Viral gain-of-function experiments uncover residues under diversifying selection in nature

Rohan Maddamsetti[1], Daniel T. Johnson[2], Stephanie J. Spielman[3], Katherine L. Petrie[2,4]
Debora S. Marks[1], Justin R. Meyer[2*]

1 – Department of Systems Biology, Harvard Medical School, Boston, MA, USA
2 – Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA
3 – Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA
4 – Earth-Life Science Institute, Tokyo Institute of Technology, Japan
* – jrmeyer@ucsd.edu

Viral gain-of-function mutations are commonly observed in the laboratory; however, it is unknown whether those mutations also evolve in nature. We identify two key residues in the host recognition protein of bacteriophage λ that are necessary to exploit a new receptor; both residues repeatedly evolved among homologs from environmental samples. Our results provide evidence for widespread host-shift evolution in nature and a proof of concept for integrating experiments with genomic epidemiology.

Many viruses can expand their host range with a few mutations[1-3] that enable the exploitation of new cellular receptors[2,3]. Such mutations may be the first steps toward an epidemic outbreak; this observation has driven an expansion of theoretical[4], experimental and surveillance studies of host-range shifts in emergent pathogens, including avian influenza[5-8], coronaviruses[9,10], HIV[11] and ebolavirus[12]. Ideally, laboratory evolution experiments could be used to accelerate our understanding of how viruses expand their host range. However, it is not clear whether viral evolution observed under the chemical and physical environments of a laboratory can faithfully inform us about viral evolution 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 virus gain-of-function laboratory experiments[13] on the grounds that they would tell us little about real-world viral evolution at the risk of constructing viral strains that are pandemic to new hosts. Here, we use a harmless virus, bacteriophage λ, to demonstrate how gain-of-function experiments can identify mutations that mimic those that occur in nature: we find that two amino acid residues that are critical for gain of function in the laboratory recurrently evolve in nature.

Typical laboratory strains of λ infect *Escherichia coli* by binding to the outer membrane protein LamB[14], but the phage rapidly evolves in the laboratory to exploit a different membrane protein, OmpF[3]. Since OmpF is not the usual *E. coli* receptor for λ, 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 non-synonymous mutations in the host-recognition gene *J*. Each OmpF$^+$ isolate had between 4 and 10 single nucleotide substitutions in *J*, and none had insertions or deletions (indels). 97% of the substitutions in 24 independently evolved OmpF$^+$ λ phage occurred in the J protein[3] between residues 960—1132, which we call the 'specificity region' of J (Figure 1A).

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[3]. In this study, we directly tested this hypothesis by using Multiplexed Automated Genome Engineering (MAGE)[15] to construct two combinatorial genetic libraries of the *J* mutations that evolved in Meyer *et al.* (2012). We constructed the first library by focusing on 10 commonly evolved *J* mutations[3], from which we sequenced 33 OmpF$^+$ isolates. All strains possessed the target mutation at residue 1012 and both target mutations at residue 1107; by contrast, some strains lacked a fourth mutation between residues 990 and 1000 (Figure 1). We call the mutations at residues 1012 and 1107 the "canonical mutations".

To test whether the three canonical 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 laboratory-evolved OmpF$^+$ phage. To find what further mutations might be needed to confer the OmpF$^+$ phenotype, we constructed a second MAGE library using the phage with the three canonical *J* mutations as the

baseline, and random combinations of 19 other *J* mutations found in the OmpF$^+$ λ evolved by Meyer *et al.* (2012). Screening this second library produced 88 OmpF$^+$ isolates. One OmpF$^+$ isolate had just a single extra mutation in amino acid position 1083 in addition to the three canonical mutations (Figure 1). However, the majority of the OmpF$^+$ genotypes 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 laboratory experiments.

We then asked whether natural J sequences reflected the host-shift evolutionary dynamics seen in our gain-of-function experiments. We collected and aligned full-length homologous J protein sequences from UniRef100[16] (1,207 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*.

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: 30% of the total amino acid variation occurred in the specificity region, despite only being 15% of the total length of J (Figure 2A). 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 engineered into the MAGE libraries were significantly more variable than randomly chosen groups of 17 residues from J (non-parametric bootstrap: $p < 10^{-5}$) and from the specificity region 960-1132 (non-parametric bootstrap: $p = 0.0054$) showing the experiments had identified evolutionary hotspots. However, the specific 17 amino acid substitutions evolved in the laboratory were not common in natural homologs suggesting that the selection experiments had the resolution to predict where changes would evolve, but not the exact change (Supplementary Figure 1).

Focusing in on the two residues critical for the OmpF$^+$ gain of function, 1012 and 1107, we find that they alone are more variable than random pairs of sites in J (non-parametric bootstrap: $p = 0.00101$) and the specificity region (non-parametric bootstrap: $p = 0.036$). But what is more remarkable is the observation that in-frame indels are common in the regions around sites 1012 and 1107 (Figure 2B; Supplementary Data 1). We measured indel variation over J homologs using gap entropy: a two-state entropy counting each position in each aligned

sequence as a gap or non-gap. Residues 1010–1013 and 1107–1109 have the highest gap entropy in J, ranging from 0.799 to 0.893 out of a maximum of 1.000 (Figure 2B; Supplementary Data 2). For comparison, the typical amino acid site in the J alignment had no indels (mode and median gap entropy = 0). That our experiments would correctly identify both regions with the most natural indel variation by chance is extremely unlikely (given the size of each region and the protein length ($2\times\frac{4}{1132}\times\frac{3}{1132}$) = $1.87\cdot10^{-5}$). Indels can have large beneficial effects on proteins, including altering specificity by changing surface loops[14,17], or causing structural rearrangements that improve function[18].

To analyze J protein evolutionary dynamics in more depth, we built maximum-likelihood phylogenetic trees for J (Figure 2C) and its specificity region (Figure 2D). The trees are very different (normalized Robinson-Foulds distance = 0.94 out of 1.00), indicating that recombination has caused the history of the specificity region to differ from the history of the rest of the protein. This means that the specificity region is an evolutionary module that commonly recombines and circulates as a unit in the phage population. This observation supports the notion that this region has elevated rates of evolution and diversification.

Using the second phylogeny, we calculated site-specific evolutionary rates for the specificity region. The 17 residues studied in the MAGE experiments evolve faster than equally-sized random samples taken from the specificity region (non-parametric bootstrap, $p = 0.0067$) and residues 1012 and 1107 evolve faster than random pairs of sites sampled from the specificity region (non-parametric bootstrap, $p = 0.0073$).

Together, the accelerated rate of change and increased entropy at residues 1012 and 1107 suggest that these sites have experienced strong diversifying selection. Our combinatorial genetic studies showed that these particular residues determine receptor tropism, so it is reasonable to conclude that their unique evolution is due to host-range evolution. Mathematical theory has demonstrated that receptor-use evolution should lead to diversification since accessing new host types alters the phages' ecological niches[19]. Indeed, diversification was observed in a related experiment where λ evolved to exploit two host cell types with different receptors[20]. To test whether indels at 1012 and 1107 cause diversification, we compared the branches where indels occur versus all others. If the indels cause ecological differentiation, then they should reduce intraspecific competition for hosts and facilitate the long-term maintenance of distinct evolutionary lineages. Indeed, the branches on which indels occur are significantly longer than

the other branches of the specificity-region phylogeny (Kruskal-Wallis test, $p < 10^{-8}$; visualized by the discrete indel clades colored in lavender and light green in Figure 2D).

The congruence we find between laboratory and natural evolution in λ contrasts with work showing that beneficial mutations in Richard Lenski's long-term experiment anti-correlate[21] with natural protein variation in *E. coli*. By design, Lenski's experiment lacks the complex and variable selection pressures found in *E. coli*'s natural gut environment, leading 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 phage in the wild. Notably, another study on experimental evolution of bacteriophage has observed residues under positive selection in both the lab and nature[22].

Together, these studies show that selection experiments on viruses cultured in the laboratory can inform evolution in nature. Based on the patterns of J sequence variation observed, 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 combat host shifts. In particular, worrisome mutations can be identified with functional genetic experiments as described here or with other laboratory techniques such as deep mutational scanning[23]. This information can be combined with genomic surveillance efforts underway[24,25] to devise more effective disease management strategies to eradicate problematic strains.

Methods

MAGE *experiments*

We modified λ strain cI857 (provided by Ing-Nang Wang, SUNY Albany) integrated into the genome of HWEC106 (provided by Harris Wang, Columbia University). We engineered the J substitutions into λ by using MAGE[15,26] on strain HWEC106. MAGE uses the λ-red recombineering system provided on the pKD46 plasmid[27]. For a description of the oligos used, see Supplementary Table S1.

For the first, 10-mutation library, we ran 18 rounds of MAGE and then screened for OmpF$^+$ isolates. For the second library, we edited λ with oligos 'a3034g' and 'g3319a t3321a' to introduce the canonical three mutations. Next, we performed a number of different MAGE trials in order to maximize the number of unique alleles we observed. See Supplementary Table S2 for our MAGE strategy.

To screen the edited phage for OmpF$^+$ genotypes, we induced the MAGE lysogen libraries and plated the phage on lawns of *lamB*$^-$ *E. coli* strain JW3996 from the KEIO collection[28]. Plaques were picked from the lawns and the C-terminus of the J gene was Sanger sequenced at the Genewiz La Jolla, CA facility. Unpurified PCR products (Forward primer: 5' CGCATCGTTCACCTCTCACT; Reverse primer: 5' CCTGCGGGCGGTTTGTCATTT) were submitted.

Analysis of natural J *variation*

We used the *evcouplings* pipeline[29] to generate a jackhmmer[30] alignment of 1,207 full-length J 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. When analyzing the specificity region (say, constructing a phylogeny), we used residues 960-1132 of this full-length alignment.

We used RAxML[31] to generate maximum-likelihood phylogenies using automatically selected amino acid substitution matrices (PROTGAMMAAUTO) based on model fit. To account for recombination when estimating site-specific evolutionary rates in the specificity region, we first ran a modified SBP algorithm[32] on the specificity region. We found a putative recombination breakpoint at residue 1000, supported by Akaike's information criterion but not the Bayesian information criterion. We therefore partitioned the specificity region at residue 1000 and re-calculated trees for each partition using FastTree[33] with an LG substitution matrix[34]. We next calculated site-specific evolutionary rates with LEISR[35], a scalable implementation of Rate4Site[36] that accounts for recombination breakpoints, using an LG substitution matrix. We calculated the Robinson-Foulds distance between our maximum-likelihood phylogenies for J and

its specificity region using *ete3-compare*[37]. We divided the Robinson-Foulds distance (2276.00) by its maximum possible value (2412.00) to get the normalized distance (0.94).

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.

For all non-parametric bootstrap calculations, we used 100,000 bootstraps, and chose groups of sites without replacement. For example, when testing for high entropy in the 17 MAGE residues, we generated an empirical null distribution by randomly choosing 17 different residues for one sample, and resampling 100,000 times. We used the same procedure for our evolutionary-rate tests.

Data Availability: All data and analysis code for this project are available at the Dryad digital repository [doi: pending publication]

References:

1 Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J. & Jiggins, F. M. The evolution and genetics of virus host shifts. *PLoS Pathog* **10**, e1004395, doi:10.1371/journal.ppat.1004395 (2014).

2 Imai, M. *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**, 420-428, doi:10.1038/nature10831 (2012).

3 Meyer, J. R. *et al.* Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science* **335**, 428-432, doi:10.1126/science.1214449 (2012).

4 Antia, R., Regoes, R. R., Koella, J. C. & Bergstrom, C. T. The role of evolution in the emergence of infectious diseases. *Nature* **426**, 658-661, doi:10.1038/nature02104 (2003).

5 Shi, Y., Wu, Y., Zhang, W., Qi, J. & Gao, G. F. Enabling the 'host jump': structural determinants of receptor-binding specificity in influenza A viruses. *Nature reviews. Microbiology* **12**, 822-831, doi:10.1038/nrmicro3362 (2014).

6 Song, H. *et al.* Avian-to-Human Receptor-Binding Adaptation by Influenza A Virus Hemagglutinin H4. *Cell reports* **20**, 1201-1214, doi:10.1016/j.celrep.2017.07.028 (2017).

7 Koel, B. F. *et al.* Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* **342**, 976-979, doi:10.1126/science.1244730 (2013).

8 Linster, M. *et al.* Identification, characterization, and natural selection of mutations driving airborne transmission of A/H5N1 virus. *Cell* **157**, 329-339, doi:10.1016/j.cell.2014.02.040 (2014).

9 de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nature reviews. Microbiology* **14**, 523-534, doi:10.1038/nrmicro.2016.81 (2016).

10 Lu, G. *et al.* Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. *Nature* **500**, 227-231, doi:10.1038/nature12328 (2013).
11 Rambaut, A., Posada, D., Crandall, K. A. & Holmes, E. C. The causes and consequences of HIV evolution. *Nature Reviews Genetics* **5**, 52, doi:10.1038/nrg1246 (2004).
12 Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature* **538**, 193-200, doi:10.1038/nature19790 (2016).
13 Casadevall, A. & Imperiale, M. J. 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, doi:10.1128/mBio.01730-14 (2014).
14 Chatterjee, S. & Rothenberg, E. Interaction of bacteriophage l with its E. coli receptor, LamB. *Viruses* **4**, 3162-3178, doi:10.3390/v4113162 (2012).
15 Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894-898, doi:10.1038/nature08187 (2009).
16 Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* **31**, 926-932, doi:10.1093/bioinformatics/btu739 (2015).
17 Porcek, N. B. & Parent, K. N. Key residues of S. flexneri OmpA mediate infection by bacteriophage Sf6. *Journal of molecular biology* **427**, 1964-1976, doi:10.1016/j.jmb.2015.03.012 (2015).
18 Arpino, J. A., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J. & Jones, D. D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure (London, England : 1993)* **22**, 889-898, doi:10.1016/j.str.2014.03.014 (2014).
19 Weitz, J. S., Hartman, H. & Levin, S. A. Coevolutionary arms races between bacteria and bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9535-9540, doi:10.1073/pnas.0504062102 (2005).
20 Meyer, J. R. *et al.* Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science* **354**, 1301-1304, doi:10.1126/science.aai8446 (2016).
21 Maddamsetti, R. *et al.* Core Genes Evolve Rapidly in the Long-term Evolution Experiment with Escherichia coli. *Genome biology and evolution*, doi:10.1093/gbe/evx064 (2017).
22 Wichman, H. A., Scott, L. A., Yarber, C. D. & Bull, J. J. Experimental evolution recapitulates natural evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **355**, 1677-1684, doi:10.1098/rstb.2000.0731 (2000).
23 Bloom, J. D. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology direct* **12**, 1, doi:10.1186/s13062-016-0172-z (2017).
24 Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369-1372, doi:10.1126/science.1259657 (2014).
25 Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401-405, doi:10.1038/nature22400 (2017).

26 Wang, H. H. & Church, G. M. Multiplexed genome engineering and genotyping methods applications for synthetic biology and metabolic engineering. *Methods in enzymology* **498**, 409-426, doi:10.1016/b978-0-12-385120-8.00018-8 (2011).
27 Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6640-6645, doi:10.1073/pnas.120163297 (2000).
28 Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* **2**, 2006.0008, doi:10.1038/msb4100050 (2006).
29 Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature biotechnology* **35**, 128-135, doi:10.1038/nbt.3769 (2017).
30 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS computational biology* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).
31 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
32 Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular biology and evolution* **23**, 1891-1901, doi:10.1093/molbev/msl051 (2006).
33 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).
34 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Molecular biology and evolution* **25**, 1307-1320, doi:10.1093/molbev/msn067 (2008).
35 Spielman, S. J. & Pond, S. L. K. Relative evolutionary rate inference in HyPhy with LEISR. *bioRxiv* (2017).
36 Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S71-77 (2002).
37 Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular biology and evolution* **33**, 1635-1638, doi:10.1093/molbev/msw046 (2016).

Acknowledgments: We thank John Ingraham, Adam Riesselman, Kelly Brock, David Ding, and Alita Burmeister for helpful discussions. K.L.P. is 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.

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

Competing Interests: The authors declare no competing financial interests.

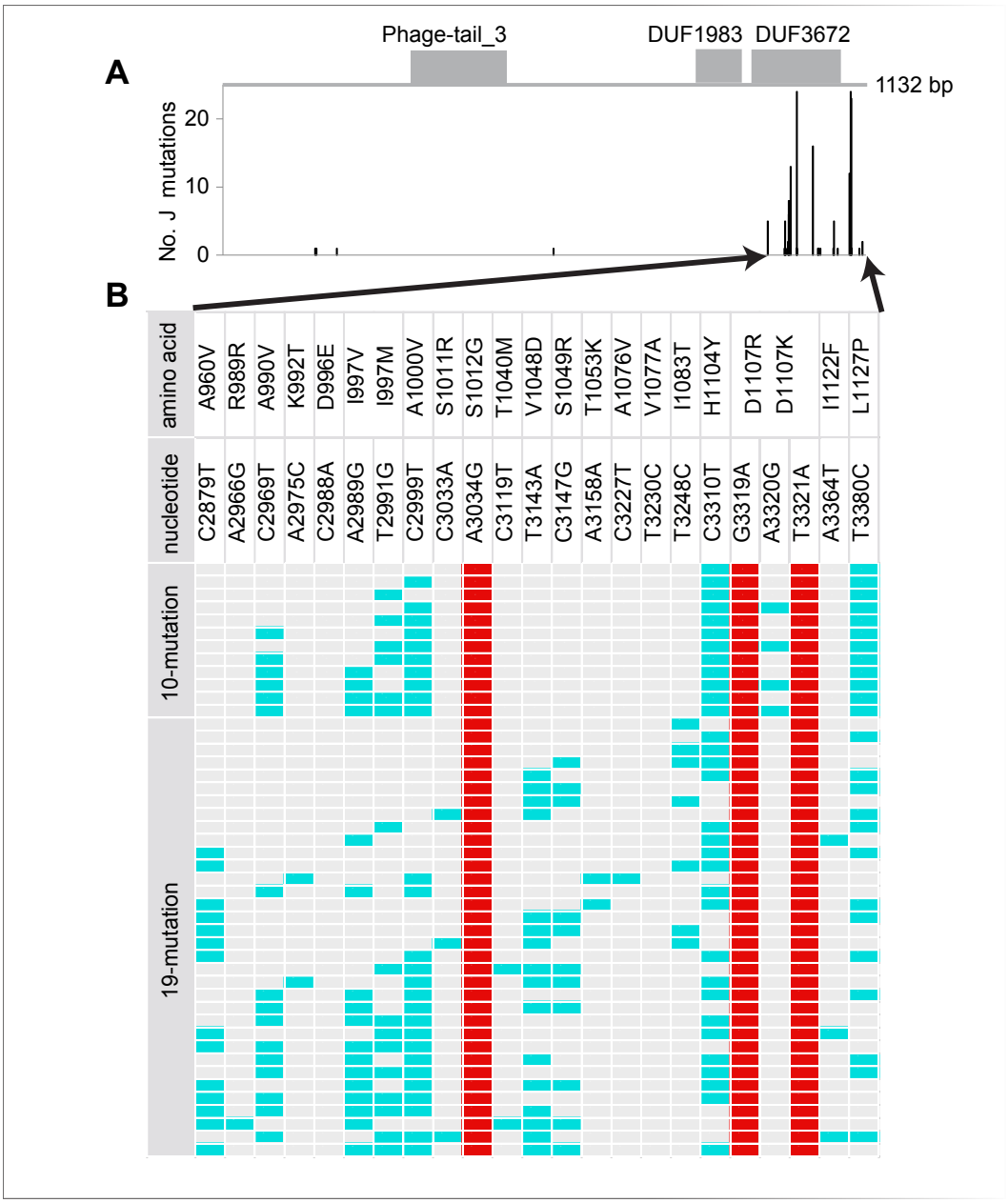

Figure 1 | A) Distribution of J mutations evolved *en route* to OmpF$^+$ and B) synthetic phage genotypes capable of using OmpF**.** A) Mutations summed across 24 genomes independently evolved to exploit OmpF in Meyer *et al.* (2012). Protein domains as annotated in the Pfam database are shown in gray. The majority of mutations either occur in the DUF3672 domain, or in the C-terminus past the annotated boundary of DUF3672. B) Synthetic OmpF$^+$ genotypes indicated by colored cells (red marks the canonical three) observed after combinatorial engineering of 10 common mutations or 19 mutations when the canonical 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.

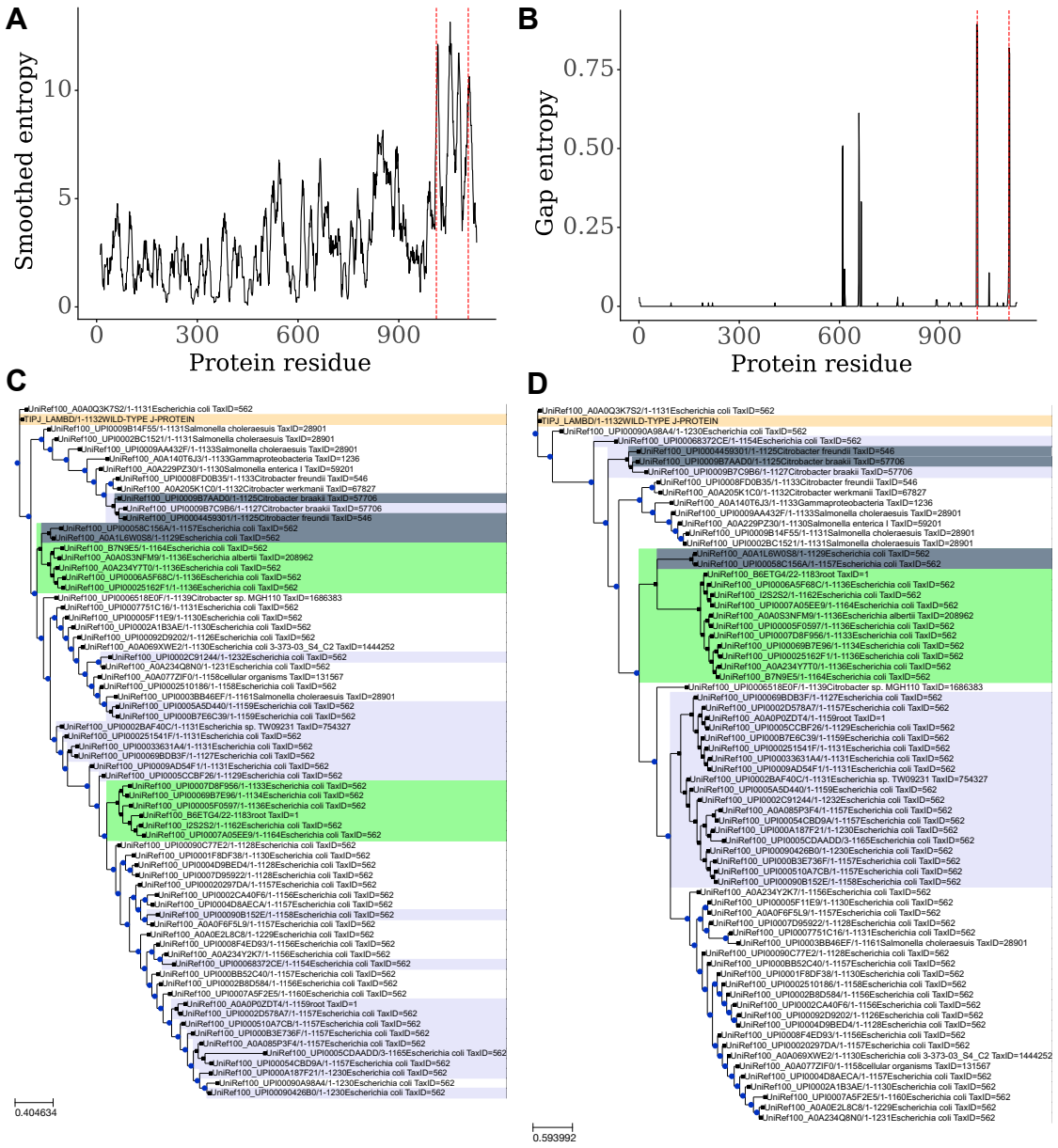

Figure 2 | A) Sequence entropy averaged over a 10-residue window in J. Dashed red lines indicate residues 1012 and 1107. B) Gap entropy over the J protein. Dashed red lines indicate residues 1012 and 1107. C) Phylogeny for J protein. Only a subset of the operational taxonomic units (OTUs) are displayed: just OTUs with sequences ≥ 5% divergence from each other. The wild-type J protein sequence is labeled in tan. Clades with indels at residue 1012 are labeled in lavender. Clades with indels at residue 1107 are labeled in light green. Sequences with indels at both positions are labeled in light steel gray. D) Phylogeny for residues 960–1132 of J protein. OTUs have at least 5% divergence from each other, and are labeled as in (C).

Supplementary Table S1 | Oligonucleotides (oligos) used to edit λ genomes. Oligos were designed slightly differently for the two libraries. For the first, we inserted a synonymous change near each target. This was done as a silent tag to improve our confidence that an edit had been made successfully. We did not do this for the second library because we found that all intentional nonsynonymous changes were accompanied by the synonymous change. Additionally, for the first library, when mutations fell within 90 bases, we designed a separate oligo for each possible mutation combination. For the second library, we ordered a mixed pool of oligos synthesized that had all combinations of the mutations. Asterisks indicate phosphorothioated bases.

10-mutation library		
1_c2969tg2970c	c2969t g2970c	A*A*A*G*CCGCGCTCGCCGCCTTTACAATGTCCCCGACGAT TTTTTCgaCCCTCAGCGTACCGTTTATCGTACAGTTTTCAGCT ATCGTCACA
2_c2969tg2970t	c2969t g2970t	A*A*A*G*CCGCGCTCGCCGCCTTTACAATGTCCCCGACGAT TTTTTCaaCCCTCAGCGTACCGTTTATCGTACAGTTTTCAGCT ATCGTCACA
3_a2989gg2985t	a2989g g2985t	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCCGCCTTTACAAcGTCaCCGACGATTT TTTCCGCCCT
4_a2989gc2988t	a2989g c2988t	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCCGCCTTTACAAcaTCCCCGACGATTT TTTCCGCCCT
5_t2991ga2994c	t2991g a2994c	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCCGCCTTgACcATGTCCCCGACGATTT TTTCCGCCCT
6_t2991ga2994g	t2991g a2994g	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCCGCCTTcACcATGTCCCCGACGATTT TTTCCGCCCT
7_c2999tg3000a	c2999t g3000a	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCtaCCTTTACAATGTCCCCGACGATTT TTTCCGCCCT
8_c2999tg3000c	c2999t g3000c	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCgaCCTTTACAATGTCCCCGACGATTT TTTCCGCCCT
9_a2989gg2985tt2991ga2994c	a2989g g2985t t2991g a2994c	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCCGCCTTgACcAcGTCaCCGACGATTT TTTCCGCCCT
10_a2989gc2988tc2999tg3000a	a2989g c2988t c2999t g3000a	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCtaCCTTTACAAcaTCCCCGACGATTTT TTCCGCCCT
11_t2991ga2994gc2999tg3000c	t2991g a2994g c2999t g3000c	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCgaCCTTcACcATGTCCCCGACGATTT TTTCCGCCCT
12_a2989gc2988tt2991ga2994cc2999tg3000c	a2989g c2988t t2991g a2994c c2999t g3000c	A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCgaCCTTcACcAcaTCCCCGACGATTTT TTCCGCCCT
13_a3034gc3033t	a3034g c3033t	C*A*T*C*GGTCACGGTGACAGTACGGGTACCTGACGGCCA GTCCACACcaCTTTCACGCTGGCGCGGAAAAGCCGCGCTCG CCGCCTTTACAA
14_a3034gt3036a	a3034g t3036a	C*A*T*C*GGTCACGGTGACAGTACGGGTACCTGACGGCCA GTCCACtCcGCTTTCACGCTGGCGCGGAAAAGCCGCGCTCG CCGCCTTTACAA
15_c3310tg3309a	c3310t g3309a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATATCTGCCGAATatCGTGTGGACGTA AGCGTGAACGT
16_c3310tg3309t	c3310t g3309t	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATATCTGCCGAATaaCGTGTGGACGTA AGCGTGAACGT

17_g3319ag3315c	g3319a g3315c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATATtTGCgGAATGCCGTGTGGACGTA AGCGTGAACGT
18_g3319ag3315a	g3319a g3315a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATATtTGCtGAATGCCGTGTGGACGTA AGCGTGAACGT
19_a3320ga3318c	a3320g a3318c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATAcCgGCCGAATGCCGTGTGGACGTA AGCGTGAACGT
20_a3320ga3318g	a3320g a3318g	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATAcCcGCCGAATGCCGTGTGGACGTA AGCGTGAACGT
21_t3321at3324a	t3321a t3324a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGtATtTCTGCCGAATGCCGTGTGGACGTA AGCGTGAACGT
22_t3321at3324c	t3321a t3324c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGgATtTCTGCCGAATGCCGTGTGGACGTA AGCGTGAACGT
23_g3319ag3315ca3320ga3318c	g3319a g3315c a3320g a3318c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATActgGCgGAATGCCGTGTGGACGTA AGCGTGAACGT
24_g3319ag3315at3321at3324c	g3319a g3315a t3321a t3324c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGgATtTtTGCtGAATGCCGTGTGGACGTAA GCGTGAACGT
25_a3320ga3318ct3321at3324a	a3320g a3318c t3321a t3324a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGtATtcCgGCCGAATGCCGTGTGGACGTAA GCGTGAACGT
26_g3319ag3315ca3320ga3318ct3321at3324c	g3319a g3315c a3320g a3318c t3321a t3324c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGgATtctgGCgGAATGCCGTGTGGACGTAA GCGTGAACGT
27_t3380cg3378a	t3380c g3378a	T*T*C*C*CGGACGAACCTCTGTAACACACTCAGACCACGCT GATGCCCgGtGCCTGTTTCTTAATCACCATAACCTGCACATC GCTGGCAAAC
28_t3380cg3378c	t3380c g3378c	T*T*C*C*CGGACGAACCTCTGTAACACACTCAGACCACGCT GATGCCCgGgGCCTGTTTCTTAATCACCATAACCTGCACATC GCTGGCAAAC
29_g3381a	g3381a	T*T*C*C*CGGACGAACCTCTGTAACACACTCAGACCACGCT GATGCCtAGCGCCTGTTTCTTAATCACCATAACCTGCACATC GCTGGCAAAC
31_wt		A*A*A*G*CCGCGCTCGCCGCCTTTACAATGTCCCCGACGAT TTTTTCCGCCCTCAGCGTACCGTTTATCGTACAGTTTTCAGCT ATCGTCACA
32_wt		A*C*C*T*GACGGCCAGTCCACACTGCTTTCACGCTGGCGCG GAAAAGCCGCGCTCGCCGCCTTTACAATGTCCCCGACGATT TTTTCCGCCCT
33_wt		C*A*T*C*GGTCACGGTGACAGTACGGGTACCTGACGGCCA GTCCACACTGCTTTCACGCTGGCGCGGAAAAGCCGCGCTCG CCGCCTTTACAA

35_wt		C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATATCTGCCGAATGCCGTGTGGACGTA AGCGTGAACGT
36_wt		T*T*C*C*CGGACGAACCTCTGTAACACACTCAGACCACGCT GATGCCCAGCGCCTGTTTCTTAATCACCATAACCTGCACATC GCTGGCAAAC
37_c3310tg3309ag3319ag3315a	c3310t g3309a g3319a g3315a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATATTtTGCtGAATatCGTGTGGACGTAA GCGTGAACGT
38_c3310tg3309ta3320ga3318g	c3310t g3309t a3320g a3318g	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATAcCcGCCGAATaaCGTGTGGACGTA AGCGTGAACGT
39_c3310tg3309at3321at3324a	c3310t g3309a t3321a t3324a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGtATtTCTGCCGAATatCGTGTGGACGTAA GCGTGAACGT
40_c3310tg3309tg3319ag3315ca3320ga3318g	c3310t g3309t g3319a g3315c a3320g a3318g	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATActcGCgGAATaaCGTGTGGACGTAA GCGTGAACGT
41_c3310tg3309ag3319ag3315at3321at3324c	c3310t g3309a g3319a g3315a t3321a t3324c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGgATtTtTGCtGAATatCGTGTGGACGTAA GCGTGAACGT
42_c3310tg3309ta3320ga3318gt3321at3324a	c3310t g3309t a3320g a3318g t3321a t3324a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGtATtcCcGCCGAATaaCGTGTGGACGTAA GCGTGAACGT
43_c3310tg3309ag3319ag3315ca3320ga3318ct3321at3324c	c3310t g3309a g3319a g3315c a3320g a3318c t3321a t3324c	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGgATtctgGCgGAATatCGTGTGGACGTAA GCGTGAACGT
3 canonical mutations		
a3034g	a3034g	C*A*T*C*GGTCACGGTGACAGTACGGGTACCTGACGGCCA GTCCACACcGCTTTCACGCTGGCGCGGAAAAGCCGCGCTCG CCGCCTTTACAA
g3319a t3321a	g3319a t3321a	C*T*G*T*TTCTTAATCACCATAACCTGCACATCGCTGGCAA ACGTATACGGCGGAATTtTtTGCCGAATGCCGTGTGGACGTA AGCGTGAACGT
19-mutation library		
OLIGO#1: J_01_C2879T	C2879T	T*A*C*T*GAGCGTCCCGGAGTTCGCATTCACACTGCCACTG ATATCCaCATTTTTAGCGGTCAGCTTTCCGTCCGGTGTCAGG GAAAAGGCCG
OLIGO#2: J_02_2966R_67Y_75M_88M_89R_91K_2999Y	A2966R, C2967Y, A2975M, C2988M, A2989R, T2991K, C299Y	C*G*G*A*AAAGCCGCGCTCGCCrCCTTTACmAykTCCCCGA CGATTkTTTCCrCCCyCAGCGTACCGTTTATCGTACAGTTTTC AGCTATCGT
OLIGO#3: J_03_C3033A	C3033A	C*T*G*G*CGATCAAAAGGATGGTCATCGGTCACGGTGACA GTACGGGTACCTGACGGCCAGTCCACACctCTTTCACGCTGG CGCGGAAAAGC

OLIGO#4: J_04_C3119Y_43W_47S_58M Corrected	C3119Y, T3143W, C3147S, C3158M	T*T*C*A*TCAGTACTTTCAGATAACACATCGAATACKTTGTCCTGCCSCTGWCAGTACGCTTACTTCCGCGAAACRTCAGCGGAAGCACCACT
OLIGO#5: J_05_C3227Y_30Y_48Y	C3227Y, T3230Y, T3248Y	A*T*C*A*CGTTTCCCCGACCCGCTGGCATGTCAACArTACGGGAGAACACCTGTrCCrCCTCGTTCGCCGCGCCATCATAAATCACCGCACCG
OLIGO#6: J_06_C3310T	C3310T	A*A*A*C*GTATACGGCGGAATtTtTGCCGAATaCCGTGTGGACGTAAGCGTGAACGTCAGGATCACGTTTCCCCGACCCGCTGGCATGTCAAC
OLIGO#7: J_07_A3364W_80Y Corrected	A3364W, T3380Y	A*A*A*A*CGCCCGTTCCCGGACGAACCTCTGTAACACACTCAGACCACGCTGATGCCCRGCGCCTGTTTCTTAAWCACCATAACCTGCACATC

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

Oligos	No. cycles	No. replicates	No. plaques sequenced	Notes
All 7	1	2	16	This was a pilot study
All 7	2	2	15	This was a pilot study
OLIGO#1	1	1	0	No plaques formed on *lamB*$^-$ lawn
OLIGO#2	1	1	4	
OLIGO#3	1	1	0	No plaques formed on *lamB*$^-$ lawn
OLIGO#4	1	1	0	No plaques formed on *lamB*$^-$ lawn
OLIGO#5	1	1	4	
OLIGO#6	1	1	0	No plaques formed on *lamB*$^-$ lawn
OLIGO#7	1	1	0	No plaques formed on *lamB*$^-$ lawn
All 7	1	3	10	
All 7	2	3	12	
All 7	3	3	8	
All 7	4	3	10	
All 7	5	3	15	
All 7	6	3	13	
All 7	7	3	11	
All 7	8	3	11	

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

960	990	1000	1040	1048	1053	1076	1083	1104	1107	
.	.	.	.	.	.	.	.	Y	.	4
.	.	.	.	.	.	.	T	.	.	5
.	.	.	.	.	.	V	.	.	.	117
.	.	.	.	.	.	V	.	.	K	2
.	.	.	.	.	.	V	T	.	.	23
.	.	.	.	D	.	.	.	.	.	266
.	.	.	.	D	K	.	.	.	.	1
.	.	.	M	.	.	.	.	.	.	8
.	.	.	M	.	.	.	.	.	K	1
.	.	V	.	.	.	V	T	.	.	1
.	V	.	.	.	.	.	.	.	.	1
V	.	.	.	D	.	.	.	.	.	1

Supplementary Data File 1 | Alignment of 1207 full-length J homologs analyzed in this paper.

Supplementary Data File 2 | Spreadsheet of gap entropies for each site in the alignment of 1207 full-length J homologs.