthetic biology and metabolic engineering. *Methods Enzymol.* 498:409–426.
- Wang, H. H., F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu, C. R. Forest, and G. M. Church. 2009. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460:894–898.
- Weinreich, D. M., Y. Lan, C. S. Wylie, and R. B. Heckendorn. 2013. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* 23:700–707.
- Wichman, H. A., L. A. Scott, C. D. Yarber, and J. J. Bull. 2000. Experimental evolution recapitulates natural evolution. *Philos. Trans. R Soc. Lond. Ser B Biol. Sci.* 355:1677–1684.

Associate Editor: V. Cooper
Handling Editor: M. Servedio

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Oligonucleotides (oligos) used to edit λ genomes.

Table S2. MAGE experiment design for 19-mutation library.

Figure S1. Occurrence of engineered MAGE substitutions in 1207 natural J sequences.

Figure S2. Phylogeny and alignment for residues 960–1132 of J protein.

Figure S3. Evolutionary rates for sites in the specificity region, calculated using LEISR.

Figure S4. Distribution of \log_{10} non-zero branch lengths in the full phylogeny for residues 960–1132 including all 1,207 J homologs.

Figure S5. Visualization of geodesic distances between λ protein phylogenies by metric multidimensional scaling.

Data S1. Alignment of 1207 full-length J homologs analyzed in this paper.