PROTEIN EVOLUTION

Pervasive degeneracy and epistasis in a protein-protein interface

Anna I. Podgornaia[1,2]* and Michael T. Laub[2,3]†

Mapping protein sequence space is a difficult problem that necessitates the analysis of 20^N combinations for sequences of length *N*. We systematically mapped the sequence space of four key residues in the *Escherichia coli* protein kinase PhoQ that drive recognition of its substrate PhoP. We generated a library containing all 160,000 variants of PhoQ at these positions and used a two-step selection coupled to next-generation sequencing to identify 1659 functional variants. Our results reveal extensive degeneracy in the PhoQ-PhoP interface and epistasis, with the effect of individual substitutions often highly dependent on context. Together, epistasis and the genetic code create a pattern of connectivity of functional variants in sequence space that likely constrains PhoQ evolution. Consequently, the diversity of PhoQ orthologs is substantially lower than that of functional PhoQ variants.

Protein-protein interactions drive the operation and function of cells. These interactions involve a molecular interface formed by a subset of amino acids from each protein. Interfacial residues often vary between orthologs, indicating some mutational tolerance or degeneracy (*1*, *2*), but such natural variability may not capture the full plasticity of interfaces. Thus, it remains unclear how many combinations of interface residues will support a given interaction and how these combinations are distributed and connected in sequence space (*3*) (fig. S1A). It is also unknown whether all functional variants can be reached through a series of mutations that retain function, or whether evolution is fundamentally constrained, limiting the natural diversity in orthologous proteins. Several studies have examined the mutational intermediates separating two proteins, but these typically exclude residues not present in either protein (*4*). Assays that couple genotype to phenotype along with deep sequencing have enabled the interrogation of large numbers of mutants, including saturation mutagenesis of individual positions. These deep mutational scans have also tested many double and higher-order mutants, although not comprehensively, impeding the systematic analysis of functional variants and mutational paths in sequence space (*5*–*10*).

We mapped the sequence space underlying the interface formed by bacterial two-component signaling proteins in vivo (Fig. 1A and fig. S1, B and C). *E. coli* PhoQ is a sensor histidine kinase that is stimulated by low extracellular magnesium concentrations to phosphorylate the response regulator PhoP (*11*). When not stimulated to autophosphorylate, PhoQ binds to and drives the dephosphorylation of PhoP. The interface formed by two-component signaling proteins, such as PhoQ-PhoP, involves a limited number of residues from each protein (*12*, *13*) (fig. S1B). For histidine kinases, mutating just three or four interfacial residues to match those in another kinase is often sufficient to reprogram partner specificity (*12*, *14*).

[1]Computational & Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*Present address: Merck Research Laboratories, Boston, MA 02115, USA. †Corresponding author. E-mail: laub@mit.edu

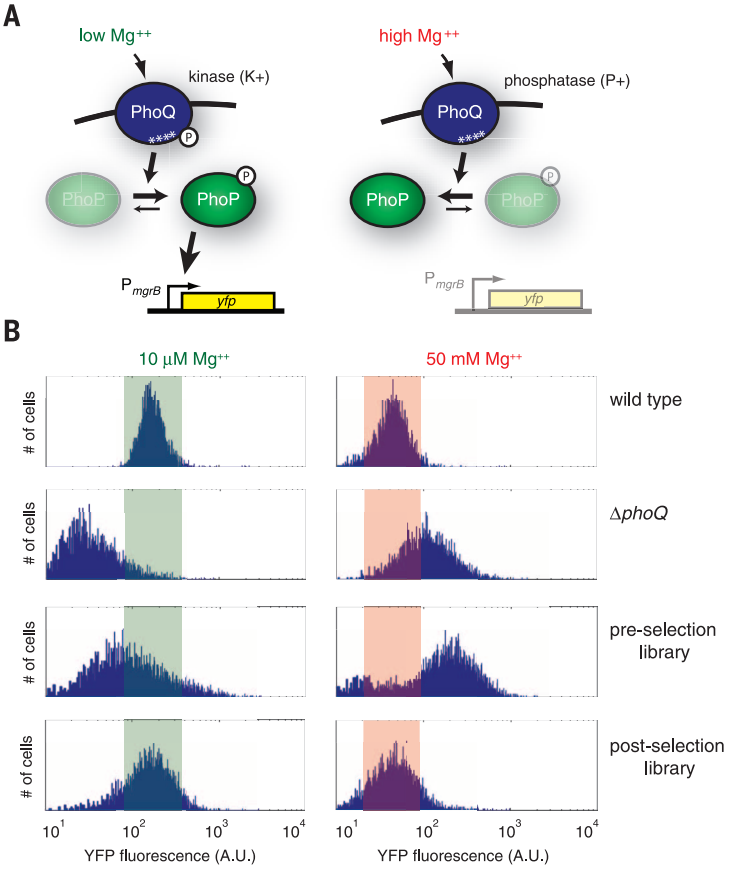

Fig. 1. Mapping sequence space. (**A**) PhoQ phosphorylates or dephosphorylates PhoP depending on extracellular magnesium concentration. White asterisks indicate interfacial residues randomized in the *phoQ* library. (**B**) YFP levels measured by flow cytometry for cells expressing wild-type *phoQ*, lacking *phoQ*, or harboring the *phoQ* library, before and after selection. Shaded regions indicate wild-type YFP levels.

For PhoQ, these key interfacial residues are Ala^{284}, Val^{285}, Ser^{288}, and Thr^{289} (AVST), which is just one of 160,000 possible combinations at these four positions. To assess the ability of each combination to promote a functional PhoQ-PhoP interface in *E. coli*, we developed a high-throughput screen using a strain where *yfp* is expressed from a PhoP-dependent promoter, P_{mgrB} (Fig. 1A and fig. S1, B and C). When extracellular Mg^{2+} concentration is low, PhoQ is predominantly a kinase, driving PhoP phosphorylation and YFP (yellow fluorescent protein) production; in high extracellular Mg^{2+}, PhoQ is mainly a phosphatase, stimulating PhoP dephosphorylation and preventing YFP production (*11*) (fig. S2A). To systematically probe the PhoQ interface, we constructed a library in which the four key interfacial residues were fully randomized and then transformed this library into a Δ*phoQ* strain harboring the P_{mgrB}-*yfp* reporter. The library was grown for 6 hours in medium with low or high extracellular Mg^{2+} to stimulate PhoQ kinase or phosphatase activity, respectively, and was then subjected to fluorescence-activated cell sorting to isolate those mutants that behaved similarly to wild-type PhoQ (fig. S3A).

This screen proved more stringent in selecting mutants with wild-type phosphatase activity because cells deficient in PhoQ kinase activity still accumulate some phosphorylated PhoP via acetyl-phosphate (fig. S2B). We therefore performed a second screen in which cells selected for phosphatase activity were starved of extracellular Mg^{2+} for 18 hours and then recovered in Mg^{2+}-replete medium for 6 hours (figs. S4 and S5). Kinase activity comparable to that of the wild type was required to induce the PhoP regulon and survive without Mg^{2+} (fig. S2C).

To identify interfacial residues that promote a PhoQ-PhoP interaction, we deep-sequenced the relevant region of *phoQ* from cells that passed our two-step selection (Fig. 1B and fig. S4). The starting library used NNS codons to randomize the four interfacial residues (where N = any nucleotide and S = G or C). Hence, the theoretical diversity of the library is 194,481, with 160,000 combinations lacking stop codons. Sequencing of the starting library indicated that 93% of 160,000 possible protein variants had more than three reads (fig. S3). For the postselection library, we used an unbiased binary classifier to identify sequences that were enriched relative to the starting library. The training set for the classifier consisted of (i) library data for the 34,481 PhoQ variants harboring one or more stop codons, each of which produces a nonfunctional PhoQ, and (ii) data collected individually for each single mutant of PhoQ at the first three positions randomized in our library. Of these 57 single point mutants, 13 exhibited nearly wild-type flow cytometry profiles and could successfully compete with the wild type under conditions of Mg^{2+} starvation (Fig. 2A and fig. S6). The binary classifier identified 1659 unique, functional PhoQ variants with an estimated false positive rate of <0.01% and false negative rate of ~7% (table S1).

To validate the functionality of the PhoQ variants identified, we isolated and tested 20 individual mutants. Each mutant enabled PhoP-dependent gene expression at approximately wild-type levels in the presence of low Mg^{2+} and each suppressed PhoP activity in high Mg^{2+},

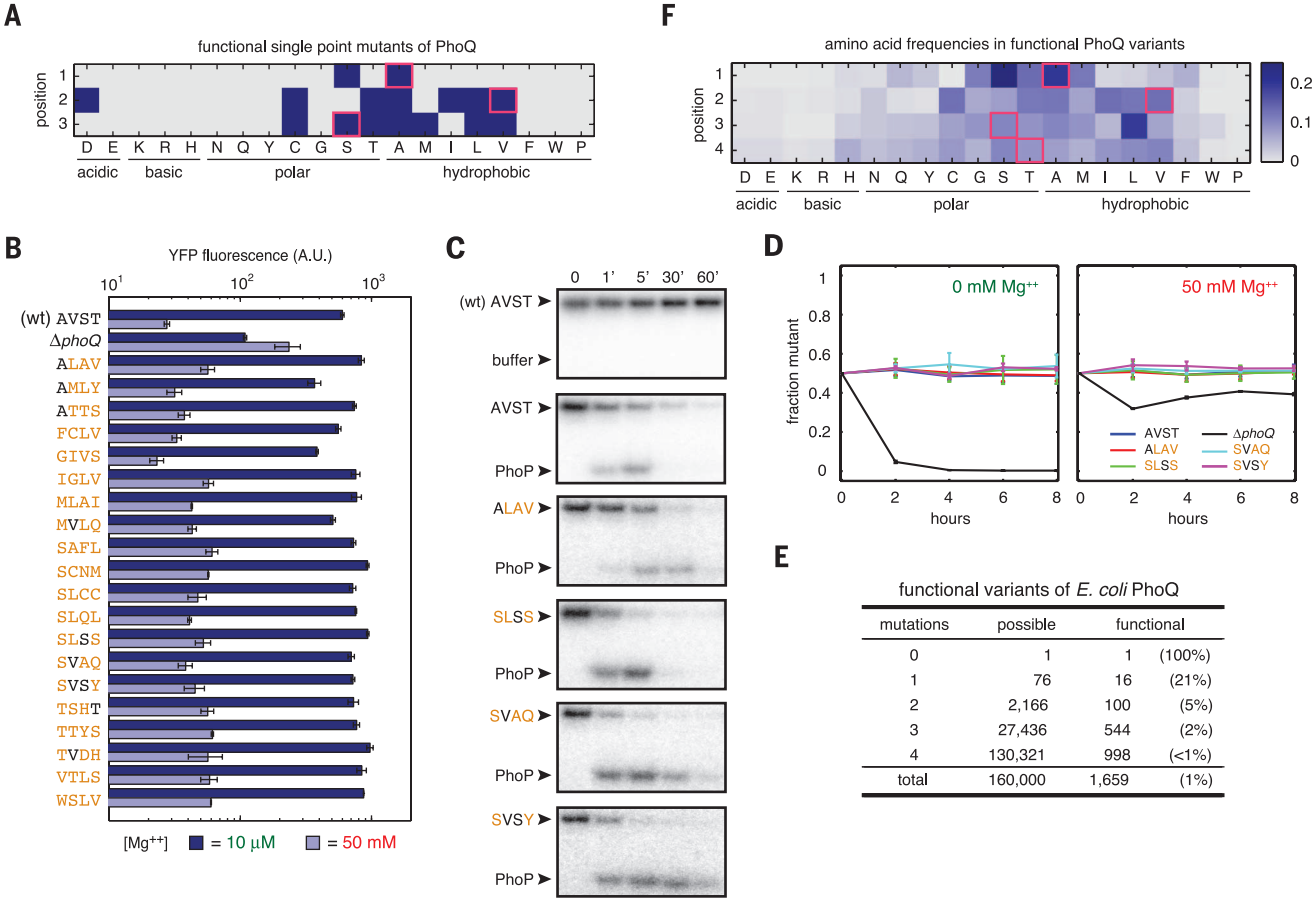

Fig. 2. Functional degeneracy of PhoQ interfacial residues. (**A**) Functionality of point mutations assessed individually. Blue indicates functional variants; magenta boxes indicate wild-type residues. Amino acid abbreviations: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr. (**B**) Flow cytometry measurements of YFP levels for 20 PhoQ variants. Error bars indicate SD; *n* = 2. (**C**) PhoQ variants indicated were autophosphorylated in vitro and tested for phosphotransfer to and dephosphorylation of PhoP. (**D**) Head-to-head competitions of wild-type against strains expressing the indicated PhoQ variant. (**E**) Summary of functional PhoQ variants. (**F**) Heat map indicating amino acid frequencies in the 1659 functional PhoQ variants.

indicating that these PhoQ variants harbored kinase and phosphatase activity (Fig. 2B). We purified four of these variants, harboring the residues ALAV, SLSS, SVAQ, and SVSY, and confirmed that each exhibited kinase and phosphatase activity in vitro and in vivo (Fig. 2, C and D, and fig. S2D).

The identification of 1659 functional PhoQ variants that are signal-responsive in vivo and that survived magnesium starvation indicates extensive degeneracy of the PhoQ-PhoP interfacial residues. In addition to 16 single mutants, there were 100 double, 544 triple, and 998 quadruple mutants; this finding demonstrates that many diverse combinations of residues can support a functional interaction with PhoP (Fig. 2E and table S1). The set of 1659 functional variants showed an enrichment of hydrophobic and small polar residues at each position (Fig. 2F). Most bulky and charged residues appeared at low frequencies, which indicates that they can only be tolerated in certain contexts (Fig. 3, A to C). For example, the substitution A284H (variant HVST) abolished the phosphatase activity of PhoQ. However, the mutant HSLV was functional, indicating that A284H can be tolerated in this context. Similarly, the substitution S288Q alone (AVQT) was nonfunctional but permissible in some variants harboring the A284S substitution, such as SIQC.

Conversely, some substitutions were permissible individually but not in combination. For instance, A284S, V285T, S288L, and T289V individually support a functional PhoQ-PhoP interface but are nonfunctional when combined (STLV) (Fig. 3D). Similarly, the combinations ACLV, TISV, and SILS, each involving residues found individually at high frequency, were severely impaired in competition against wild-type PhoQ (Fig. 3D and fig. S7A). Thus, the effects of individual substitutions are often context-dependent, or epistatic (*15*–*19*).

This epistasis implies that the functionality of variants with multiple substitutions cannot be easily predicted from the behavior of single point mutants or site-saturation mutagenesis. If each position contributed independently, our single-mutant data (Fig. 2A) would predict $2 \times 7 \times 7 \times 5 = 490$ functional combinations: [AS][ACDILTV][ACLMSTV][RTVWY]. However, our screen recovered only 104 of these and revealed an additional 1555 functional combinations (Fig. 3E), emphasizing the interdependency of individual positions.

To further assess interdependencies in PhoQ, we measured mutual information between each pair of positions in the set of 1659 functional variants (Fig. 3F and fig. S7B). The strongest coupling occurs between positions 1 and 2 and positions 1 and 3. For instance, a histidine at position 1 in functional PhoQ variants is highly correlated with a leucine or methionine at position 3 (Fig. 3, F and G), occurring three times as often as expected if these substitutions occurred independently. Similarly, a glutamine at position 3 correlates with a serine or threonine at position 1 (Fig. 3, F and G). These dependencies likely arise from constraints on the packing of adjacent residues. Positions 1 and 3 are separated by three residues in the primary sequence but are adjacent on α helix 1 in PhoQ.

The epistasis observed significantly constrains the mutational paths that PhoQ can follow through sequence space, assuming that PhoQ must retain a productive interaction with PhoP. For instance, of the 100 functional double mutants of PhoQ, only 23 represent cases where both single mutants are functional (Fig. 4A and fig. S8A). In 46 cases, only one of the single mutants is functional; thus, the mutational paths to these double mutants

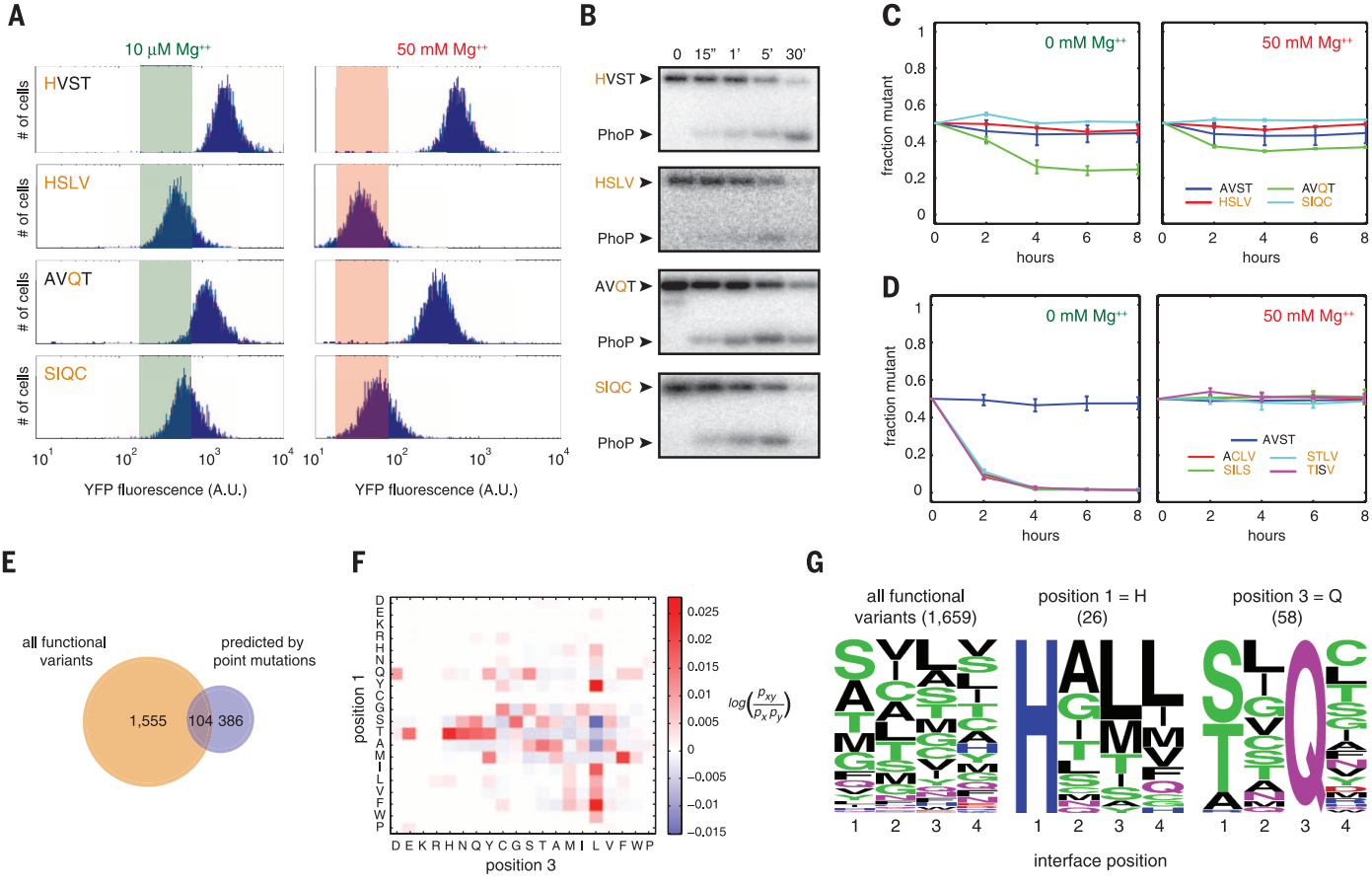

Fig. 3. Epistasis of PhoQ interfacial residues. (**A** and **B**) Flow cytometry (A) and in vitro analysis (B) of the PhoQ variants indicated. Shaded regions indicate wild-type YFP levels. (**C** and **D**) Head-to-head competitions of the wild type against strains producing the indicated PhoQ variant. (**E**) Venn diagram comparing the number of functional PhoQ variants identified with that predicted from single mutants, assuming position independence. (**F**) Heat map showing frequency of residue pairs at positions 1 and 3 relative to frequency expected if residues occurred independently. (**G**) Frequency logos for residues at each position in the PhoQ sets indicated. The height of each letter is proportional to its frequency in each set.

are constrained, requiring a certain order of substitutions. In the remaining 31 cases, the double mutant is functional even though neither single mutant is functional. Paths connecting the wild-type combination AVST to such double mutants, if they exist, must be indirect.

To systematically explore the impact of epistasis on mutational paths, we quantified the shortest path connecting the wild-type combination AVST with each of the 1658 functional variants (Fig. 4B). There are 79 PhoQ variants that cannot be reached from AVST without passing through a nonfunctional intermediate. For 428 variants, the Hamming distance from AVST equals the shortest path length. For example, AVST can convert to MLAI using four consecutive substitutions, with each intermediate being functional. However, because of epistasis, only 3 of 24 possible direct paths are permissible (Fig. 4C). Of the 1658 functional PhoQ variants, 1151 (~70%) would require more mutational steps than their Hamming distance from AVST (Fig. 4B), indicating that many paths in sequence space are indirect and require the transient introduction of other residues (Fig. 4, B and D). Similar results were obtained when considering all possible starting points, not just AVST (fig. S8B).

To visualize connectivity in sequence space, we generated a force-directed graph in which nodes represent functional PhoQ variants and edges connect nodes differing by one residue (Fig. 4E). This graph revealed five primary clusters, each with high internal connectivity. Functional variants have an average of seven neighbors (Hamming distance 1), ranging from 0 to 20 (fig. S8). The wild-type combination AVST has 16 neighbors and resides within a densely interconnected region of sequence space. Epistasis helps drive the separation of clusters; for example, clusters 1 and 2, featuring an A or S at position 1, are distal to cluster 5, which contains an L at position 3 that is incompatible with A/S at position 1 (Figs. 3F and 4E).

We also generated a force-directed graph in which edges connect functional variants separated by a single nucleotide substitution, following the codon table (Fig. 4F and fig. S9, A and B). This resulted in 260 variants that could not be reached from AVST without passing through

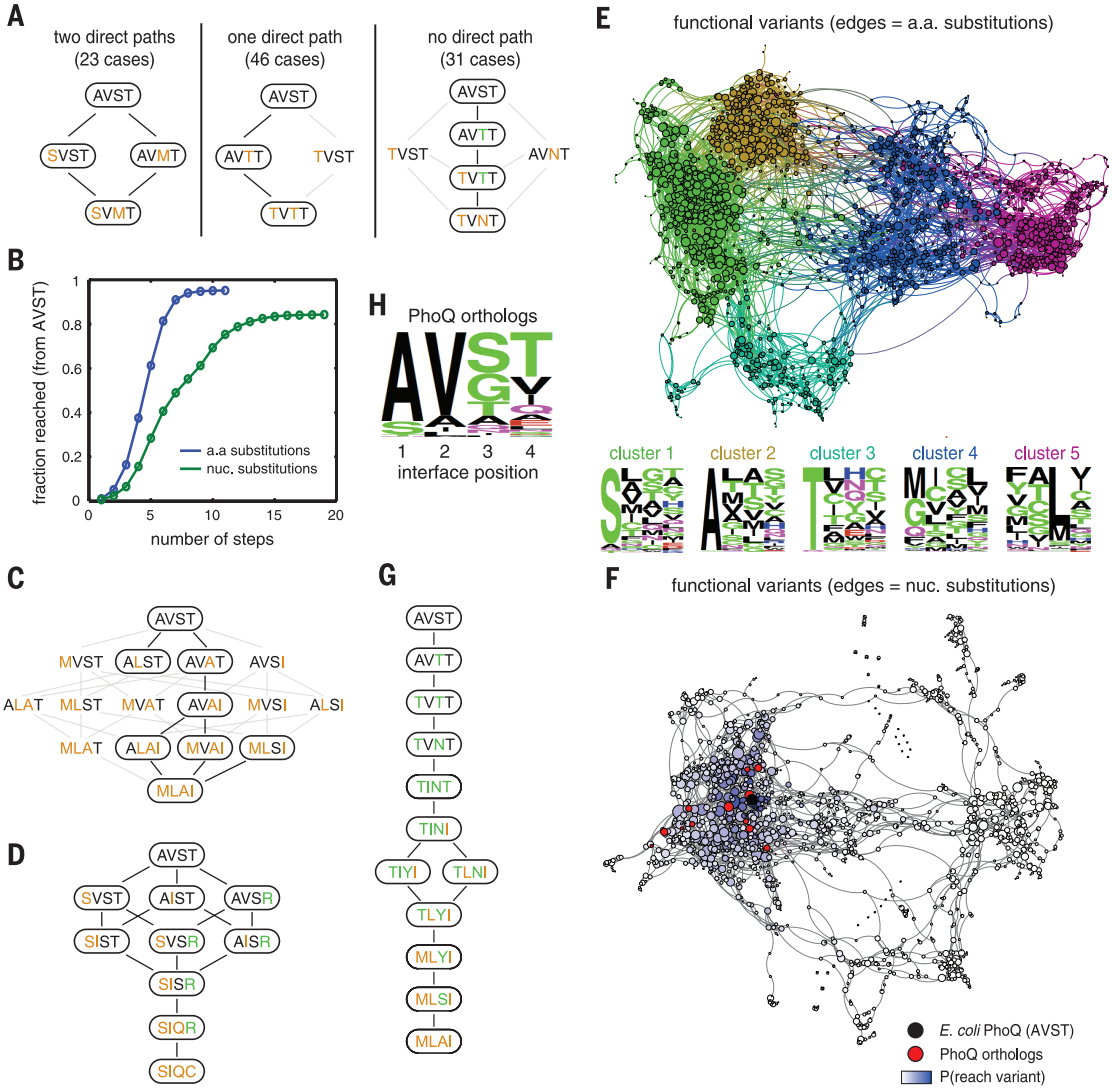

Fig. 4. PhoQ sequence space. (**A**) Tabulation and examples of double mutants reached by 2, 1, or 0 direct paths from AVST. Functional variants are circled. Lines connect variants differing by one residue (black, accessible paths; gray, inaccessible paths). (**B**) Cumulative fraction of functional variants reached from wild-type PhoQ in a given number of amino acid (blue) or nucleotide (green) substitutions. (**C** and **D**) Examples of shortest paths connecting AVST (wild type) to MLAI and SIQC. Green text indicates residues not in either terminal node. (**E** and **F**) Force-directed graphs of functional PhoQ variants (nodes) with edges connecting variants differing by one residue (E) or one nucleotide (F). Node size is proportional to number of neighbors. In (E), clusters are colored with corresponding frequency logos shown. In (F), the color scale represents the probability of reaching a node after 20 mutational steps, with red nodes indicating variants found in PhoQ orthologs. (**G**) Shortest paths connecting AVST and MLAI via nucleotide substitutions. (**H**) Frequency logo for interfacial residues of PhoQ orthologs from γ-proteobacteria.

nonfunctional intermediates, as well as a substantial increase in the length of shortest paths connecting functional variants (Fig. 4B and fig. S8). For example, the shortest path from AVST to MLAI increased from 4 to 10 (Fig. 4G). Shortest path lengths now exceeded Hamming distances for >97% of all connected variant pairs (fig. S8B). Together, the genetic code and epistasis severely constrain mutational paths in sequence space for PhoQ.

The set of functional variants includes 13 residue combinations found in PhoQ orthologs. Some residue combinations found in PhoQ orthologs are not included, possibly because these orthologs have widely divergent PhoP partners (fig. S9C) and are thus constrained differently from *E. coli* PhoQ. In general, the natural diversity in PhoQ orthologs (Fig. 4H and fig. S9C), even those with divergent PhoP partners, is much more limited than the diversity in our selected, functional variants. This difference may indicate that some PhoQ variants identified as functional have subtle defects that confer a disadvantage in the wild on long time scales. However, there was no obvious correlation between the enrichment ratios of variants after magnesium starvation and their sequences (figs. S5 and S10) (*20*). Alternatively, mutational paths may be fundamentally constrained by the nonuniform interconnectivity of variants in sequence space, such that nature has not sampled certain sequences. To test this idea, we simulated PhoQ mutational paths starting from AVST and making one nucleotide change at each step. Even after 20 simulated steps, a relatively limited region of sequence space was explored (Fig. 4F and fig. S9D), with the region most densely sampled including all of the 13 PhoQ ortholog residue combinations. Collectively, our results suggest greater functional degeneracy for PhoQ than would be expected by site-saturation mutagenesis. However, the interconnectivity of functional variants, which results from epistasis and the structure of the genetic code, has likely limited nature's exploration of sequence space (*2*, *15*–*18*, *21*), as reflected in the limited diversity of PhoQ orthologs (Fig. 4H).

REFERENCES AND NOTES

1. M. J. Harms, J. W. Thornton, *Nat. Rev. Genet.* **14**, 559–571 (2013).
2. M. A. DePristo, D. M. Weinreich, D. L. Hartl, *Nat. Rev. Genet.* **6**, 678–687 (2005).
3. J. Maynard Smith, *Nature* **225**, 563–564 (1970).
4. D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, *Science* **312**, 111–114 (2006).
5. D. M. Fowler *et al.*, *Nat. Methods* **7**, 741–746 (2010).
6. T. A. Whitehead *et al.*, *Nat. Biotechnol.* **30**, 543–548 (2012).
7. R. N. McLaughlin Jr., F. J. Poelwijk, A. Raman, W. S. Gosal, R. Ranganathan, *Nature* **491**, 138–142 (2012).
8. R. T. Hietpas, J. D. Jensen, D. N. Bolon, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7896–7901 (2011).
9. D. M. Fowler, S. Fields, *Nat. Methods* **11**, 801–807 (2014).
10. D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, S. Fields, *RNA* **19**, 1537–1551 (2013).
11. E. A. Groisman, *J. Bacteriol.* **183**, 1835–1842 (2001).
12. J. M. Skerker *et al.*, *Cell* **133**, 1043–1054 (2008).
13. P. Casino, V. Rubio, A. Marina, *Cell* **139**, 325–336 (2009).
14. E. J. Capra *et al.*, *PLOS Genet.* **6**, e1001220 (2010).
15. M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, F. A. Kondrashov, *Nature* **490**, 535–538 (2012).
16. B. Lehner, *Trends Genet.* **27**, 323–331 (2011).
17. E. A. Ortlund, J. T. Bridgham, M. R. Redinbo, J. W. Thornton, *Science* **317**, 1544–1548 (2007).
18. D. M. Weinreich, R. A. Watson, L. Chao, *Evolution* **59**, 1165–1174 (2005).
19. D. M. Weinreich, Y. Lan, C. S. Wylie, R. B. Heckendorn, *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
20. C. Bank, R. T. Hietpas, A. Wong, D. N. Bolon, J. D. Jensen, *Genetics* **196**, 841–852 (2014).
21. I. S. Povolotskaya, F. A. Kondrashov, *Nature* **465**, 922–926 (2010).

ACKNOWLEDGMENTS

We thank B. Fiske, A. Murray, B. Sauer, and the Laub laboratory for discussions. Supported by Human Frontier Science Program and Office of Naval Research grants (M.T.L.) and an NSF Graduate Fellowship (A.I.P.). Author contributions: A.I.P. and M.T.L. conceived the approach, analyzed data, and wrote the paper. A.I.P. performed all experiments.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/347/6222/673/suppl/DC1
Materials and Methods
Tables S1 and S2
Figs. S1 to S10
References (*22*–*32*)

12 June 2014; accepted 8 December 2014
10.1126/science.1257360

Science

Pervasive degeneracy and epistasis in a protein-protein interface

Anna I. Podgornaia and Michael T. Laub

Science **347** (6222), 673-677.
DOI: 10.1126/science.1257360

Exploring the limits of protein sequence space

Exploring the variability of individual functional proteins is complicated by the vast number of combinations of possible amino acid sequences. Podgornaia and Laub take on this challenge by analyzing four amino acids critical for the interaction between two signaling proteins in *Escherichia coli.* They build all the possible 160,000 variants of one of the two proteins and find that over 1650 are functional. Even though there can be very high variability in the composition of the interface between the two proteins, there are nonetheless strong context-dependent constraints for some amino acids, which suggests why many functional variants are not seen in nature.

Science, this issue p. 673