Engineering orthogonal signaling pathways to probe sequence space capacity

By

Conor James McClune

B.A. Molecular and Cell Biology
University of California, Berkeley (2012)

Submitted to the Department of Biology in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN BIOLOGY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2019

Conor James McClune Author August 7th, 2019	Michael T. Laub Professor of Biology Thesis Advisor
Christopher A. Voigt Professor of Biological Engineering Thesis Advisor	Stephen P. Bell Professor of Biology and Biological Engineering Co-Chair, Biology Graduate Committee

Engineering orthogonal signaling pathways to probe sequence space capacity

by

Conor McClune

Submitted to the Graduate Program in Biology
on September 3rd, 2019 in partial fulfillment of the requirement for the degree of
Doctor of Philosophy in Biology at the Massachusetts Institute of Technology

ABSTRACT

Gene duplication is a common and powerful mechanism by which cells create new signaling pathways, but recently duplicated proteins typically must become insulated from each other, and from other paralogs, to prevent unwanted cross-talk. A similar challenge arises when new sensors or synthetic signaling pathways are engineered within cells or transferred between genomes. How easily new pathways can be introduced into cells depends on the density and distribution of paralogous pathways in the sequence space defined by their specificity-determining residues. Here, I directly probe how crowded sequence space is by generating novel two-component signaling proteins in *Escherichia coli* using cell sorting coupled to deep-sequencing to analyze large libraries designed based on coevolution patterns. I produce 58 new insulated pathways, in which functional kinase-substrate pairs have different specificities than the parent proteins, and demonstrate that several new pairs are orthogonal to all 27 paralogous pathways in *E. coli*. Additionally, I readily identify sets of 6 novel kinase-substrate pairs that are mutually orthogonal to each other, significantly increasing the two-component signaling capacity of *E. coli*. These results indicate that sequence space is not densely occupied. The relative sparsity of paralogs in sequence space suggests that new, insulated pathways can easily arise during evolution or be designed de novo. I demonstrate the latter by engineering a new signaling pathway in *E. coli* that responds to a plant cytokinin without cross-talk to extant pathways. The work in this thesis also demonstrates how coevolution-guided mutagenesis and sequence-space mapping can be used to design large sets of orthogonal protein-protein interactions.

Thesis Co-supervisor: Michael T. Laub
Title: Professor of Biology

Thesis Co-supervisor: Christopher A. Voigt
Title: Professor of Biological Engineering

Table of Contents

Figure Index

Acknowledgements

First and foremost, I would like to thank my two advisors, Mike Laub and Chris Voigt. The independence you gave me to work at the interface of your interests and to immerse myself in two labs at the forefront of disparate fields, was not only an incredible education and a true joy.

I would also like to thank my committee: Amy Keating and Jing-Ke Weng. I appreciate the sincerity with which you gave your time and attention.

To my labmates in the Laub and Voigt labs: you are too talented to be so kind and fun. Brian Caliando, Lei Yang, Mike Smanski, Mike Salazar, Anjana Badrinarayanan, Anna Podgornaia, Chris Aakre and Diane Bearonsen – you were all incredible technical mentors and taught me so much. I want to thank my baymates for your unreasonable ability to make me smile. Katie, you were always there to hop fences or commiserate about cell sorting. David, just do it. Thuy Lan, thank you for Frankie. Frankie, thank you for everything. Peter, you were always there to keep me in shape, both on the Charles and on the server. Aurora, everyone was jealous that I got the good UROP.

I want to thank my colleagues in the Voigt lab for making a huge lab feel like a small family. Jenn Brophy, thanks for testing out labs before I join then. Daniel Anderson, I am never getting on a tandem bike with you again. Working with team DERPA has been an incredibly fun and enriching experience – thank you Andrew, Thomas, Emerson and Adam. Especially Emerson, who taught me the value of wearing real shoes on a construction site. Brian Bonk and Felix Moser, I cannot believe we won so many biotech pitches, but it was an incredibly rewarding experience and I never would have done it without you. Alec, you taught me the proper way to do biophysics. Barbara and Terry, I don't know how you managed to be so aggressively productive in such a compassionate way.

The Biology Department is such a supportive, unpretentious community. Thank you to all my classmates, friends and neighbors for sustaining my sanity. Betsey Walsh, thank you for taking care of me whenever I wondered into your office with questions not worth your time.

To my family, thank you for your love and support from the opposite corner of the country. It never feels like we are very far away.

Thank you, Djenet, for every adventure, scientific and otherwise, during this PhD.

Chapter 1 – Introduction

Biological systems evolve through reuse and rewiring of a common set of parts

Most cellular functions consist of maintaining homeostasis and linking environmental signals to appropriate reactions. The logic that governs these behaviors is defined by the network of regulatory proteins, RNAs, and other molecules, large and small. Connectivity between these interacting genes describes how information flows for sensors to responses and from disturbances to corrections. Thus, the evolution of new cellular behavior primarily arises from either the gain or loss of genes or of the interactions between the proteins they encode.

Most new genes derive from copies of other genes (Long et al., 2013). Gene duplication, chromosome duplication and horizontal gene transfer has produced genomes that contain numerous copies of homologous proteins, various proteins composed of structurally similar domains, and even diverse domains composed of shared peptide “themes” (Nepomnyachiy et al., 2017). As our ability to sequence genomes and metagenomes has accelerated, we have learned the extent to which evolution has repurposed similar protein building blocks. Though the rate of gene discovery has increased, the discovery of new protein folds has declined (Khafizov et al., 2014). This saturation highlights how much of molecular biology’s diversity comes from variants on similar structural themes, rather than entirely novel molecular components. The total number of extant protein folds is estimated to be in the thousands (Liu et al., 2004), though UniProt TrEMBL currently (June 2019) contains greater than 150 million protein sequences (UniProt, 2015). Furthermore, a small subset of just 20 highly common protein ‘superfolds’ describes the structure of 46% of all protein domains (Cuff et al., 2009). Such widespread structural conservation supports the intuitive theory that it is easier to evolve new proteins from sequences that already fold into stable geometries, as opposed to random sequences.

While new nodes in a cell's regulatory network generally arise through gene duplication, new connections form as evolutionary processes rewire protein interaction specificity. Gene birth and rewiring are intrinsically linked. After a gene duplicates, one of the first functionally significant changes that can occur to differentiate these identical duplicates is an alteration of how they interact with other cellular components. Duplicated paralogs evolve different binding specificities for the proteins (Schreiber and Keating, 2011), RNAs (Hogan et al., 2015) or DNA sequences (Pougach et al., 2014) with which they interact. Enzyme paralogs evolve different substrate specificities or diverge in their enzymatic products (O'Maille et al., 2008) (Sunden et al., 2015). In this chapter, I will address, in mechanistic detail, how interaction specificity can be rewired, with an emphasis on protein-protein interactions.

I will also address how repeated use of structural motifs affects how biological systems can evolve and ultimately how they can be engineered. Once paralogs have adopted different roles, the maintenance of insulation through separate specificities can be crucial for each protein's independent function. This is especially critical for proteins involved in signaling pathways – where the set of interaction partners defines the core function of a protein. For example, mutations that change the specificity of human kinases can lead to cancer (Creixell et al., 2015b) and mutations that relax the specificity of bacterial kinases can dramatically reduce fitness (Capra et al., 2012). The mechanisms that determine specificity for a given protein-protein complex also determine the total variations of that complex with orthogonal specificity, thereby limiting the repeated use of those structural domains.

Evolutionary plasticity of protein interactions

Regulatory flexibility arising through rewiring of protein-DNA interactions

The physical interaction between transcription factors and the regulatory DNA of the genes they regulate is the first layer of transcriptional circuits and the location of agile evolutionary rewiring. Because it is tractable to identify DNA sequence motifs and to rapidly characterize the optimal motif for a DNA-binding protein, we have gained extensive knowledge about the mechanisms of these evolutionary dynamics. More recently, this has been aided by tools like ChIP-seq, which captures information about a protein's global occupancy across the genome (Johnson et al., 2007).

Some regulatory networks have evolved through the emergence and disappearance of a transcription factor's binding sites (Nocedal et al., 2017). Others emerge as the specificity of a regulatory protein drifts (Nocedal and Johnson, 2015). Matα1, a transcription factor that controls the well-studied yeast mating pathway, has undergone a dramatic drift in DNA recognition specificity (Baker et al., 2011). Johnson and colleagues have demonstrated that paralogs can differentiate themselves by developing different recognition sequences, different half-site spacing, or different interactions with cofactors (Perez et al., 2014). Many of these transcriptional rewiring events are coupled to the emergence of new transcription factor paralogs. Pougach and Verstrepen identified two recently duplicated paralogs of the transcription factor MalS that have differentiated through alternative mechanisms of DNA specificity. One drifted toward a novel recognition sequence; the other lost DNA binding affinity and thus relied on repetitive recognition sequences to recruit TFs cooperatively (Pougach et al., 2014). When duplicated transcription factors homodimerize or form activation complexes with cofactor proteins, they can quickly subfunctionalize if each sister paralog loses different DNA or protein interactions (Baker et al., 2013). In *Kluyveromyces lactis*, Mcm1 regulates transcription of both arginine biosynthesis enzymes and mating genes. In the lineage that includes *S. cerevisiae*, however, Mcm1 duplicated, then developed two causal mutations breaking different contacts and allowing the daughter paralogs to independently regulate the arginine biosynthesis and mating regulons (Baker et al., 2013). Mutations that cause specificity drift of transcription factors allow organisms to sample different regulatory connectivity, and gene duplication is a powerful mechanism for increasing the complexity of these networks.

The scale and plasticity of protein-protein interaction networks

Much of the connectivity of regulatory networks is defined by the mutual affinity of protein domains and peptides that drive the interactions between components of each pathway. The proteins of the *S. cerevisiae* genome weave a network of approximately ~18,000 ± 4,500 binary protein-protein interactions (Yu et al., 2008). Multicellular organisms have an order of magnitude more interactions per genome: human proteins make ~130,000 binary interactions (Venkatesan et al., 2009) and *Arabidopsis thaliana* proteins make ~299,000 interactions (Arabidopsis Interactome Mapping, 2011). The connectivity of these networks follows a power law; rare hub-proteins are connected to dramatically more partners than average proteins interactions (Arabidopsis Interactome Mapping, 2011; Venkatesan et al., 2009; Yu et al., 2008). By looking at the connectivity of paralogous proteins within these networks, we can get a sense for how increasing complexity evolved. In general, sister paralogs diverge rapidly after duplication, both in terms of interaction partners and amino acid sequence, and then stabilize (Arabidopsis Interactome Mapping, 2011). The correlation of these evolutionary dynamics suggests that interaction rewiring is one of the main drivers of sequence change after duplication.

Certain domains highly prevalent within signaling architectures, such as leucine zippers, DED, SAM, SH3, PDZ and WW domains, are responsible for producing many of the specific interactions that associate each subsequent protein in signaling cascades (Ingham et al., 2005; Pawson and Nash, 2003). Interaction networks can be rewired by swapping or fusing these scaffolding domains (Howard et al., 2003). Paralogous binding domains also undergo extensive reshuffling of binding specificity. The leucine zippers of bZIP proteins, which are helical domains that can form both homodimers and heterodimers, have dramatically changed their interaction network during metazoan evolution (Reinke et al., 2013). Systematic *in vitro* binding assays show that amongst the 14 bZIP families common in metazoans, nematode and Drosophila bZIPs have lost many interactions present in the ancestral network, while humans and anemones have gained extensive new connections (Reinke et al., 2013). Many of these specificity changes could be explained and reversed by a single amino acid substitution. Rewiring of the small, modular domains that enforce pathway connectivity plays a role in the evolution of new biological programs.

Molecular mechanisms of rewiring protein interactions

Cells are extremely crowded environments where proteins can comprise up to 30% of dry weight (Elcock, 2010; Zimmerman and Trach, 1991). Every protein must correctly discriminate their various binding partners from this background noise. The binding affinities relevant in protein-protein interactions range from femtomolar to millimolar and depend heavily on their physiological concentration, solute conditions, and physical access to one another (Schreiber and Keating, 2011). For some interactions, the specificity between two proteins is driven by their co-localization, compartmentalization, and coexpression (Schreiber and Keating, 2011). However, specificity is often encoded in physiochemical complementarity at the structural level via the residues that make direct contact at the protein-protein interface. While expression and colocalization offer dynamic regulatory options, the dimensions of temporal dynamics and cellular compartments are dramatically smaller than the combinatorial dimensionality of amino acid sequences.

While protein-protein interactions generally span 1000-2000 $Å^2$ and involve dozens of interacting residues (Janin et al., 2007), the energetic contribution of these residues is not homogeneous. A smaller subset of residues, often termed “hot-spots” and generally defined as contributing > 2 kcal/mol in binding $\Delta\Delta G$, are invaluable for the binding energy and the specificity of protein interactions (Bogan and Thorn, 1998). The “hot spot” hypothesis was formalized after an alanine-scan of the growth hormone-receptor complex (hGH-hGHbp) revealed that two tryptophans, out of the 30 residues that make interfacial contacts, contribute over 75% of the binding energy (Clackson and Wells, 1995). Since the hGHbp structure, hot spots have been identified within many dimer structures, often clustered into small regions of the interface (Keskin et al., 2005; Moreira et al., 2007).

This nonuniform contribution of surface residues to binding specificity is consistent with the observation that there are myriad strategies for constructing a tight interface between two given protein domains. An analysis of 433 binary protein complexes from the AB/AC database (non-homologous proteins B and C each interacting with homologous proteins A and A’) showed no evidence of convergent evolution in binding strategies (Martin, 2010). Non-homologous proteins used different contact points and different residues, few of which were evolutionarily conserved, to build an interface to the same domain. This suggests an intuitive hypothesis: the solution space for strong binding interfaces, built combinatorically from the twenty natural amino acids, is large.

The skew in the significance of binding residues also reduces the number of amino acid substitutions needed for proteins to gain, lose or alter binding capacity. The Dscam proteins from Drosophila provide an extreme example of the binding plasticity endowed by a few amino acids (Fig. 1.1) (Wojtowicz et al., 2004; Wojtowicz et al., 2007). These surface-displayed proteins, which homodimerize and allow insect neurons to self-discriminate, are comprised of three alternatively spliced immunoglobulin domains. The 12, 48 and 33 alternative exons for the Ig2, Ig3 and Ig7 sections, respectively, encode domains with remarkable homodimerization specificity, despite some of these domains differing by only a dozen residues (Wojtowicz et al., 2007). The result is more than 19,008 possible Dscam splice variants with unique self-recognition capacity. Even more remarkably, the specificity of these three domain classes is encoded entirely in small stretches of residues; transplanting a region of 10 amino acids can endow any Ig2 variant with the specificity of any other Ig2 variant (Wojtowicz et al., 2007). Even fewer substitutions are needed to rewire the other two domains. Dscam Ig domains illustrate how minimal substitutions can block previous binding partners and recruit new ones, independent of the domain backbone.

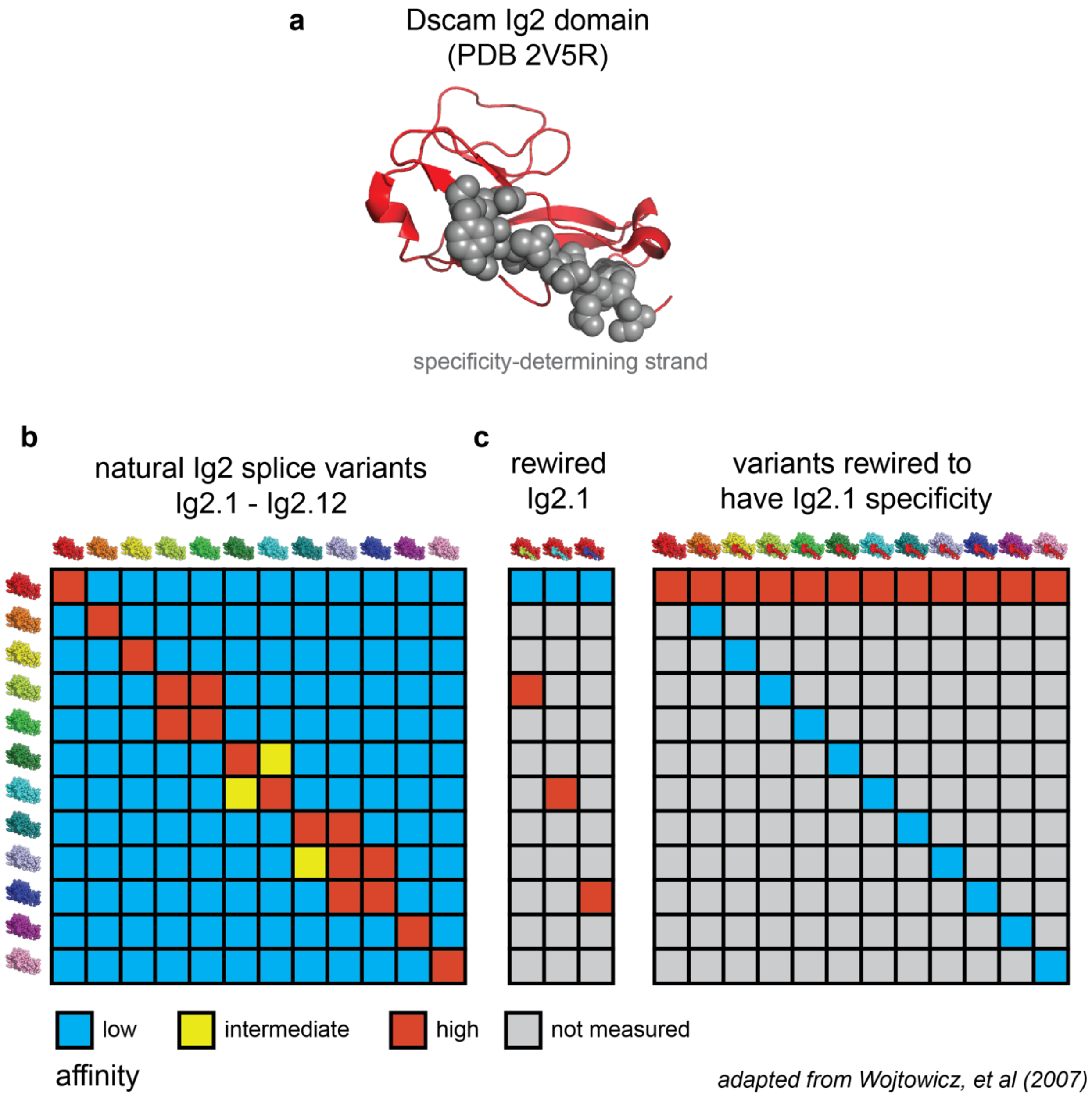


Figure 1.1 The modular specificity-determining strand of Dscam domains.

(a) A single strand (shown in spheres) of a Dscam Ig2 domains is necessary and sufficient to encode homodimerization specificity. (b) Dcscam Ig2 splice variants, the least numerous of the variable Dscam regions, display stringent homodimerization specificity. (c) Swapping the specificity strand is sufficient to rewire Ig2.1 to have the specificity of Ig2.4, Ig2.7, or Ig2.10. It is also sufficient to endow all twelve Ig2 variants with the specificity of Ig2.1.

There are numerous other homologous complexes with structurally conserved domains that have evolved highly divergent interfaces. Even when many hotspot positions are conserved, the identity of the residues at these positions diverge to the extent that structurally identical homologs are incompatible for binding. This has been observed in the beta-lactamases (TEM1, SHV, KPC2) and their binding proteins (BLIP, BLIPII, BLIP2). Despite these proteins forming structurally conserved complexes with essentially identical orientation, only two residue contacts are conserved (Gretes et al., 2009; Schreiber and Keating, 2011; Zhang and Palzkill, 2004). For BLIP, which effectively inhibits both TEM1 and KPC2 through nanomolar binding, selectivity can be introduced via two point-mutations that preserve nanomolar KPC2 inhibition while reducing TEM-1 binding by 20,000-fold (Chow et al., 2016). Similar observations have been made for antibody-antigen contacts, BCL2 proteins (Dutta et al., 2010), and colicin-immunity protein complexes (Meenan et al., 2010). A short directed-evolution experiment with the non-cognate DNase-immunity protein pair ColE7–Im9 demonstrated that complexes with similar binding modes can quickly alter their specificity to interact with previously incompatible partners (Levin et al., 2009). Other work has demonstrated that the specificity profiles of PDZ domains could be dramatically altered by single amino acid substitutions (Gee et al., 2000). Overall, rewiring specificity is rampant in evolutionary history and requires few amino acid changes.

The discovery of hotspots has also manifested in numerous successful efforts to rationally rewire the specificity of binding interactions. Kortemme and Baker conducted one of the first computationally-guided redesigns of protein-interaction specificity by first introducing a single, polarity-switching N516L mutation in a DNase (colicin E7), then generating compensatory mutations in its partner protein (Im7). This “second site suppressor” strategy, though not always trivial (Sammond et al., 2010), has now been applied many times to build protein complexes with orthogonal specificity (Netzer et al., 2018), including therapeutically relevant protein complexes such as PDL1/PD1 (Shrestha et al., 2019) and IL2 / IL-2Rβ (Sockolosky et al., 2018).

The dominant influence of a few specificity-determining residues has been highlighted by experimental efforts to transplant the minimal specificity-determining regions from one protein complex to another. Since Peter Jones and Greg Winter discovered in 1986 that the complementarity-determining regions (CDRs) of mouse antibodies could be transplanted into human antibodies to convey the same specificity (Jones et al., 1986), scientists have explored residue transplantation to test (*a*) how consistently interface residues behave when fixed to different backbones, (*b*) what are the minimal amino acid determinants of specificity and (*c*) how modular are the different sections of an interface from each other and proximal residues. Melero and Kortemme explicitly tested transferability by measuring how well a set of specificity switching mutations designed for one PDZ domain-peptide complex could be introduced five other PDZ-peptide complexes (Melero et al., 2014). In three out of five complexes, the transplanted residues to generated a functional PDZ-peptide pair with specificity orthogonal to the parent pair. However, some epistasis limited transfer to the other two transplanted pairs, which did not achieve insulation from both parent proteins. In a more dramatic example of interface transplant, nine specificity were transferred from one kinase-regulator complex, *Escherichia coli* PhoRB , into a kinase-regulator complex, HK853-RR468 from the hyperthermophile *Thermatoga maritima* (Podgornaia et al., 2013). Although the HK853-RR468 complex comes from an organism that lives in an environment 50°C warmer than *Escherichia coli*, the mutant complex HK853*-RR468* was just as functional as the wild-type complex and crystalized with approximately the same quaternary structure. Rewired interface chimeras have also been built from another large protein family: the ParDE toxin-antitoxin systems (Aakre et al., 2015). Four specificity residues were sufficient to transplant the specificity of antitoxins ParD1 or ParD2 into ParD3, such that the mutants specifically inhibited cognate toxins ParE1 or ParE2, instead of ParE3. These successful transplants highlight the apparent modularity of the specificity-determining residues in many protein complexes. They confirm that a small number of residues contribute most of the specificity of a protein interaction and do not have significant epistatic dependencies with the residues that make up the rest of the protein domain. They do not, however, demonstrate, that interface residues are generally independent from each other. Indeed, systematic mutational scanning of alkaline phosphatases (Sunden et al., 2015) has shown that specificity-determining regions make up multiple small networks that are modular from each other, yet contain a significant amount of within-network epistasis. Adjacent clustering of hotspot residues appears to be a general trend across protein-protein interactions (Keskin et al., 2005). While the network connectivity of residue epistasis is likely to vary for each system, modularity, at some scale, appears to be the trend.

Predicting specificity determinants using extant protein sequences

The growing realization that the specificity of many protein interactions could be reduced, primarily, to a small set of residues inspired the development of many computational methods aimed at identifying such residues. The most intuitive and oldest of these efforts are purely physical: each atom of the protein(s) of interest and solvent are explicitly modeled, forces are calculated and atomic movement can be simulated over short periods of time. These kinds of molecular dynamics analyses are, however, very computationally expensive, and generally only used when researchers hypothesize there is a dynamical aspect to specificity, as in TCR-peptide-MHC catch bonds (Sibener et al., 2018). Rigid body docking has been used for designing protein interaction specificity, but its accuracy is generally too poor to produce single, high-confidence predictions or designs. possibly due to the numerous degrees of freedom in protein confirmation (Kastritis and Bonvin, 2010). Instead, it is often coupled to the construction of large libraries of hypothetical designs (Boyken et al., 2016; Chevalier et al., 2017).

The past decade has seen a dramatic increase in computational approaches using coevolutionary information to extract meaningful structural and functional information about proteins-protein interactions. This approach leverages the large sets of homologous protein sequences available in ever-growing databases of genomes and metagenomes. By analyzing how homologs differ, co-evolving amino acid positions can be found: namely, pairs of positions where the amino acid identity in one position is dependent on the identity of the other. For example, if two positions always contain amino acids of opposite charge, we can infer they may interact in physical space (structural information) or be important for folding or binding (functional information). These coevolutionary relationships were first formalized by calculating the mutual information between differing columns in a multiple sequence alignment (Korber et al., 1993; Li et al., 2003; Skerker et al., 2008). To predict residue interactions across protein interfaces, the sequences of interacting proteins are concatenated into a single sequence per genome, and the same protocol can be applied. If there are many homologous complexes per genome, the members of each separate complex can be predicted, at least in bacterial genomes, by their genomic organization in separate co-operonic locations (Ovchinnikov et al., 2014).

Coevolutionary covariance analysis proved to be a revolutionary mindset for protein science in the genomic era. However, mutual information suffered from some intrinsic limitations (primarily the inability to differentiate direct and indirect dependencies between residues), leading to the development of a zoo of new coevolutionary models and metrics. These models, including Direct Coupling Analysis (Morcos et al., 2011), PSICOV (Jones et al., 2012), EVfold (Hopf et al., 2012; Marks et al., 2011) and GREMLIN (Balakrishnan et al., 2011), lean on techniques from statistical physics, machine learning and information theory. In contrast to a mutual information model, which independently measures the coupling of every pair of residues without regard to whether this coupling may be mediated be a third residue, these second-generation models fit a global Potts model of the form:

$$P(A_1, A_2, \dots A_L) = \frac{1}{Z}\exp\left[\sum_{i} h_j(A_i) + \sum_{i<j} J_{ij}(A_i, A_j)\right]$$

These models consider both the amino acid biases at each position (h_i) as well as couplings between dependent residues (J_{ij}) when calculating the probability of observing a protein sequence ($A_1 \ldots A_L$) of length L. These coupling terms (J_{ij}) provide the critical information about how residues of a protein or protein complex interact. The main technical challenge of fitting a Potts model arises from the normalizing partition function Z, which, if done precisely, would require enumeration of the model over every possible amino acid sequence of length L at a computational cost that scales intractably as 20^L. As the field has evolved, different methods have been developed to find efficient and accurate approximations of this function, with pseudolikelihood models finding widespread use in recent years (Balakrishnan et al., 2011; Kamisetty et al., 2013; Stein et al., 2015).

The increased performance of Pott's model for predicting residue couplings within proteins and across protein contacts led to the application of residue co-variation to a plethora of biological problems (Fig. 1.2) (Kamisetty et al., 2013; Marks et al., 2011; Morcos et al., 2011; Weigt et al., 2009). By connecting multiple, binary, contact maps together, larger protein complexes can be modeled entirely from sequence data (Ovchinnikov et al., 2014). Very recently, as increased protein sequence data has increased the statistical power of these models, the interconnectivity of the entire *Escherichia coli* and *Mycobacterium tuberculosis* proteomes was modeled with this strategy (Cong et al., 2019). In addition to the prediction of protein contacts and specificity determinants, the development of residue coevolution tools have been motivated by of *ab initio* structure prediction (Ovchinnikov et al., 2017). Indeed, one of the first applications of protein Pott's model tools was to model the structure of transmembrane proteins by identifying intra-protein residue contacts (Marks et al., 2011). Since then, the GREMLIN model has been used to identify 137 novel protein folds in metagenomic data (Ovchinnikov et al., 2017), and coevolution analysis has become a standard part of all successful *ab initio* structure prediction pipelines (AlQuraishi, 2019; Ovchinnikov et al., 2016; Ovchinnikov et al., 2018).

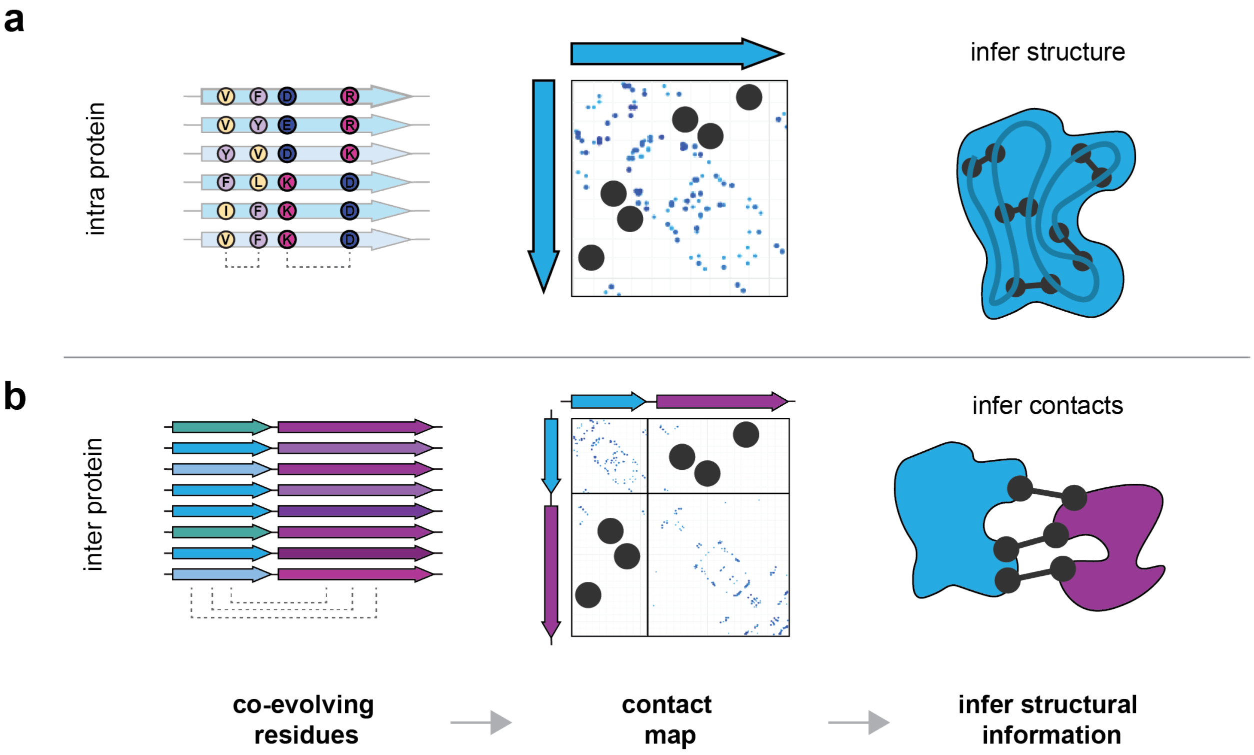


Figure 1.2 Inference from residue covariation.

(a) By tracking covariation of different positions within a protein multiple sequence alignment, direct residue dependencies can be quantified and a contact map can be inferred. This contact map can be used to constrain the problem of protein structure prediction. (b) A similar strategy can be utilized to track residue covariation between the residues of interaction proteins, as long as protein partners can be inferred directly from genomic synteny or context. Interprotein contact maps can be useful for constructing quaternary complexes and inferring specificity-determining regions.

There are certain limitations of coevolution analysis. Firstly, their statistical power scales with the number of homologous sequences available. Consequently, they are far more useful for large protein families, especially prokaryotic families where we have large amounts of metagenomic sequence information. Even though we have sequenced thousands of human genomes, the amino acid sequences are not diverse enough to provide additional information. Fortunately, this feature does mean that the accuracy of these models will improve, without any algorithmic changes, as we continue to collect more diverse genomes and metagenomes. The second limitation of coevolutionary modeling is that it generates contact maps that represent average contacts for an entire multiple sequence alignment. As such, any idiosyncratic structural or functional features that arose recently in the evolutionary history of your query protein will not be captured. There is an inherent tradeoff with these models for including more distant homologs, which increases the confidence of highly conserved contacts but reduces signal for features unique to the protein subclade of interest. One of the solutions to this problem, and a general benefit of coevolutionary models, is their compatibility with datasets from high throughput mutational assays. Recently, two groups demonstrated that deep mutational scanning of a protein could be coupled with coevolution analysis to build protein structures without either structural or genomic information (Rollins et al., 2019; Schmiedel and Lehner, 2019). The major limitation of this approach, however, is that it requires the development of selection coupled to the proper folding or function of the query protein – an integral part of deep mutational scanning.

Some of these strategies may change as more elements of deep learning approaches are incorporated into protein informatics. Already, a plethora of hybrid techniques have emerged that use traditional Potts model fitting to generate a contact map, then use deep residual networks to fine-tune these contact maps (Liu et al., 2018b; Wang et al., 2017). Deep neural network architectures are dramatically changing how bioinformaticians think about modelling complex datasets. New techniques, such as DeepSeq, rely on these new architectures to extract meaningful information from sequence data, then extract the residue coupling networks from within these models (Riesselman et al., 2018). DeepSeq exemplifies a new form of generative model that uses standard neural network architectures to fit highly complex systems through a black box of millions of parameters. The innovation then comes from interpreting the trained parameters within the model.

Coevolution inference has advanced from a niche interest to an essential component of computational protein analysis in the span of a decade. Its accuracy has improved such that it is an integral part of the best *ab initio* structure prediction algorithms, however, it still is not reliable enough to predict the effects of many novel mutations (Gray et al., 2018; Hopf et al., 2017). This may be due to the limitations outlined above, as well as these models' lack of information about residue combinations not found in nature. Time will tell if additional protein sequences, natural or experimentally generated, are sufficient to improve these methods.

There is surprisingly little use of covariation techniques for rational reengineering of protein specificity or other forms of protein engineering. Co-variation information has been successfully employed to identify broad sets of specificity-determining residues, though these must often be experimentally refined (Aakre et al., 2015; Nicoludis et al., 2015; Skerker et al., 2008). However, current engineering endeavors from coupling information are limited to transplanting sets of residues from other natural proteins, not designing novel contacts. Pott's models have been used as scoring metrics for natural (Bitbol et al., 2016; Cheng et al., 2014) or homologous protein interactions (Dimas et al., 2019), but coupling-guided specificity engineering has been limited to single point substitutions and has displayed unremarkable accuracy (Cheng et al., 2018). Despite limitations for producing single high-confidence designs, evolutionary co-variation information may be useful for dramatically reducing the search space for protein engineering, especially for protein specificity, analogous to its utility in reducing the search space of computational protein folding. In the next chapter, I demonstrate this approach.

The fitness cost of spurious interactions

Thus far, I have highlighted the plasticity of protein interaction networks. However, just as missense mutations rarely have positive consequences for protein functions, alterations in protein specificity or increased promiscuity have fitness costs due to direct alterations of biological logic. Some of the evidence for this comes from selection signatures in extant genomes. Across metazoans, there is a strong negative correlation between the number of tyrosine kinases and the frequency of tyrosine in other proteins (Tan et al., 2009), presumably because of the cost of spurious phosphorylation events. More direct evidence comes from the maintenance of exquisite insulation between extant paralogous signaling pathways. Zarrinpar and colleagues found that the SH3 binding-domain of *S. cerevisiae's* osmolarity sensor Sho1 could not be functionally exchanged with any of the genome's 26 other SH3 domains without disrupting signaling and producing strains inviable on high-osmolarity media (Zarrinpar et al., 2003). Histidine kinases purified from *Escherichia coli* and *Caulobacter crescentus* phosphorylate only a single substrate protein from all possible paralogs encoded within these genomes. Mutations to increase the promiscuity of these sensor kinases produce low-fitness strains that quickly drop out of mixed populations (Capra et al., 2012).

Similarly, calamitous consequences can be seen in humans when mutations rewire protein interactions. A recent analysis of kinase mutations in multiple ovarian cancer samples found specificity drift of multiple kinase pathways to be the causative, oncogenic perturbation (Creixell et al., 2015b) (Creixell et al., 2015a). Certain non-catalytic domains such as signal peptides, DNA-binding domains and transmembrane domains are enriched in oncogenic gene fusions, exemplifying the harmful outcomes that can result from single changes in the connectivity of regulatory networks (Latysheva and Babu, 2016). A recent analysis of The Cancer Genome Atlas found that when kinases become mutated, specificity-determining residues were more likely to be altered than catalytic, regulatory or other residues (Bradley et al., 2018). Altering signaling networks changes how cells respond to their environment – sometimes with devastating consequences.

These pathologies illuminate a particular challenge for engineering or modifying biological systems. The word "orthogonal" is rampant within synthetic biology literature, and for good reason. Engineered biological pathways must not interfere with each other or with the endogenous systems that maintain homeostasis. Spurious interactions can be difficult to predict, difficult to detect, and extremely consequential. To address this complication, synthetic biologists have successfully built replication (Ravikumar et al., 2014; Ravikumar et al., 2018a), transcriptional (Segall-Shapiro et al., 2014) and translational (An and Chin, 2009; Liu et al., 2018a) machinery that operates independently from host processes. However, there has been little work generating protein interactions that are orthogonal to host proteins, probably because of the difficulty of protein engineering and the systems scale of the challenge. The many efforts thus far to construct synthetic signaling architectures instead rely on modular chimeric proteins (Lim, 2010); by fusing either natural (Barnea et al., 2008) or engineered (Thompson et al., 2012) sets of binding domains to proteins of interest, pathways can be rewired (Howard et al., 2003) or forced to be insulated from endogenous programs (Whitaker et al., 2012). Additionally, chimeric scaffolds have been used to string together signaling proteins in novel ways and rewire signal processing (Park et al., 2003). Chimeric constructs are undoubtedly the simplest way to generate insulated protein interactions, when possible. But scaling this strategy is severely limited by two restrictions. First, our ability to engineer novel interactions is limited by the number of orthogonal interaction domains we have available – most of which are taken from extant genomes and are indeed not insulated in many organisms. Secondly, many signaling proteins are not so domain-modular that entirely new domains can be inserted without disrupting function. Therefore, methods are needed to diversify the specificity of protein-protein interactions without making crude fusions with extant domains.

The sequence space of protein-protein interactions

I have thus far addressed the minimal sequence features determining the specificity of many protein interactions, the plasticity of these interactions, the expansion of proteins into large paralogous families and the importance of avoiding cross-talk. Collectively, these observations highlight a particular challenge: is there sufficient sequence space to continually encode paralogous variants of the same proteins with different specificity? The term sequence space is used elsewhere to refer to concepts other than protein specificity, such as the set of sequences which fold or function. In this manuscript, however, sequence space is defined as the total possible sequences of a protein that define different interaction specificities. Generally, this means the possible residue combinations at the specificity-determining positions, though some combinations may be specificity-redundant. The concept of sequence space allows us to ask whether there is some coding capacity that ultimately limits evolution's reuse of the same domain, because it is difficult to generate variants with novel specificity to serve novel functions. Until sequence space is saturated, a protein family can continue to duplicate, diverge in specificity and generate insulated channels for information within the cell's network.

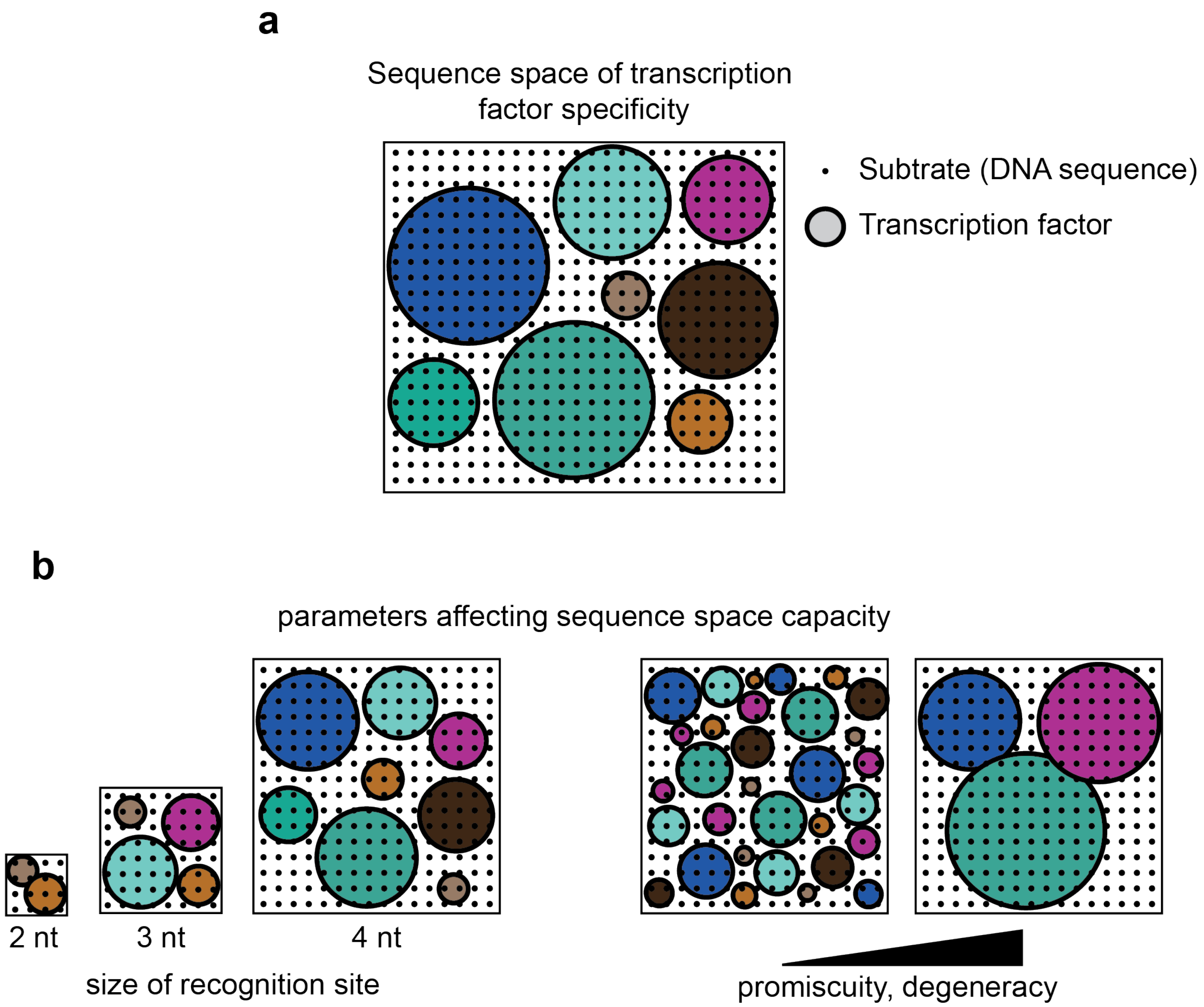


Figure 1.3 Sequence space of macromolecular interactions

(a) The sequence space of a protein family is the total set of substrates. For transcription factors, these substrates are specific DNA sequences. Most transcription factors bind multiple sequences, due to degeneracy at some positions within their recognition motif. In this abstract schematic the set of all compatible substrates for each transcription factor is contained within their respective circle, representing their 'footprint' in sequence space. (b) The capacity of sequence space to harbor orthogonal sets of interacting protein-substrate pairs is limited by the size of the recognition site and the degeneracy of each protein's recognition motif.

The sequence space of macromolecular interactions was first quantitatively approximated by Itzkovitz and Alon in the context of transcription factor DNA-binding specificity (Fig. 1.3) (Itzkovitz et al., 2006). Transcription factor families are tractable models for examining the coding capacity of binding specificity because their members generally interact with the same number of nucleotides and dsDNA has structure that is far more sequence-invariant than proteins. Itzkovitz et al, therefore, could approximate the total sequence space of each transcription factor class as a function of the total nucleotides it bound. They found that, for each class of transcription factor, this sequence space loosely correlated with the maximum number of paralogs found in a single genome. This observation suggests that the size of a transcription factor's sequence space plays some role in limiting its duplication and expansion.

The sequence space of protein-protein interactions is likely to be much larger than that of protein-DNA interactions and is certainly harder to define and measure. There are many more degrees of freedom in the structure of a protein substrate, as compared to a double-stranded DNA substrate. Even if domain backbone structure is highly conserved, a protein surface with N specificity-determining residues has up to 20^N possible specificities, as compared to the 4^N possible sequences of a DNA operator of length N.

There have been some efforts to indirectly study the coding capacity of protein interactions by examining the specificity and evolutionary history of extant proteins in large protein families. An analysis of the nine dominant classes of kinases in eukaryotes suggested that most Ser/Thr phosphorylation motifs were found across eukaryotes, suggesting that there was a burst in novel phospho-specificity that postdates the divergence of prokaryotes and eukaryotes but relatively little specificity innovation since (Bradley and Beltrao, 2019). However, we cannot infer that stasis in specificity drift is due to the limited ability for these kinase domains to encode differential specificity. The specificity of eukaryotic kinases has been shown to be an integration of motif specificity, kinase localization and the specificity of additional recruiting domains (Alexander et al., 2011).

Studies with SH3 and PDZ domains, abundant eukaryotic peptide-binding modules, have suggested that the specificity of these domains has been optimized to maximally diverge across sequence space. If true, this general observation would suggest that sequence space is a strong limiting factor in protein evolution and that global optimization is necessary to generate and maintain insulation between all paralogs in a genome. Zarrinpar and Lim found that a peptide substrate for the SH3 domain of Sho1, Pbs2 from *Saccharomyces cerevisiae,* bound none of the other 26 SH3 domains from the *S. cerevisiae* genome. In contrast to this striking specificity within the *S. cerevisiae* set, they discovered that Pbs2 could bind a number of SH3 domains from other metazoans. Thus, they claimed that global selection against cross-talk must have optimized the full set of *S. cerevisiae* SH3 domains to avoid each other's substrates, but did not endow SH3-peptides pairs with sufficient specificity to be insulated in the context of a different genome. However, the findings of this study were weakened by the fact that 4/6 non-*cerevisiae* SH3 domains compatible with Pbs2 were closely related homologs to Sho1, Pbs2's cognate.

In another study, Stiffler and colleagues used protein microarrays to quantify the peptide binding specificity of all mouse PDZ domains at an unprecedented scale. With these data, they trained a linear regression model and used principle component analysis to reduce the dimensionality of each PDZ's specificity to three parameters. The success of their linear model underscores the independent additivity of the different specificity residues of PDZ-binding peptides. However, they then claimed that mouse PDZs are optimized to span specificity space, because their PDZ domains spread out satisfyingly across their three principle component axes.

This logic, however, is somewhat circular, since principle component analysis selects the eigenvectors that maximally capture the variance in the data. The limits of these two studies make it difficult to make strong claims about the occupancy of the sequence space of paralogous protein interactions.

Deep mutational scanning has promised a new era for our ability to analyze sequence space. By combining large mutant libraries, high-throughput selections and Illumina sequencing, we have generated rich datasets that offer a new appreciation for the complexity of protein specificity (Diss and Lehner, 2018). Scrambling four specificity-determining residues of the bacterial kinase PhoQ revealed the degeneracy of this specificity code: dramatic alterations could, in an unpredictable manner, have little effect on specificity (Podgornaia and Laub, 2015). In part, this is due to epistatic effects – complex dependencies between residues that allow some mutations to compensate or exasperate other mutations. In PhoQ, only 104 of 1,555 functional interface sequences could be predicted from single point mutations using an additive model (Podgornaia and Laub, 2015). Because of the combinatorial complexity of residue interactions, deep mutational scanning gives us much more accurate tools for assessing the size and features of sequence space.

Access to different regions of sequence space may be limited by the fitness of intermediate steps along mutational trajectories. Joseph Thornton's group has shown that phenotypically-silent, epistatic mutations are vital in permitting the specificity switch of steroid receptors for DNA motfs (McKeown et al., 2014). Depending on background mutations, different mutational paths are available to different specificities – some of them much shorter than others (Starr et al., 2017). Some of these permissive mutations stabilize other mutations. Others increase non-specific binding, permitting mutations that afflict all interactions, but some disrupt some by orders of magnitude more than others. Aakre and colleagues showed that libraries of a bacterial anti-toxin, mutated at its specificity-determining residues, contained variants compatible with either the wild-type partner, a paralog, or both. About half specificity switching mutational trajectories passed through a promiscuous intermediate, highlighting the role of negative design elements in specificity divergence (Aakre et al., 2015). The scale of these datasets allows us to grasp probable evolutionary histories and evolutionary potential of protein specificity. However, no study, to date, has utilized this technology to approximate the size of sequence space relevant to protein-protein interactions, and occupancy of this space by the proteins of a single genome.

Two-component pathways: a tool for studying protein sequence space

In the main body of this thesis, I ask questions about the ease with which protein pairs can diverge into orthogonal complexes and the occupancy of the sequence space defined by specificity-determining residues. The appropriate protein-complex model system to address these questions would ideally present a certain number of characteristics. First, the pair of interacting proteins must be members of large families such that there is a broad natural diversity in specificity and sequence. Secondly, there must be a high copy number of paralogs per genome, with different paralogous complexes displaying different specificity and stringent insulation. Finally, it is ideal if paralogous pairs can be predicted by genome context, and if each paralog only interacts with a single, predictable partner. By these parameters, there is perhaps no better model system than two-component signaling pathways. One of the best studied classes of signaling architectures, these pathways are found throughout the genomes of plants, fungi, archaea and especially prokaryotes (Capra and Laub, 2012; Stock et al., 2000). Their frequency across many genomes has made them a vital tool for studying protein evolution. Furthermore, their diverse sensory capacity has made them attractive sources of protein biosensors for biotechnological applications (Daeffler et al., 2017; Tabor et al., 2011).

In their canonical form, two-component pathways comprise two proteins: a sensor histidine kinase, and a downstream effector known as a response regular. Upon stimulation by an environmental signal, a histidine kinase autophosphorylates, then transfers the phosphoryl group to its partner response regulator, thereby activating it (Fig 1.4). The response regulator actuates the appropriate response to the signal, often by acting as a transcription factor. Most histidine kinases are also bifunctional phosphatases, dephosphorylating their cognate regulator in the absence of signal input. This phosphatase activity removes phosphoryl groups that may accumulate due to promiscuous phosphorylation from other kinases or phosphorylated metabolites, e.g. acetyl phosphate (Atkinson et al., 2003; Groban et al., 2009; Podgornaia and Laub, 2015; Wanner and Wilmes-Riesenberg, 1992).

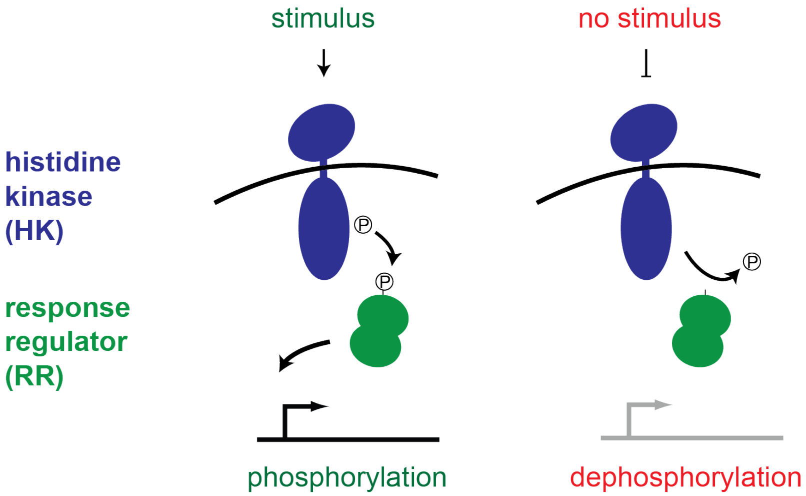


Figure 1.4 - The canonical two-component system.

Sensor histidine kinases exist in an equilibrium between a phosphatase and kinase state, biased towards the former in absence of an input signal. Upon exposure to a specific stimulus, this equilibrium shifts towards the kinase state. In this state, they autophosphorylate, then transfer phosphotransfer to their cognate response regulator, eliciting a downstream response.

The frequency of two component signaling proteins in bacterial genomes is striking (Fig 1.5). Most prokaryotic genomes encode dozens of histidine kinase – response regulator pairs: *Escherichia coli* has 32, *Pseudomonas aeruginosa* has 72 (Gooderham et al., 2008), *Myxococcus xanthus* has 119 (Lee et al., 2010). They are responsible for sensing a diverse array of environmental stimuli. Though the input signal for most two-component systems is unknown, some characterized histidine kinases sense nutrients, inorganic ions, the antibiotic vancomycin, quorum signaling peptides, and specific wavelengths of light (Cheung and Hendrickson, 2010; Schmidl et al., 2014). They also monitor cellular states such as membrane stress, temperature, and nitrogen levels. Each sensor kinase typically interacts with a single response regulator, and vice versa. The specificity of this interaction allows these diverse information channels to operate without interference.

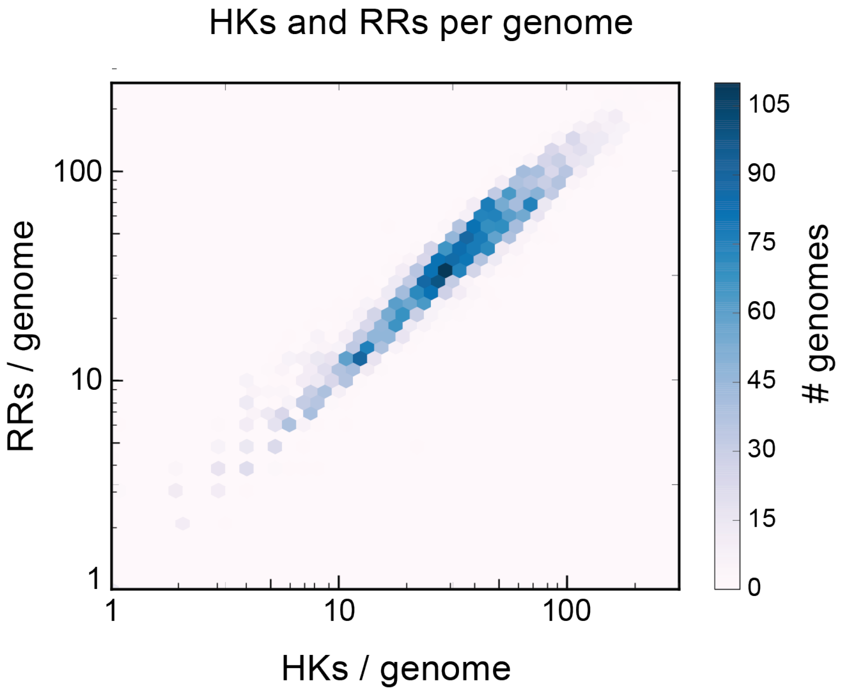


Figure 1.5 Two-component proteins per bacterial genome.

A density map of the total number response regulator and histidine kinase genes found per genome within the RefSeq Prokaryotic Genomes database of 5,506 total bacterial genomes (September 2017).

Mechanism of action

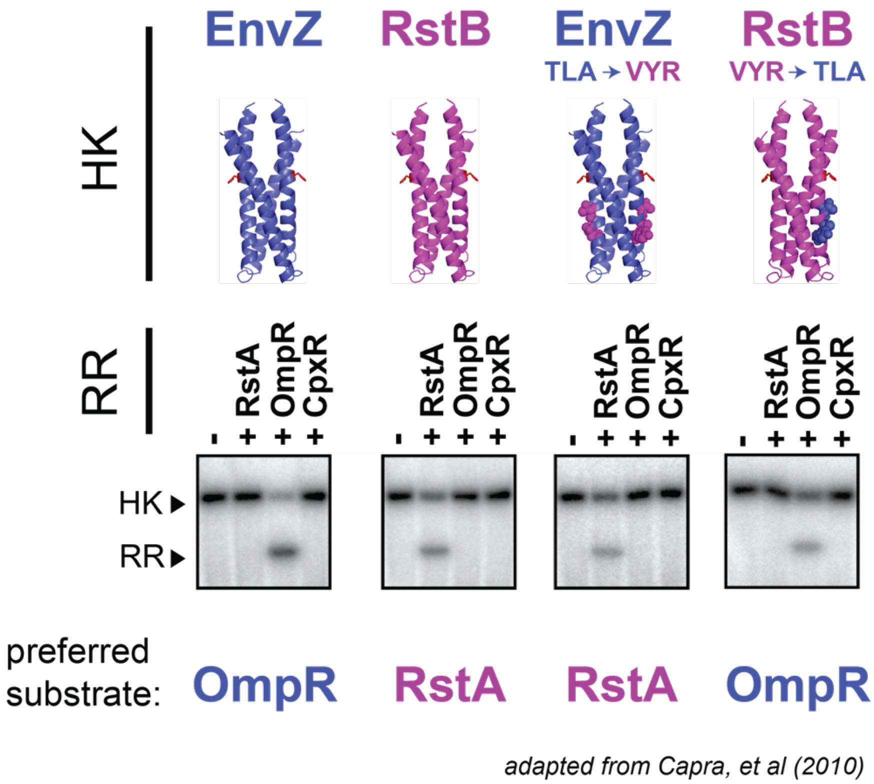
Histidine kinases are able to sense a broad array of inputs due to their highly variable domain architectures. The canonical architecture consists of one or more N-terminal sensory domains, a signal amplifying region such as a HAMP, GAF or PAS domain, and then the defining domains of the histidine kinase: a dimerizing and histidine-phosphotransfer (DHp) domain and a catalytic ATP-binding (CA) domain. Despite diverse layouts, these kinases activate through a conserved mechanism of action. In the absence of signal, the DHp domain remains a rigid, symmetrical, four-helix bundle, holding the CA domain far from the catalytic histidine. Structures of one histidine kinase, DesK from *Bacillus subtilis*, show that the catalytic histidine becomes buried within the core of the bundle during this state, protected from any phosphodonor (Trajtenberg et al., 2016). Activation of the sensory domain(s) causes subtle twisting and piston movement of one of the helices N-terminal to the DHp domain (Gushchin et al., 2017). This twist destabilizes the coiled-coil at the beginning of the domain (Saita et al., 2015; Trajtenberg et al., 2016) eliciting two conformational effects: (1) the catalytic histidine is rotated out from the core, becoming exposed, and (2) the catalytic CA domain is released from the now destabilized coiled-coil and can swing down to phosphorylate the histidine.

Once phosphorylated, histidine kinases can transfer to the phosphoryl group onto their cognate regulator to elicit a conserved confirmation change in its receiver domain. Catalyzed by a magnesium coordinated in the regulator's activate site, the phosphoryl group is transferred from the kinase's histidine to an aspartic acid on the response regulator. The phosphorylated aspartate forms a hydrogen bond with a conserved Ser/Thr switch residue and pulls a second aromatic switch residue (Gao and Stock, 2015). The implications of this conserved conformational change vary slightly between the REC domains of different regulators, but all involve some perturbation of the α4-β5-α5 surface of the REC domain's αβ-fold. Generally, the α4 helix slides, exposing a previously-buried interface of the REC domain and allowing new interactions and behaviors.

Domain modularity

The domain modularity of two-component systems has likely contributed to their widespread adoption to sense diverse stimuli (Jacob-Dubuisson et al., 2018). For histidine kinases, the minimal functional requirement for activation (just destabilizing the coiled coil proximal the kinase domain) has allowed these systems to adopt hundreds of different sensory domains; InterPro currently lists 391 domain architecture variations (Hunter et al., 2009). Many sensory domains are extracellular or transmembrane, but control a cytoplasmic kinase domain through manipulation of transmembrane helices. The membrane separation between sensory and kinase domains indicates that all signaling information is encoded in relative movement of a few transmembrane helices. Protein engineers have demonstrated the high degree of the domain modularity here by constructing a plethora of functional kinase fusions (Bi et al., 2016; Hori et al., 2017; Lehning et al., 2017; Levskaya et al., 2005; Ravikumar et al., 2018b). Some of these chimeras enlist sensory domains from other histidine kinases, but others do not. The Moglich and Muir groups have beautifully demonstrated that the linkers between sensory domains and DHp domains have a critical periodicity - adding 7n coil residues maintains function and adding 7n+1 residues generates a sensor with opposite signaling logic (Ohlendorf et al., 2016; Wang et al., 2014).

The domain modularity of response regulators is possibly even more flexible than the modularity of kinases. The alpha/beta Rossman-like fold of the REC domain exists in equilibrium between two conformations – one stable when phosphorylated and one stable when unmodified. Phosphorylation generally catalyzes the movement of one alpha helix, burying one surface and exposing another. For many regulators, such as the OmpR/PhoB class (29% of known response regulators) (Gao and Stock, 2010), this transition exposes a homo-dimerizing surface and generates an active transcription factor complex with two DNA-binding domains capable of targeting palindromic regulatory sequences. Many other oligomerization-dependent domains found in response regulators, such as RNA-binding proteins, and di-GMP cyclases, AAA+ ATPases and phosphodiesterases, highlight the flexibility of this activation strategy (Gao and Stock, 2010). In other transcription-modulating regulators, such as the NarL class (19% of known response regulators) (Gao and Stock, 2010), the REC domain covers its own DNA-binding domain in an auto-inhibitory fashion until phosphorylation blocks this interaction. These intuitive and common mechanisms demonstrate the versatility of REC phosphorylation, but probably represent a small fraction of the activation strategies used by response regulators, given the enormous variability in domain architecture. Highlighting this versatility, biological engineers have recently constructed large sets of functional regulator chimeras by swapping REC domains (Schmidl et al., 2019).



adapted from Capra, et al (2010)

Figure 1.6 Rational rewiring of histidine kinase specificity.

Kinases EnvZ and RstB preferentially phosphorylate their respective regulator substrates, OmpR and RstA, as seen by *in vitro* phosphotransfer assays. Gel images show radiolabeled phosphate moving from kinases to regulators only when proteins are compatible. Switching three specificity residues is sufficient to swap the specificity of EnvZ and RstB, such that they phosphorylate only each other's natural substrate, not their own.

Specificity and insulation of two-component signaling pathways

Most bacterial genomes encode dozens of two-component signaling pathways, each functioning through structurally conserved interactions between a kinase's DHp domain and regulator's REC domain. The specificity of this interaction is thus critical for wiring environmental signals to the appropriate response. Most histidine kinases and response regulators interact with only a single partner, such that cells are comprised of many parallel, insulated pathways. Perhaps because bacteria lack the organelle complexity of eukaryotes, the specificity of these interactions is generally determined by lock-and-key molecular recognition (Skerker et al., 2005). Phosphoprofiling assays, in which individual kinase domains were tested *in vitro* against panels of response regulators, revealed that these kinases have a global kinetic preference for a single target. A kinase's phosphorylation efficiency (k_{cat} / K_m) with its cognate regulator is often a thousand-fold higher than with the next best target. In addition to molecular recognition, secondary buffering mechanisms help to enforce specificity (Podgornaia and Laub, 2013). Deletion of a kinase can result in cross-talk from other kinases onto the now orphaned regulator, highlighting the importance of a kinase's phosphatase activity in reducing spurious phosphorylation (Groban et al., 2009; Siryaporn and Goulian, 2008). Deletion of a response regular can also cause its cognate kinase to inappropriately phosphorylate other regulators, suggesting that competitive binding also plays a role in specificity (Groban et al., 2009; Siryaporn and Goulian, 2008).

The molecular recognition of this interaction is governed by a set of critical interfacial residues that co-evolve. On the histidine kinase, these are located in a short stretch of positions starting seven amino acids away to the catalytic histidine. Transplanting this short stretch from five *E. coli* kinases, RstB, CpxA, PhoR , AtoS and PhoQ, into the DHp domain of a sixth kinase EnvZ endowed EnvZ with identical specificity of each of these kinases (Skerker et al., 2008). Using mutual information to quantify the coevolutionary dependencies between residues, the specificity-determining residues were further narrowed to a small set of noncontiguous residues. Transplanting as few as three residues from this patch between kinases can effectively swap their phosphorylation specificity (Fig 1.6) (Capra et al., 2010). In addition to swapping their phosphotransferase preference, rewiring kinases results in a change of binding affinity from > 50uM to the 1uM range typical for cognate pairs (Willett et al., 2013).

The role of these residues in specificity was later supported by crystal structures of kinase-regulator complexes (Fig 1.7) (Casino et al., 2009). The main specificity-determining contacts occur between the surface of two kinase alpha helices with, primarily, a single alpha helix on the regulator. This interaction patch is central to the buried surface within the complex, though slightly too low for specificity residues to make direct contact with catalytic residues. Though specificity residues are tightly packed and form a dense network of dependencies across the interface, the physical separation between the catalytic and specificity-determining regions highlights how these sections behave modularly. This separation is likely why these systems seem amenable to rewiring specificity without losing function. In the most direct example of this modularity, Podgornaia and Laub were able to transfer the majority of specificity residues (five from the kinase, four from the regulator) from one kinase-regulator complex, PhoR-PhoB, into HK853-RR468 to generate a functional new complex HK853*-RR468* (Podgornaia et al., 2013). As previously mentioned, this rewired pair no longer interacted with parent proteins, yet crystalized with a similar quaternary structure to the parent complex. When the parent kinase was crystalized with the incompatible, rewired regulator, they were found to bind at a significantly altered angle, such that the catalytic residues were four times further apart than in the functional complexes. Seemingly, the role of the specificity residues is not only to contribute sufficient binding energy, but also to align the helices of the two proteins such that upstream catalytic residues can interact.

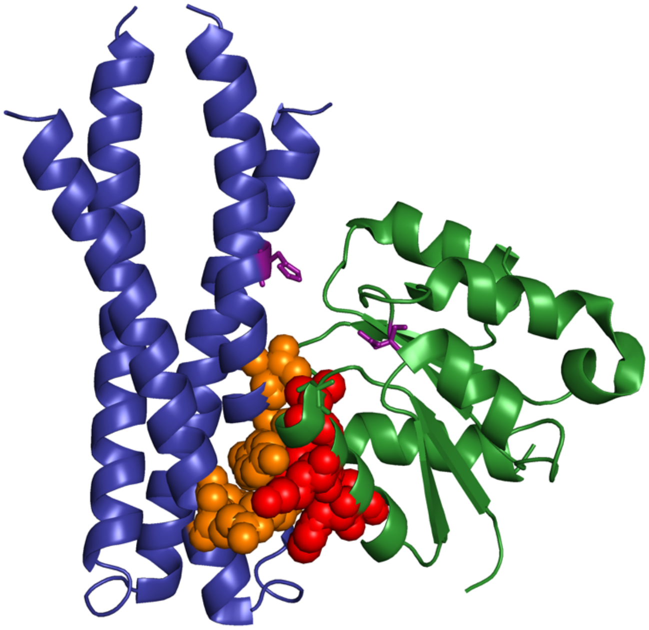
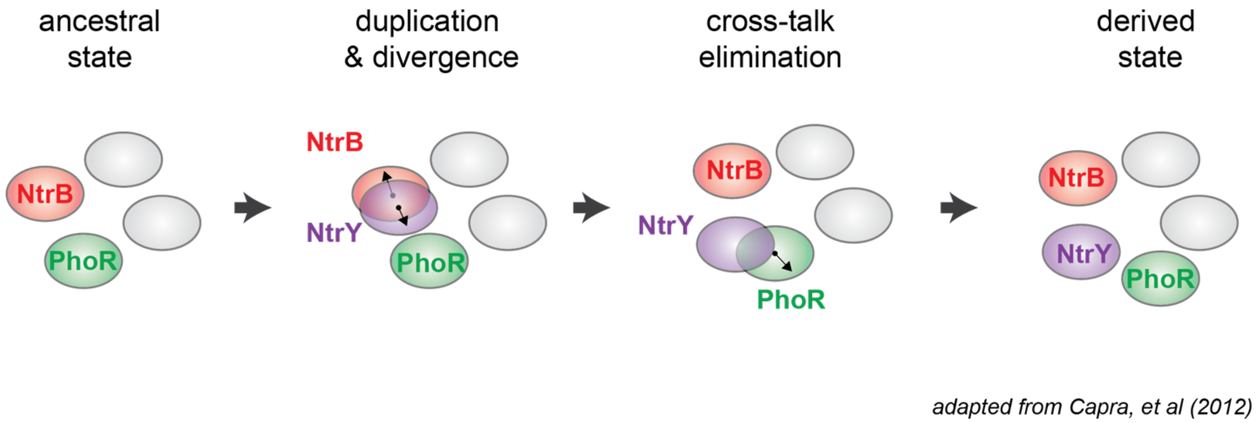


Figure 1.7 Specificity residues of the histidine kinase - response regulator complex

Structure of the DHp domain of a histidine kinase, TM0853 (blue) in complex with its cognate response regulator, RR0468 (green). Residues that dictate specificity are spacefilled in orange (kinase) and red (regulator). Catalytic histidine and aspartate residues are colored in purple.

Sequence space of specificity-determining residues

The observation that a small set of residues is sufficient for determining the specificity of two-component systems raises questions about the coding capacity of these paralogous interactions. The specificity residues define a finite sequence space. Does the size of this sequence space limit for the number of two-component pathways a cell can use? Or is it trivial to continue evolving more paralogous pathways with orthogonal specificity? The answer determines whether the evolutionary expansion of these pathways, via either horizontal transfer or gene duplication, will result in cross-talk between unrelated pathways. If sequence space is crowded, it could limit both how interacting proteins can evolve and our ability to engineer novel organisms with additional pathways.



adapted from Capra, et al (2012)

Figure 1.8 Paralog interference during the duplication of NtrBC.

Duplication of NtrBC produced NtrXY, which caused crosstalk with the unrelated two-component pathway PhoBR. Specificity residues of PhoR changed in order to reestablish global insulation. Schematic illustrates occupancy and movement of each histidine kinase through the sequence space defined by their specificity residues.

By examining two-component system phylogeny, Capra and colleagues identified one historical example where interference between two-component pathways constrained their evolution (Capra et al., 2012). Duplication of the NtrBC pathway in α-proteobacteria produced the two daughter systems: NtrBC and NtrXY. Interestingly, this divergence event coincided with a change in the specificity residues of PhoBR, an unrelated pathway. Biochemical investigations with the ancestral PhoBR showed that it cross-talked dramatically with NtrXY. Consequently, the researchers were elucidated that the duplication event had forced PhoBR to change its specificity residues, to move in sequence space, in order to make room for the new NtrBC pair and resolve NtrY-PhoB cross talk (Fig. 1.8). However, we do not know how often these types of paralog interference events occur. This observation also does not mean that sequence space is saturated. Sequence space collisions such as these could result from both global saturation, i.e. no additional two-component pairs can be generated with novel specificity, or from local saturation, i.e. there are just too many systems similar to NtrBC.

Systematic analysis of the histidine kinase PhoQ has provided a sense for the functional degeneracy of specificity residues. Here we define degeneracy as the phenomenon where of different amino acid sequences, at the positions that determine binding, still encode the same specificity. For example, in some contexts, the substitution of leucine for isoleucine may have profound changes for binding specificity, but in many contexts this alteration is too physiochemically subtle to have consequences and is thus functionally degenerate. Degeneracy is an important parameter in establishing the coding capacity of a particular protein-protein interaction; more degeneracy results in smaller coding capacity because fewer orthogonal interactions are possible (Fig. 1.3b). Podgornaia and Laub screened a mutant library in which four of the PhoQ's specificity residues were mutated all amino acids (Podgornaia and Laub, 2015). This systematic mutation revealed that a diverse mixture of residues was permitted at each specificity position, without breaking interaction with the cognate regulator PhoP. However, of the 160,000 possible residue combinations, only 1% remained functional and compatible with PhoP. Thus, despite some degeneracy, PhoP is specific for only a small fraction of possible PhoQ surfaces. This fraction is likely much smaller than 1% because additional residues, not mutated in this study, play a significant role in the specificity of histidine kinases (Skerker et al., 2008).

Conclusion

Proteins are complex molecules. Their diverse structures and functions are defined - in ways we cannot always predict, despite decades of effort - entirely by their sequence of amino acids. They form complex, rapidly-evolving networks of interactions that determine how cells process information into behavioral responses. Much of this network complexity is thought to have evolved as genes have duplicated and diverged, producing large families of structurally-conserved protein domains with different interfaces. Yet, despite their complexity, the specificity of these interactions, and therefore the crucial connectivity of the network, are often determined by a small set of residues. Bioinformatic strategies leveraging residue co-variation have proved increasingly successful at predicting these specificity-determining residues from available genomic databases. For a given pair of interacting protein domains, the specificity-determining residues define a specificity space; the size of this space limits how much biology can re-use these domains in homologous proteins. Paralogs utilizing the same space can interact, preventing isolated information channels from functioning independently, sometimes with devastating consequences. The occupancy of sequence space by the paralogous proteins of a particular organism, therefore, will establish if global rewiring is necessary to pack new paralogous domains into that genome. Measuring this sequence space occupancy is critical for understanding the systems-level restrictions on gene duplication or horizonal transfer. It is also one of the prerequisites for high-confidence engineering of biological systems. This is because biological engineers, like evolutionary processes, reuse extant protein structures as components for novel designs, risking unintended cross-talk with endogenous pathways or other engineered circuits.

Two component signaling pathways are ideal model systems for studying the capacity and occupancy of sequence space. This is primarily because their components, histidine kinases and response regulators, are some of the most frequent genes within sequenced genomes. It is also extremely convenient that these interacting proteins generally form highly specific, one-to-one pairs, and that pairing can be inferred directly from gene synteny. Furthermore, there is great engineering interest in two-component systems because many extant protein sensors, for both organic and inorganic environmental signatures, are sensor histidine kinases. Finally, the residues that determine the specificity of the kinase-regular interaction have been identified and verified through many independent examples of specificity transplantation. Nonetheless, we know little about the size and occupancy of specificity space. In the following chapters, we will pursue a better understanding of two-component specificity space and how it may limit the installation of new pathways in a cell.

References

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell *163*, 594-606.

Alexander, J., Lim, D., Joughin, B.A., Hegemann, B., Hutchins, J.R., Ehrenberger, T., Ivins, F., Sessa, F., Hudecz, O., Nigg, E.A., *et al.* (2011). Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. Sci Signal *4*, ra42.

AlQuraishi, M. (2019). AlphaFold at CASP13. Bioinformatics.

An, W., and Chin, J.W. (2009). Synthesis of orthogonal transcription-translation networks. Proc Natl Acad Sci U S A *106*, 8477-8482.

Arabidopsis Interactome Mapping, C. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Atkinson, M.R., Savageau, M.A., Myers, J.T., and Ninfa, A.J. (2003). Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli. Cell *113*, 597-607.

Baker, C.R., Hanson-Smith, V., and Johnson, A.D. (2013). Following gene duplication, paralog interference constrains transcriptional circuit evolution. Science *342*, 104-108.

Baker, C.R., Tuch, B.B., and Johnson, A.D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc Natl Acad Sci U S A *108*, 7493-7498.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061-1078.

Barnea, G., Strapps, W., Herrada, G., Berman, Y., Ong, J., Kloss, B., Axel, R., and Lee, K.J. (2008). The genetic design of signaling cascades to record receptor activation. Proc Natl Acad Sci U S A *105*, 64-69.

Bi, S., Pollard, A.M., Yang, Y., Jin, F., and Sourjik, V. (2016). Engineering Hybrid Chemotaxis Receptors in Bacteria. ACS Synth Biol *5*, 989-1001.

Bitbol, A.F., Dwyer, R.S., Colwell, L.J., and Wingreen, N.S. (2016). Inferring interaction partners from protein sequences. Proc Natl Acad Sci U S A *113*, 12180-12185.

Bogan, A.A., and Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. J Mol Biol *280*, 1-9.

Boyken, S.E., Chen, Z., Groves, B., Langan, R.A., Oberdorfer, G., Ford, A., Gilmore, J.M., Xu, C., DiMaio, F., Pereira, J.H., *et al.* (2016). De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. Science *352*, 680-687.

Bradley, D., and Beltrao, P. (2019). Evolution of protein kinase substrate recognition at the active site. PLoS Biol *17*, e3000341.

Bradley, D., Viéitez, C., Rajeeve, V., Cutillas, P.R., and Beltrao, P. (2018). Global analysis of specificity determinants in eukaryotic protein kinases. bioRxiv, 195115.

Capra, E.J., and Laub, M.T. (2012). Evolution of two-component signal transduction systems. Annu Rev Microbiol *66*, 325-347.

Capra, E.J., Perchuk, B.S., Lubin, E.A., Ashenberg, O., Skerker, J.M., and Laub, M.T. (2010). Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. PLoS Genet *6*, e1001220.

Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. Cell *150*, 222-232.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Cheng, R.R., Haglund, E., Tiee, N.S., Morcos, F., Levine, H., Adams, J.A., Jennings, P.A., and Onuchic, J.N. (2018). Designing bacterial signaling interactions with coevolutionary landscapes. PLoS One *13*, e0201734.

Cheng, R.R., Morcos, F., Levine, H., and Onuchic, J.N. (2014). Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Natl Acad Sci U S A *111*, E563-571.

Cheung, J., and Hendrickson, W.A. (2010). Sensor domains of two-component regulatory systems. Curr Opin Microbiol *13*, 116-123.

Chevalier, A., Silva, D.A., Rocklin, G.J., Hicks, D.R., Vergara, R., Murapa, P., Bernard, S.M., Zhang, L., Lam, K.H., Yao, G*., et al.* (2017). Massively parallel de novo protein design for targeted therapeutics. Nature *550*, 74-79.

Chow, D.C., Rice, K., Huang, W., Atmar, R.L., and Palzkill, T. (2016). Engineering Specificity from Broad to Narrow: Design of a beta-Lactamase Inhibitory Protein (BLIP) Variant That Exclusively Binds and Detects KPC beta-Lactamase. ACS Infect Dis *2*, 969-979.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. Science *267*, 383-386.

Cong, Q., Anishchenko, I., Ovchinnikov, S., and Baker, D. (2019). Protein interaction networks revealed by proteome coevolution. Science *365*, 185-189.

Creixell, P., Palmeri, A., Miller, C.J., Lou, H.J., Santini, C.C., Nielsen, M., Turk, B.E., and Linding, R. (2015a). Unmasking determinants of specificity in the human kinome. Cell *163*, 187-201.

Creixell, P., Schoof, E.M., Simpson, C.D., Longden, J., Miller, C.J., Lou, H.J., Perryman, L., Cox, T.R., Zivanovic, N., Palmeri, A*., et al.* (2015b). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. Cell *163*, 202-217.

Cuff, A., Redfern, O.C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A*., et al.* (2009). The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. Structure *17*, 1051-1062.

Daeffler, K.N., Galley, J.D., Sheth, R.U., Ortiz-Velez, L.C., Bibb, C.O., Shroyer, N.F., Britton, R.A., and Tabor, J.J. (2017). Engineering bacterial thiosulfate and tetrathionate sensors for detecting gut inflammation. Mol Syst Biol *13*, 923.

Dimas, R.P., Jiang, X.L., Alberto de la Paz, J., Morcos, F., and Chan, C.T.Y. (2019). Engineering repressors with coevolutionary cues facilitates toggle switches with a master reset. Nucleic Acids Res *47*, 5449-5463.

Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. Elife *7*.

Dutta, S., Gulla, S., Chen, T.S., Fire, E., Grant, R.A., and Keating, A.E. (2010). Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. J Mol Biol *398*, 747-762.

Elcock, A.H. (2010). Models of macromolecular crowding effects and the need for quantitative comparisons with experiment. Curr Opin Struct Biol *20*, 196-206.

Gao, R., and Stock, A.M. (2010). Molecular strategies for phosphorylation-mediated regulation of response regulator activity. Curr Opin Microbiol *13*, 160-167.

Gao, R., and Stock, A.M. (2015). Temporal hierarchy of gene expression mediated by transcription factor binding affinity and activation dynamics. MBio *6*, e00686-00615.

Gee, S.H., Quenneville, S., Lombardo, C.R., and Chabot, J. (2000). Single-amino acid substitutions alter the specificity and affinity of PDZ domains for their ligands. Biochemistry *39*, 14638-14646.

Gooderham, W.J., Bains, M., McPhee, J.B., Wiegand, I., and Hancock, R.E. (2008). Induction by cationic antimicrobial peptides and involvement in intrinsic polymyxin and antimicrobial peptide resistance, biofilm formation, and swarming motility of PsrA in Pseudomonas aeruginosa. J Bacteriol *190*, 5624-5634.

Gray, V.E., Hause, R.J., Luebeck, J., Shendure, J., and Fowler, D.M. (2018). Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Syst *6*, 116-124 e113.

Gretes, M., Lim, D.C., de Castro, L., Jensen, S.E., Kang, S.G., Lee, K.J., and Strynadka, N.C. (2009). Insights into positive and negative requirements for protein-protein interactions by crystallographic analysis of the beta-lactamase inhibitory proteins BLIP, BLIP-I, and BLP. J Mol Biol *389*, 289-305.

Groban, E.S., Clarke, E.J., Salis, H.M., Miller, S.M., and Voigt, C.A. (2009). Kinetic buffering of cross talk between bacterial two-component sensors. J Mol Biol *390*, 380-393.

Gushchin, I., Melnikov, I., Polovinkin, V., Ishchenko, A., Yuzhakova, A., Buslaev, P., Bourenkov, G., Grudinin, S., Round, E., Balandin, T*., et al.* (2017). Mechanism of transmembrane signaling by sensor histidine kinases. Science *356*.

Hogan, G.J., Brown, P.O., and Herschlag, D. (2015). Evolutionary Conservation and Diversification of Puf RNA Binding Proteins and Their mRNA Targets. PLoS Biol *13*, e1002307.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. Cell *149*, 1607-1621.

Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Scharfe, C.P., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. Nat Biotechnol *35*, 128-135.

Hori, M., Oka, S., Sugie, Y., Ohtsuka, H., and Aiba, H. (2017). Construction of a photo-responsive chimeric histidine kinase in Escherichia coli. J Gen Appl Microbiol *63*, 44-50.

Howard, P.L., Chia, M.C., Del Rizzo, S., Liu, F.F., and Pawson, T. (2003). Redirecting tyrosine kinase signaling to an apoptotic caspase pathway through chimeric adaptor proteins. Proc Natl Acad Sci U S A *100*, 11267-11272.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L*., et al.* (2009). InterPro: the integrative protein signature database. Nucleic Acids Res *37*, D211-215.

Ingham, R.J., Colwill, K., Howard, C., Dettwiler, S., Lim, C.S., Yu, J., Hersi, K., Raaijmakers, J., Gish, G., Mbamalu, G*., et al.* (2005). WW domains provide a platform for the assembly of multiprotein networks. Mol Cell Biol *25*, 7092-7106.

Itzkovitz, S., Tlusty, T., and Alon, U. (2006). Coding limits on the number of transcription factors. BMC Genomics *7*, 239.

Jacob-Dubuisson, F., Mechaly, A., Betton, J.M., and Antoine, R. (2018). Structural insights into the signalling mechanisms of two-component systems. Nat Rev Microbiol *16*, 585-593.

Janin, J., Rodier, F., Chakrabarti, P., and Bahadur, R.P. (2007). Macromolecular recognition in the Protein Data Bank. Acta Crystallogr D Biol Crystallogr *63*, 1-8.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Jones, D.T., Buchan, D.W., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184-190.

Jones, P.T., Dear, P.H., Foote, J., Neuberger, M.S., and Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature *321*, 522-525.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A *110*, 15674-15679.

Kastritis, P.L., and Bonvin, A.M. (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res *9*, 2216-2225.

Keskin, O., Ma, B., and Nussinov, R. (2005). Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol *345*, 1281-1294.

Khafizov, K., Madrid-Aliste, C., Almo, S.C., and Fiser, A. (2014). Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A *111*, 3733-3738.

Korber, B.T., Farber, R.M., Wolpert, D.H., and Lapedes, A.S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A *90*, 7176-7180.

Latysheva, N.S., and Babu, M.M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res *44*, 4487-4503.

Lee, B., Schramm, A., Jagadeesan, S., and Higgs, P.I. (2010). Two-component systems and regulation of developmental progression in Myxococcus xanthus. Methods Enzymol *471*, 253-278.

Lehning, C.E., Heidelberger, J.B., Reinhard, J., Norholm, M.H.H., and Draheim, R.R. (2017). A Modular High-Throughput In Vivo Screening Platform Based on Chimeric Bacterial Receptors. ACS Synth Biol *6*, 1315-1326.

Levin, K.B., Dym, O., Albeck, S., Magdassi, S., Keeble, A.H., Kleanthous, C., and Tawfik, D.S. (2009). Following evolutionary paths to protein-protein interactions with high affinity and selectivity. Nat Struct Mol Biol *16*, 1049-1055.

Levskaya, A., Chevalier, A.A., Tabor, J.J., Simpson, Z.B., Lavery, L.A., Levy, M., Davidson, E.A., Scouras, A., Ellington, A.D., Marcotte, E.M.*, et al.* (2005). Synthetic biology: engineering Escherichia coli to see light. Nature *438*, 441-442.

Li, L., Shakhnovich, E.I., and Mirny, L.A. (2003). Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. Proc Natl Acad Sci U S A *100*, 4463-4468.

Lim, W.A. (2010). Designing customized cell signalling circuits. Nat Rev Mol Cell Biol *11*, 393-403.

Liu, C.C., Jewett, M.C., Chin, J.W., and Voigt, C.A. (2018a). Toward an orthogonal central dogma. Nat Chem Biol *14*, 103-106.

Liu, X., Fan, K., and Wang, W. (2004). The number of protein folds and their distribution over families in nature. Proteins *54*, 491-499.

Liu, Y., Palmedo, P., Ye, Q., Berger, B., and Peng, J. (2018b). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. Cell Syst *6*, 65-74 e63.

Long, M., VanKuren, N.W., Chen, S., and Vibranovski, M.D. (2013). New gene evolution: little did we know. Annu Rev Genet *47*, 307-333.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

Martin, J. (2010). Beauty is in the eye of the beholder: proteins can recognize binding sites of homologous proteins in more than one way. PLoS Comput Biol *6*, e1000821.

McKeown, A.N., Bridgham, J.T., Anderson, D.W., Murphy, M.N., Ortlund, E.A., and Thornton, J.W. (2014). Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. Cell *159*, 58-68.

Meenan, N.A., Sharma, A., Fleishman, S.J., Macdonald, C.J., Morel, B., Boetzel, R., Moore, G.R., Baker, D., and Kleanthous, C. (2010). The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc Natl Acad Sci U S A *107*, 10080-10085.

Melero, C., Ollikainen, N., Harwood, I., Karpiak, J., and Kortemme, T. (2014). Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. Proc Natl Acad Sci U S A *111*, 15426-15431.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A *108*, E1293-1301.

Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2007). Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins *68*, 803-812.

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2017). Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. Proc Natl Acad Sci U S A *114*, 11703-11708.

Netzer, R., Listov, D., Lipsh, R., Dym, O., Albeck, S., Knop, O., Kleanthous, C., and Fleishman, S.J. (2018). Ultrahigh specificity in a network of computationally designed protein-interaction pairs. Nat Commun *9*, 5286.

Nicoludis, J.M., Lau, S.Y., Scharfe, C.P., Marks, D.S., Weihofen, W.A., and Gaudet, R. (2015). Structure and Sequence Analyses of Clustered Protocadherins Reveal Antiparallel Interactions that Mediate Homophilic Specificity. Structure *23*, 2087-2098.

Nocedal, I., and Johnson, A.D. (2015). How Transcription Networks Evolve and Produce Biological Novelty. Cold Spring Harb Symp Quant Biol *80*, 265-274.

Nocedal, I., Mancera, E., and Johnson, A.D. (2017). Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. Elife *6*.

O'Maille, P.E., Malone, A., Dellas, N., Andes Hess, B., Jr., Smentek, L., Sheehan, I., Greenhagen, B.T., Chappell, J., Manning, G., and Noel, J.P. (2008). Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. Nat Chem Biol *4*, 617-623.

Ohlendorf, R., Schumacher, C.H., Richter, F., and Moglich, A. (2016). Library-Aided Probing of Linker Determinants in Hybrid Photoreceptors. ACS Synth Biol *5*, 1117-1126.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife *3*, e02030.

Ovchinnikov, S., Kim, D.E., Wang, R.Y., Liu, Y., DiMaio, F., and Baker, D. (2016). Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. Proteins *84 Suppl 1*, 67-75.

Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F., and Baker, D. (2018). Protein structure prediction using Rosetta in CASP12. Proteins *86 Suppl 1*, 113-121.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. Science *355*, 294-298.

Park, S.H., Zarrinpar, A., and Lim, W.A. (2003). Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. Science *299*, 1061-1064.

Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science *300*, 445-452.

Perez, J.C., Fordyce, P.M., Lohse, M.B., Hanson-Smith, V., DeRisi, J.L., and Johnson, A.D. (2014). How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes. Genes Dev *28*, 1272-1277.

Podgornaia, A.I., Casino, P., Marina, A., and Laub, M.T. (2013). Structural basis of a rationally rewired protein-protein interface critical to bacterial signaling. Structure *21*, 1636-1647.

Podgornaia, A.I., and Laub, M.T. (2013). Determinants of specificity in two-component signal transduction. Curr Opin Microbiol *16*, 156-162.

Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science *347*, 673-677.

Pougach, K., Voet, A., Kondrashov, F.A., Voordeckers, K., Christiaens, J.F., Baying, B., Benes, V., Sakai, R., Aerts, J., Zhu, B.*, et al.* (2014). Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nat Commun *5*, 4868.

Ravikumar, A., Arrieta, A., and Liu, C.C. (2014). An orthogonal DNA replication system in yeast. Nat Chem Biol *10*, 175-177.

Ravikumar, A., Arzumanyan, G.A., Obadi, M.K.A., Javanpour, A.A., and Liu, C.C. (2018a). Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. Cell *175*, 1946-1957 e1913.

Ravikumar, S., David, Y., Park, S.J., and Choi, J.I. (2018b). A Chimeric Two-Component Regulatory System-Based Escherichia coli Biosensor Engineered to Detect Glutamate. Appl Biochem Biotechnol *186*, 335-349.

Reinke, A.W., Baek, J., Ashenberg, O., and Keating, A.E. (2013). Networks of bZIP protein-protein interactions diversified over a billion years of evolution. Science *340*, 730-734.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat Methods *15*, 816-822.

Rollins, N.J., Brock, K.P., Poelwijk, F.J., Stiffler, M.A., Gauthier, N.P., Sander, C., and Marks, D.S. (2019). Inferring protein 3D structure from deep mutation scans. Nat Genet *51*, 1170-1176.

Saita, E., Abriata, L.A., Tsai, Y.T., Trajtenberg, F., Lemmin, T., Buschiazzo, A., Dal Peraro, M., de Mendoza, D., and Albanesi, D. (2015). A coiled coil switch mediates cold sensing by the thermosensory protein DesK. Mol Microbiol *98*, 258-271.

Sammond, D.W., Eletr, Z.M., Purbeck, C., and Kuhlman, B. (2010). Computational design of second-site suppressor mutations at protein-protein interfaces. Proteins *78*, 1055-1065.

Schmidl, S.R., Ekness, F., Sofjan, K., Daeffler, K.N., Brink, K.R., Landry, B.P., Gerhardt, K.P., Dyulgyarov, N., Sheth, R.U., and Tabor, J.J. (2019). Rewiring bacterial two-component systems by modular DNA-binding domain swapping. Nat Chem Biol *15*, 690-698.

Schmidl, S.R., Sheth, R.U., Wu, A., and Tabor, J.J. (2014). Refactoring and optimization of light-switchable Escherichia coli two-component systems. ACS Synth Biol *3*, 820-831.

Schmiedel, J.M., and Lehner, B. (2019). Determining protein structures using deep mutagenesis. Nat Genet *51*, 1177-1186.

Schreiber, G., and Keating, A.E. (2011). Protein binding specificity versus promiscuity. Curr Opin Struct Biol *21*, 50-61.

Segall-Shapiro, T.H., Meyer, A.J., Ellington, A.D., Sontag, E.D., and Voigt, C.A. (2014). A 'resource allocator' for transcription based on a highly fragmented T7 RNA polymerase. Mol Syst Biol *10*, 742.

Shrestha, R., Garrett, S.C., Almo, S.C., and Fiser, A. (2019). Computational Redesign of PD-1 Interface for PD-L1 Ligand Selectivity. Structure *27*, 829-836 e823.

Sibener, L.V., Fernandes, R.A., Kolawole, E.M., Carbone, C.B., Liu, F., McAffee, D., Birnbaum, M.E., Yang, X., Su, L.F., Yu, W*., et al.* (2018). Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding. Cell *174*, 672-687 e627.

Siryaporn, A., and Goulian, M. (2008). Cross-talk suppression between the CpxA-CpxR and EnvZ-OmpR two-component systems in E. coli. Mol Microbiol *70*, 494-506.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Sockolosky, J.T., Trotta, E., Parisi, G., Picton, L., Su, L.L., Le, A.C., Chhabra, A., Silveria, S.L., George, B.M., King, I.C*., et al.* (2018). Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. Science *359*, 1037-1042.

Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories in the sequence space of an ancient protein. Nature *549*, 409-413.

Stein, R.R., Marks, D.S., and Sander, C. (2015). Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. PLoS Comput Biol *11*, e1004182.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-component signal transduction. Annu Rev Biochem *69*, 183-215.

Sunden, F., Peck, A., Salzman, J., Ressl, S., and Herschlag, D. (2015). Extensive site-directed mutagenesis reveals interconnected functional units in the alkaline phosphatase active site. Elife *4*.

Tabor, J.J., Levskaya, A., and Voigt, C.A. (2011). Multichromatic control of gene expression in Escherichia coli. J Mol Biol *405*, 315-324.

Tan, C.S., Pasculescu, A., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2009). Positive selection of tyrosine loss in metazoan evolution. Science *325*, 1686-1688.

Thompson, K.E., Bashor, C.J., Lim, W.A., and Keating, A.E. (2012). SYNZIP protein interaction toolbox: in vitro and in vivo specifications of heterospecific coiled-coil interaction domains. ACS Synth Biol *1*, 118-129.

Trajtenberg, F., Imelio, J.A., Machado, M.R., Larrieux, N., Marti, M.A., Obal, G., Mechaly, A.E., and Buschiazzo, A. (2016). Regulation of signaling directionality revealed by 3D snapshots of a kinase:regulator complex in action. Elife *5*.

UniProt, C. (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I*., et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Wang, B., Zhao, A., Novick, R.P., and Muir, T.W. (2014). Activation and inhibition of the receptor histidine kinase AgrC occurs through opposite helical transduction motions. Mol Cell *53*, 929-940.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol *13*, e1005324.

Wanner, B.L., and Wilmes-Riesenberg, M.R. (1992). Involvement of phosphotransacetylase, acetate kinase, and acetyl phosphate synthesis in control of the phosphate regulon in Escherichia coli. J Bacteriol *174*, 2124-2130.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A *106*, 67-72.

Whitaker, W.R., Davis, S.A., Arkin, A.P., and Dueber, J.E. (2012). Engineering robust control of two-component system phosphotransfer using modular scaffolds. Proc Natl Acad Sci U S A *109*, 18090-18095.

Willett, J.W., Tiwari, N., Muller, S., Hummels, K.R., Houtman, J.C., Fuentes, E.J., and Kirby, J.R. (2013). Specificity residues determine binding affinity for two-component signal transduction systems. MBio *4*, e00420-00413.

Wojtowicz, W.M., Flanagan, J.J., Millard, S.S., Zipursky, S.L., and Clemens, J.C. (2004). Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. Cell *118*, 619-633.

Wojtowicz, W.M., Wu, W., Andre, I., Qian, B., Baker, D., and Zipursky, S.L. (2007). A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. Cell *130*, 1134-1145.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676-680.

Zhang, Z., and Palzkill, T. (2004). Dissecting the protein-protein interface between beta-lactamase inhibitory protein and class A beta-lactamases. J Biol Chem *279*, 42860-42866.

Zimmerman, S.B., and Trach, S.O. (1991). Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. J Mol Biol *222*, 599-620.

Chapter 2 – Engineering orthogonal signaling pathways reveals the sparse occupancy of sequence

This work is in revision.

Introduction

Many promising new therapies, such as CAR-T cells (Brentjens et al., 2013) and engineered probiotics (Mimee et al., 2016; Riglar and Silver, 2018), require an ability to transfer signaling pathways into a new genomic context. Such efforts typically involve the repurposing of natural signaling pathways (Bashor et al., 2008; Dueber et al., 2007; Morsut et al., 2016; Riglar et al., 2017; Sockolosky et al., 2018). Similarly, new pathways arise during evolution through the duplication, diversification, and repurposing of existing signaling mechanisms. For both engineered and evolved signaling proteins to execute independent functions within cells, they must avoid, or minimize, detrimental cross-talk, a significant challenge for proteins with multiple paralogs that are often very similar in sequence and structure.

For signaling proteins, such as protein kinases and their substrates, specificity is enforced primarily at the amino-acid level (Creixell et al., 2015a; Skerker et al., 2005). The specificity-determining residues of signaling proteins define a finite sequence space. How cells globally organize paralogous protein families in such a sequence space and whether the specificity-determining residues of individual members have been optimally distributed to minimize cross-talk during evolution remain open questions. Work with SH3 and PDZ domains in eukaryotes suggested that paralogs are densely packed in sequence space, with intense negative selection against cross-talk leading to a global optimization of specificity (Stiffler et al., 2007; Zarrinpar et al., 2003). However, sequence space is vast, even for a limited set of residues, and nature may not have fully occupied or explored it.

To assess how crowded paralogs are in sequence space, we sought to engineer protein complexes that are functional, yet insulated from extant paralogs. If sequence space is very densely occupied by existing paralogs, it should be difficult to introduce new, insulated pathways (Fig.

2.1a). However, if sequence space is sparsely occupied, new pathways should be easy to introduce, with a low probability of cross-talk to endogenous paralogs. The design of orthogonal interacting proteins remains a major challenge. Given the difficulty of programming protein interfaces, particularly compared to nucleic acids, prior efforts have only generated 3-4 orthogonal pairs (Boyken et al., 2016; Reinke et al., 2010; Thompson et al., 2012).

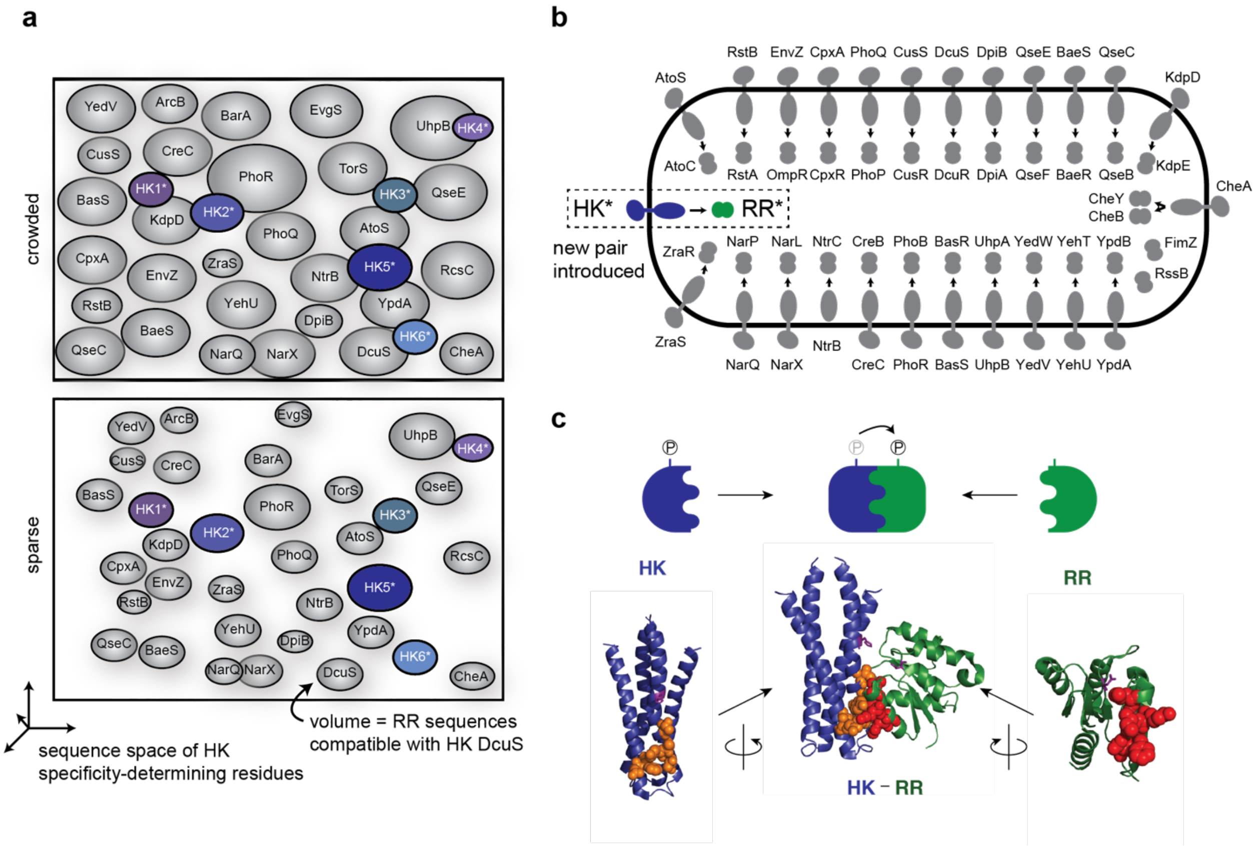


Figure 2.1 The density of paralogs in sequence space is unknown.

(a) Two models for the distribution of paralogous proteins in sequence space. For histidine kinases, each oval is a niche representing the set of substrates that a kinase can interact with given its specificity-determining residues. These niches do not overlap, reflecting the insulation of most pathways in E. coli (except NarQ and NarX). These niches could be densely packed in sequence space (top) or more sparsely distributed (bottom), making the introduction of new, insulated kinases (HK*) relatively difficult or easy, respectively. (b) Schematic summarizing the 25 canonical two-component signaling pathways in E. coli K12 that a new, orthogonal pathway must avoid cross-talk to. (c) Diagram of the DHp domain of a histidine kinase (HK), TM0853 (blue) in complex with its cognate response regulator (RR), RR0468 (green). Residues that dictate specificity and were varied in our libraries are spacefilled in orange (kinase) and red (substrate).

We focused on bacterial two-component signaling proteins, which involve a sensor histidine kinase that, upon activation, autophosphorylates and then transfers its phosphoryl group to a cognate response regulator to effect changes in cellular behavior (Fig. 2.1b-c) (Laub and Goulian, 2007; Stock et al., 2000). Most histidine kinases are bifunctional, acting as phosphatases in the absence of a signal to promote dephosphorylation of their cognate regulator. Bacteria usually encode dozens of two-component pathways (Fig. 2.1b) that are insulated from one another, with the vast majority of kinase-regulator pairs forming exclusive one-to-one relationships (Capra and Laub, 2012). Both kinase and phosphatase activities, which involve the same protein-protein interface, contribute to pathway specificity (Groban et al., 2009; Siryaporn and Goulian, 2008; Skerker et al., 2005).

Diversifying coevolving residues to engineer new signaling protein pairs

The specificity of the kinase-regulator interaction is driven by a limited set of interfacial residues in each protein that strongly coevolve (Fig. 2.1c, 2.2) (Capra et al., 2010; Podgornaia and Laub, 2015; Skerker et al., 2008). To identify combinations of these interface residues that are functional and insulated from existing two-component signaling pathways in *E. coli*, we constructed a dual library of mutants in which the key, coevolving interface residues of a canonical two-component system, PhoQ and PhoP, were randomized (Fig. 2.1c, 2.2, 2.3a-b) (Casino et al., 2009; Skerker et al., 2008). Our library, with 11 randomized positions, has a theoretical diversity of $\sim 10^{14}$. After electroporation into a Δ*phoPQ* strain of *E. coli*, individual transformants were pooled to produce a library of $\sim 5 \times 10^{8}$ variants.

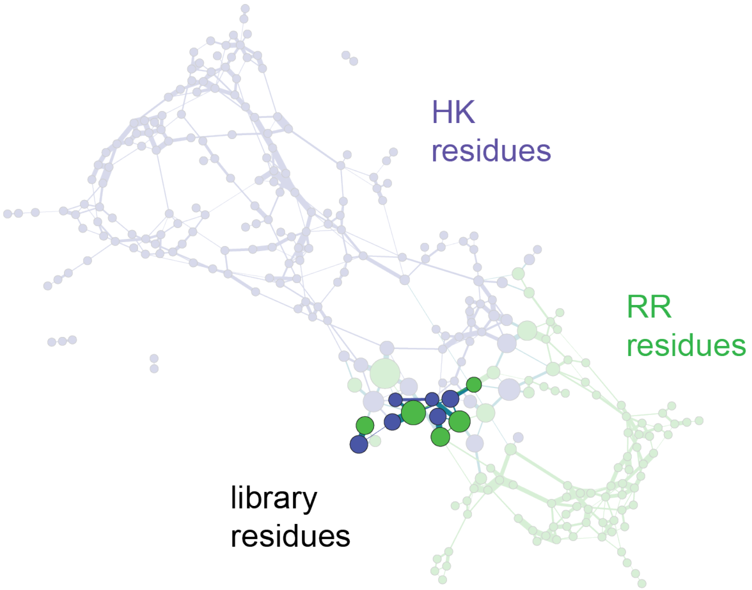


Figure 2.2 Coevolutionary relationships between PhoQ-PhoP residues

Visualization of the GREMLIN model representing the coevolutionary dependencies between the residues of cognate histidine kinases and response regulators. Blue nodes indicate PhoQ residues, green nodes indicate PhoP residues, and the darker nodes are the 11 residues randomized in the dual PhoQ-PhoP library. Edge widths indicate the strength of coevolutionary signal, while node size of each residue represents the total coevolutionary signal to residues on the other protein.

To identify functional combinations of residues, we first grew the library of PhoQ-PhoP variants overnight in medium with low Mg^{2+}, which activates PhoQ. Because cells must phosphorylate PhoP to grow when extracellular Mg^{2+} is limiting, this step enriches for functional PhoQ-PhoP variants (Fig. 2.3b). Variants that survived limiting Mg^{2+} were then subjected to Sort-seq, using fluorescence-activated cell sorting (FACS) and deep sequencing to quantify the signal responsiveness of variants in the library (Fig. 2.3b, 2.4a-f). To gauge the phosphorylation of PhoP *in vivo*, we used a fluorescent transcriptional reporter, P_{mgrB}-*yfp*. In low extracellular Mg^{2+}, functional PhoQ promotes the phosphorylation of PhoP and production of YFP, whereas in high concentrations of Mg^{2+}, PhoQ drives the dephosphorylation of PhoP, limiting YFP accumulation (Fig. 2.3a).

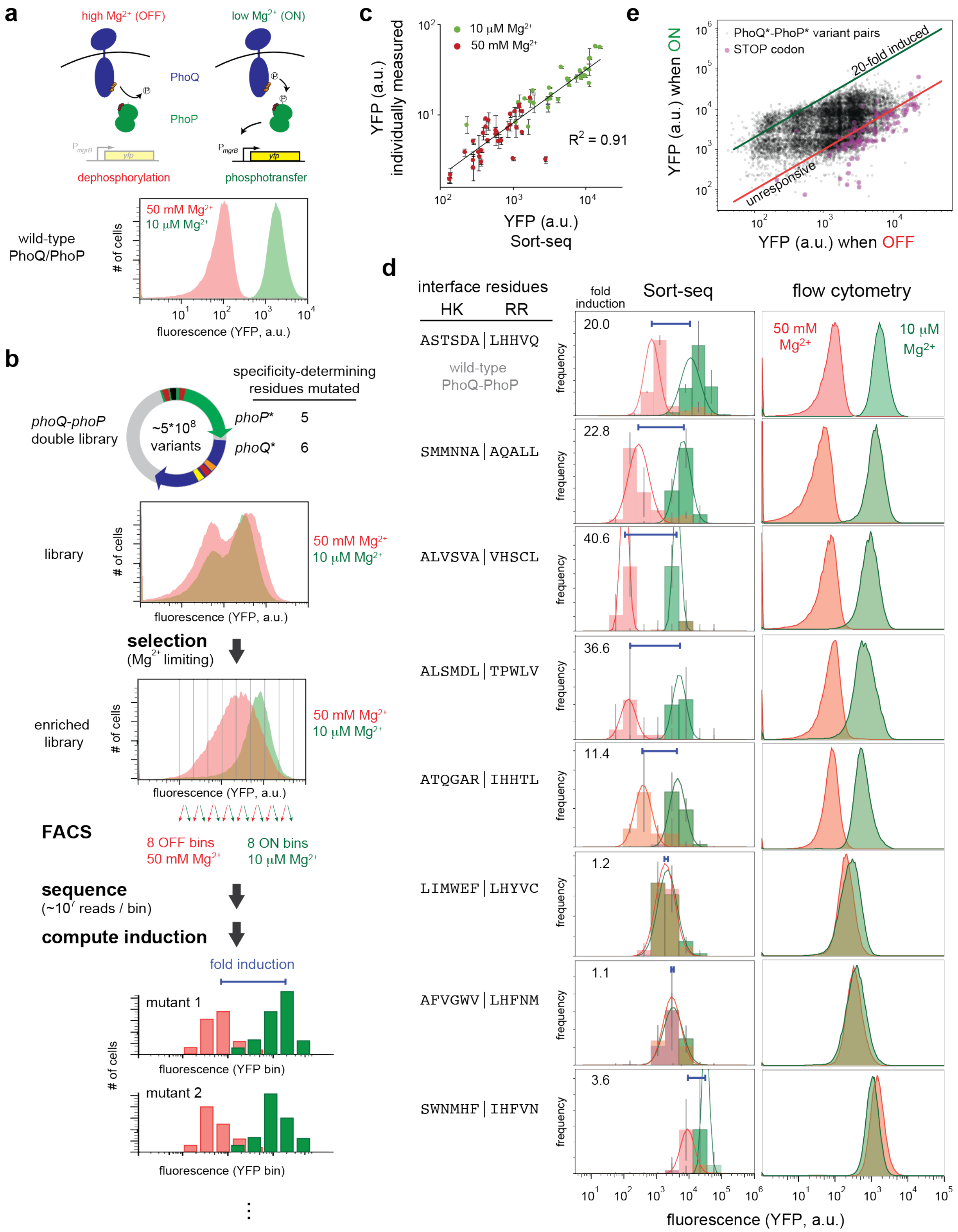
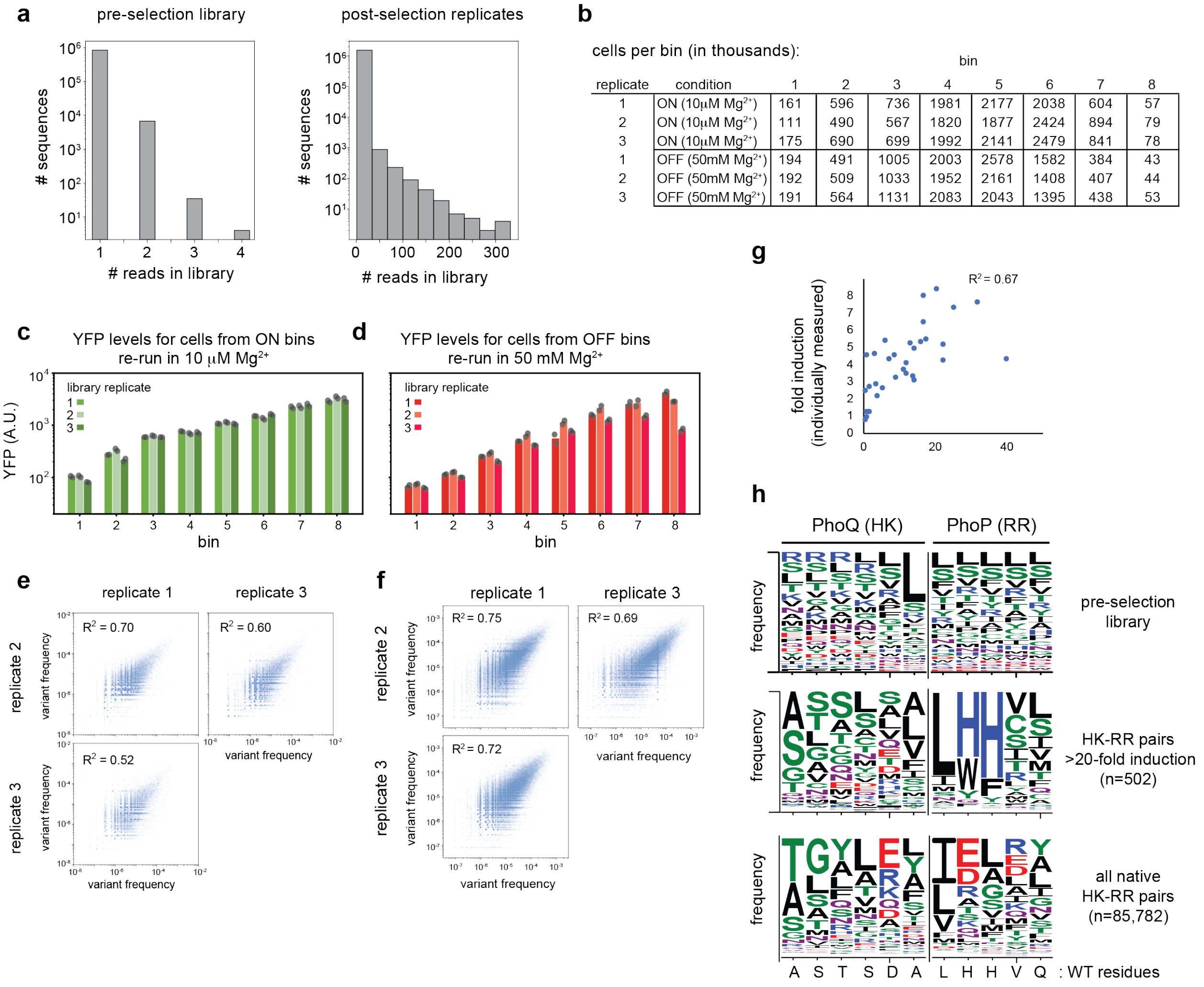


Figure 2.3 Identification of new, orthogonal signaling protein pairs.

(figure legend continued on following page)

(a) Schematic indicating that E. coli PhoQ (blue) can either phosphorylate or dephosphorylate PhoP (green), depending on extracellular Mg^{2+} levels, to stimulate or repress, respectively, gene expression, including a YFP reporter that can be quantified in individual cells by flow cytometry (bottom). (b) A library of ~5 x 10^8 PhoQ-PhoP variants, described in the text, was first examined by flow cytometry before being grown overnight with low extracellular Mg^{2+}, followed by outgrowth with high or low Mg^{2+} and then FACS, with deep sequencing of 8 consecutive, non-overlapping bins. For each variant pair in the library, the number of reads in each bin in each condition was plotted to infer the signaling profile and fold-induction value. (c) Scatterplot showing the Pearson correlation between YFP values inferred by Sort-seq and measured individually. (d) For each variant pair indicated, including the wild-type pair (top), the signaling profile inferred by Sort-seq is shown (left) in comparison to the profile measured for that pair in isolation. Fold-induction values were calculated as the difference between means of the Gaussian fits to distributions in the ON (10mM) and OFF (50 mM) states. (e) Plot of the mean YFP level inferred by Sort-seq for each variant pair in the OFF and ON states. Parallel lines indicate thresholds for pairs exhibiting > or < 20-fold induction.



b

cells per bin (in thousands):

replicate	condition	1	2	3	4	5	6	7	8
1	ON (10μM Mg^{2+})	161	596	736	1981	2177	2038	604	57
2	ON (10μM Mg^{2+})	111	490	567	1820	1877	2424	894	79
3	ON (10μM Mg^{2+})	175	690	699	1992	2141	2479	841	78
1	OFF (50mM Mg^{2+})	194	491	1005	2003	2578	1582	384	43
2	OFF (50mM Mg^{2+})	192	509	1033	1952	2161	1408	407	44
3	OFF (50mM Mg^{2+})	191	564	1131	2083	2043	1395	438	53

Figure 2.4 Summary Statistics for the dual PhoQ-PhoP library

(a) (Left) Histogram of the read counts for the pre-selection PhoQ-PhoP library. The vast majority of reads are unique, indicating that the library size is larger than Illumina sequencing coverage and no variants are over-represented. (Right) Histogram of the read counts for one replicate of the PhoQ-PhoP library after overnight growth in low Mg^{2+} conditions. (b) Counts of cells sorted into each bin, for each replicate and growth condition. (c) The cells sorted into each bin were grown overnight, diluted back to mid-exponential phase, shifted to media with 10 μM Mg^{2+} and their YFP levels verified by flow cytometry. n = 2 independent biological replicates. (d) Same as panel (c), but with cells retained in media with 50 mM Mg^{2+}. n = 2 independent biological replicates. (e) Scatterplots displaying the correlations between bin frequencies of individual variant pairs measured in independent replicates. Only 10^6 data points are shown for clarity. R^2 values indicate the Pearson correlation coefficients, calculated using all data points. (f) Same as panel (e), but displaying only the 10,595 variants with sufficient sequencing coverage and fit quality (see Methods) to be included in analysis. (g) Scatterplot displaying the Pearson correlation between YFP fold-induction measured by Sort-seq and that measured individually by flow cytometry for 32 individual variant pairs. (h) Sequence logos summarizing the amino acid frequencies at each position varied in the pre-selected library (top), set of pairs with >20-fold induction (middle), and all native HK-RR pairs (bottom). The residues found at these positions in wild-type PhoQ and PhoP are listed below.

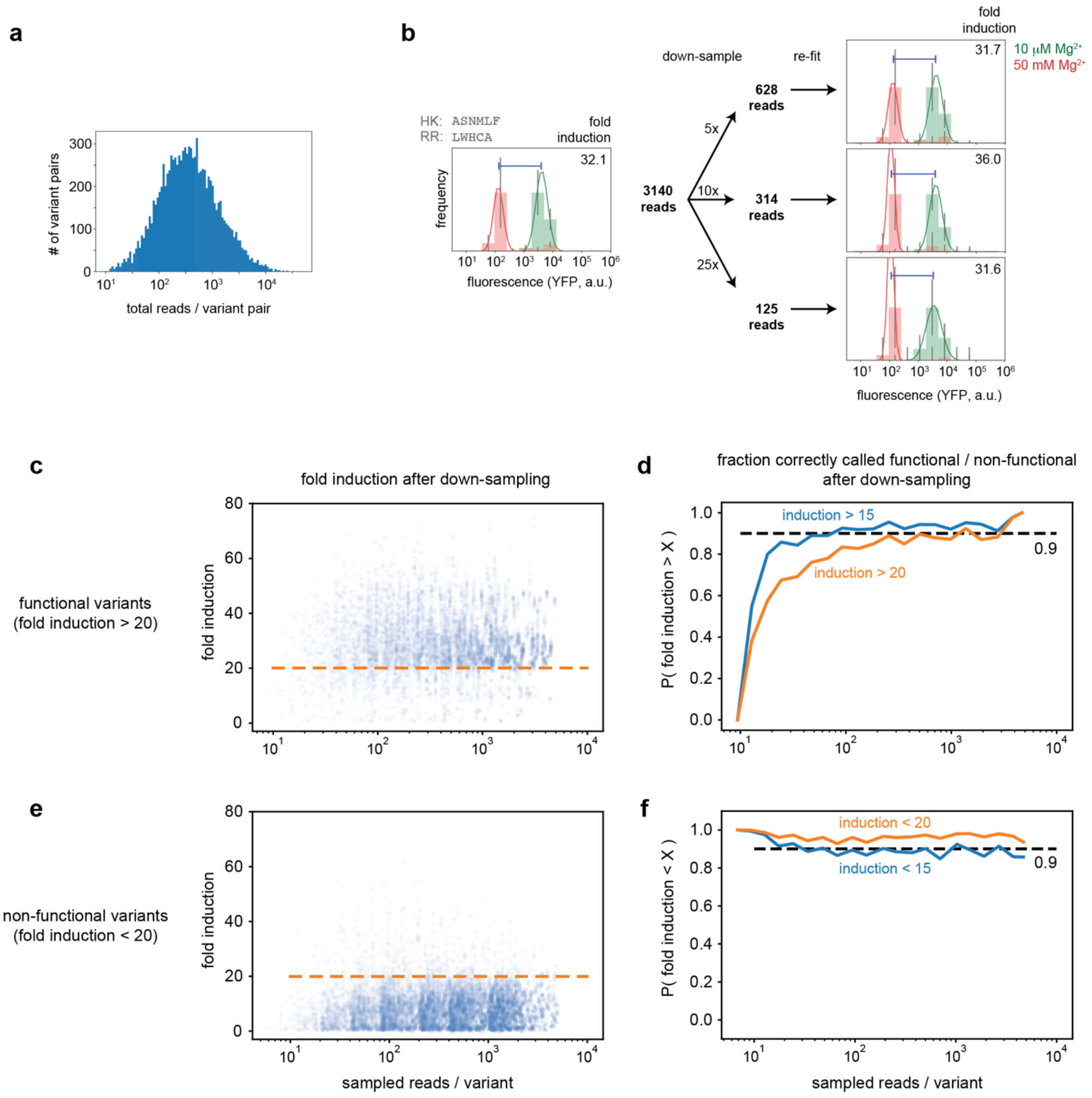


Figure 2.5 Sensitivity of Sort-seq pipeline to read coverage.

(a) Histogram of the total read counts for the 10,595 variants with sufficient coverage and fit quality to be included in analysis. (b) A schematic example of the down-sampling method used to simulate low read coverage using variants with high read coverage. (c) Fold induction of high coverage, functional PhoQ*-PhoP* variants (fold induction > 20) after simulating lower read coverage via down-sampling and re-fitting. (n = 100 independent downsampling simulations). (d) A quantification of how read coverage in (c) affects the calculated fold induction of functional variants. The fraction of functional variants that still display high fold induction at lower read coverage is plotted with respect to read coverage. (e) Same as panel (c), but for non-functional (fold induction < 20) PhoQ*-PhoP* pairs with high read coverage. (n = 100 independent downsampling simulations). (f) Same as panel (d), but for non-functional variants. The fraction of non-functional variants that still display low fold induction at lower read coverage is plotted with respect to read coverage.

To identify those variants that are signal responsive and drive YFP production specifically in low Mg^{2+}, we sorted cells from each condition into 8 separate bins and deep sequenced the randomized regions of variants in each bin (Fig. 2.3b). We then calculated the frequency of variants in each bin to yield distributions of individual variants in low and high Mg^{2+}, which were fit to Gaussians. From these fits, we assessed the mean level of YFP in each condition and the fold-induction, or signal responsiveness, of each variant.

To validate our selection scheme, we first isolated at random 48 individual clones from our starvation-enriched library and measured the distribution of YFP levels in low and high Mg^{2+} (Fig. 2.3c). Of these 48 clones, 32 had sufficient sequencing coverage and Gaussian fits for quantification by Sort-seq. Each of the individual flow cytometry profiles showed strong similarity to the distribution inferred from Sort-seq, including for variants with a fold-induction similar to wild-type and those that were constitutively ON (Fig. 2.3d). The constitutively ON behavior likely arises when a PhoQ variant lacks phosphatase activity or misfolds as PhoP can then accumulate phosphoryl groups from other sources, *e.g.* acetyl-phosphate, leading to signal-independent activity; such behavior is seen with a Δ*phoQ* strain (Podgornaia and Laub, 2015) or library variants harboring a stop codon at one of the randomized positions in *phoQ* (Fig. 2.3e).

To select variants that are signal responsive like wild-type PhoQ-PhoP, we identified combinations of residues that produced fold-induction values > 20; there were 502 such sequences, hereafter referred to as functional PhoQ*-PhoP* variants (Fig. 2.3f, 2.4h). Most of the residue combinations identified as functional shared few identities with wild-type PhoQ or PhoP, although the PhoP* variants exhibited some enrichment for wild-type residues at three positions. This enrichment likely reflects constraints imposed by residues not randomized in our library.

Nevertheless, many interfacial sequences were functional and differed substantially from wild-type PhoQ-PhoP (Fig. 2.4h, 2.6a).

Orthogonality of PhoQ*-PhoP* variants to parent proteins

We isolated and characterized 41 diverse PhoQ*-PhoP* variants that showed >20-fold induction and shared fewer than 5 identities with the wild-type PhoQ-PhoP at the 11 randomized positions (Fig. 2.6a). Although these 41 variants differ substantially from the wild-type pair, they could still cross-talk to the wild-type PhoQ-PhoP pair. Previous work demonstrated a high degree of degeneracy in PhoQ interface residues with some highly-divergent combinations still capable of interacting with wild-type PhoP (Podgornaia and Laub, 2015). To test whether the 41 selected PhoQ*-PhoP* pairs were insulated from the wild-type proteins, we built strains in which each PhoQ* and PhoP* variant was tested for interaction with a wild-type partner, *i.e.* for a given pair, we tested the signal responsiveness of PhoQ*-PhoP and PhoQ-PhoP* relative to PhoQ*-PhoP* and PhoQ-PhoP, by growing cells harboring each combination in both low and high Mg^{2+} conditions and then measuring activity of our P_{mgrB}-*yfp* reporter. For 16 of the 41 pairs tested, there was substantially higher Mg^{2+}-dependent signaling with the mutant protein pair than for either mutant paired with a wild-type partner (Fig. 2.6a); the fold-induction values for non-cognate pairings were never greater than 30% that of cognate pairs. Thus, these 16 pairs do not cross-talk with the wild-type PhoQ-PhoP proteins, indicating that our selection scheme produced functional, insulated signaling pathways, at least relative to the parent proteins.

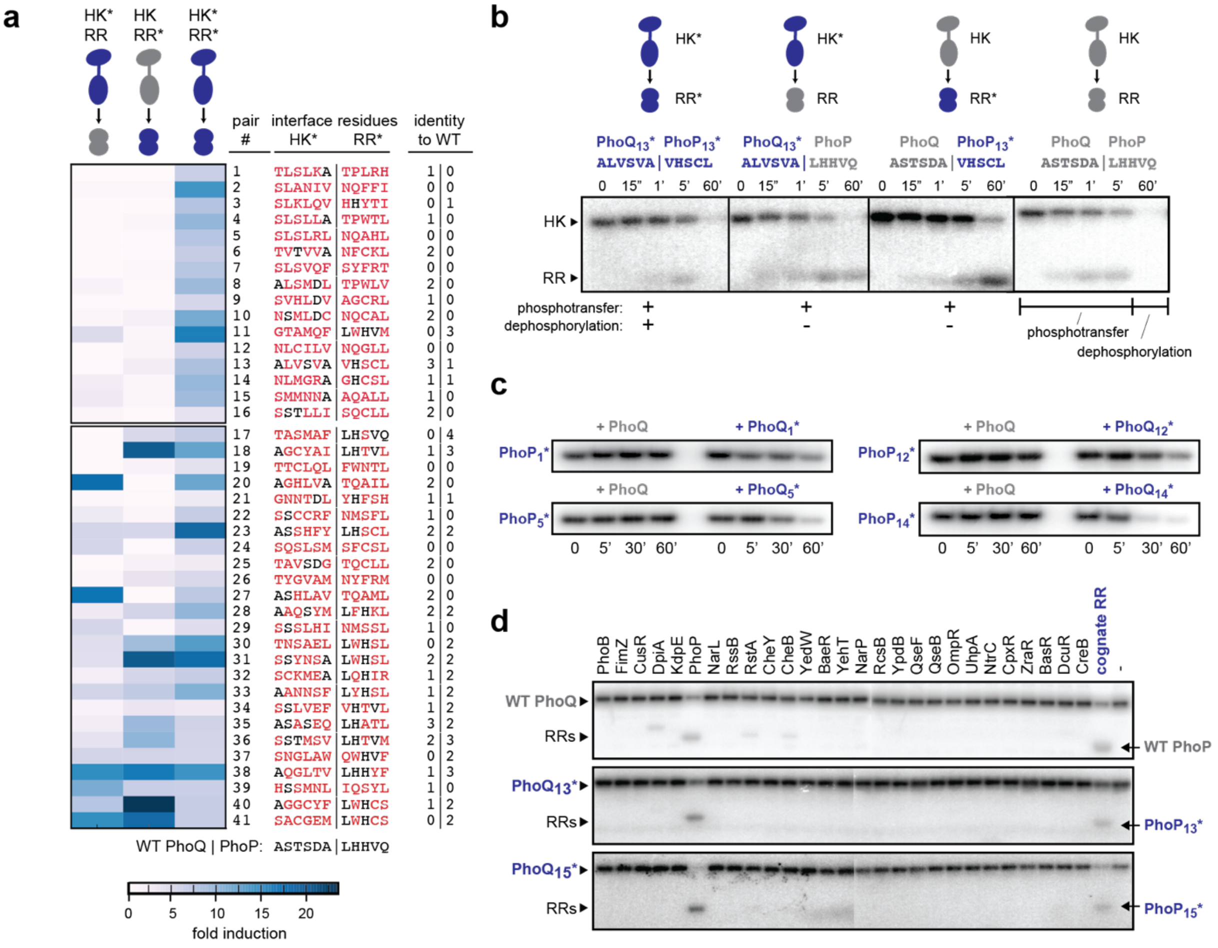


Figure 2.6 Orthogonality of selected PhoQ*-PhoP* variant pairs

(a) Fold-induction values measured by flow cytometry for PhoQ*-PhoP* pairs (right), compared to PhoQ* with wild-type PhoP (left) and wild-type PhoQ with PhoP* (middle). The interface residues of each pair are listed with the number of identities to the wild-type proteins. (b) Time-course of autophosphorylated PhoQ or $PhoQ_{13}$* incubated with PhoP or $PhoP_{13}$*. For PhoQ*-PhoP* and PhoQ-PhoP, phosphotransfer leads initially to phosphorylated PhoP and a concomitant decrease in phosphorylated PhoQ. PhoQ eventually becomes unphosphorylated and drives dephosphorylation of PhoP, yielding a blank lane. (c) PhoQ* variants with no detectable autokinase activity specifically dephosphorylate their cognate PhoP* variants. Phosphorylated PhoP* variants were incubated with their cognate PhoQ* or wild-type PhoQ for the times indicated. (d) Phosphotransfer profiles of wild-type PhoQ (top) and three PhoQ* variants. In each case, autophosphorylated kinase was incubated for 5 min. with each of 27 *E. coli* response regulators, and its selected PhoP* variant. All gels are representative of >=2 replicates.

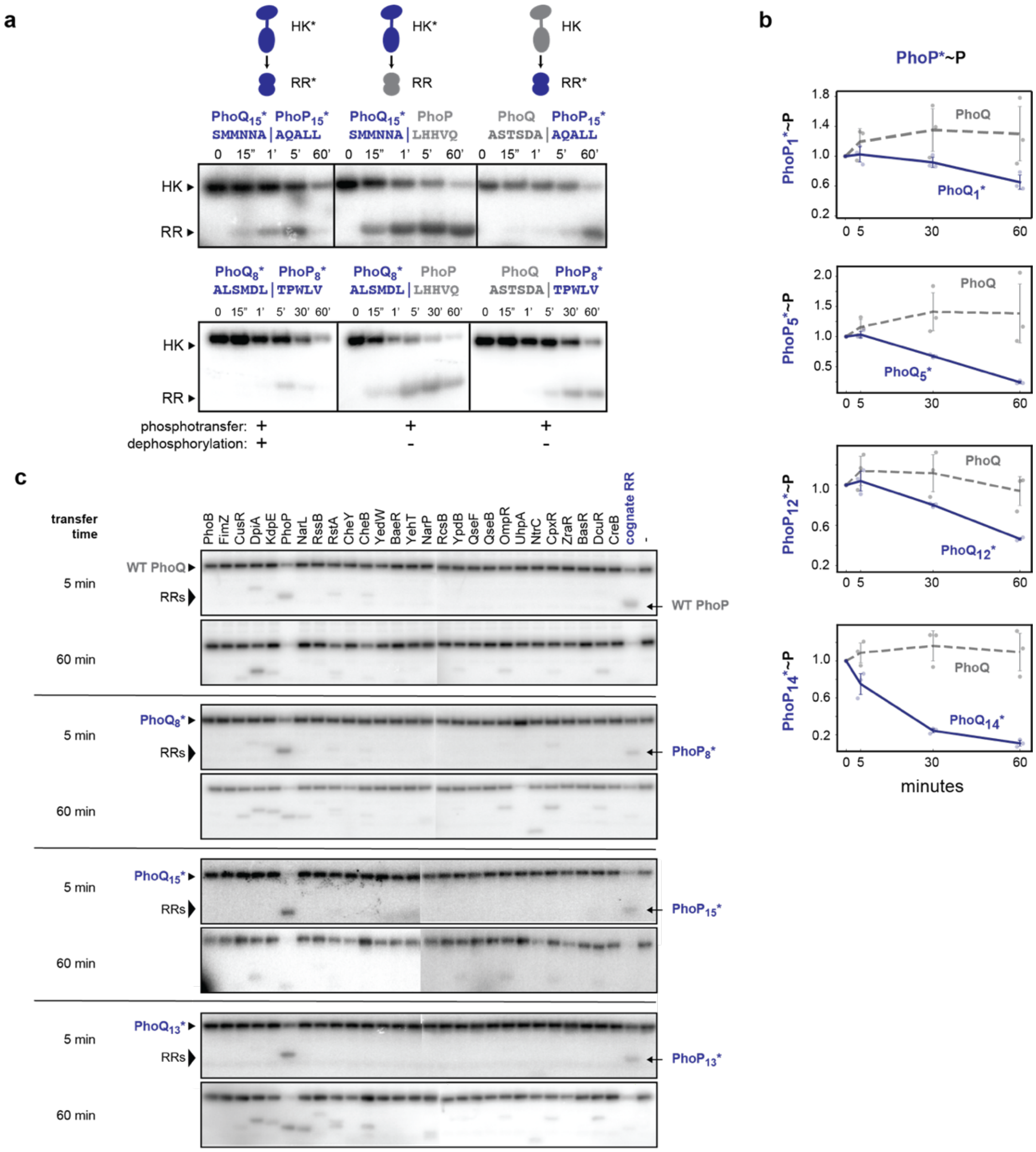


Figure 2.7 Additional in vitro characterizations of selected PhoQ*-PhoP* pairs.

(a) Phosphotransfer reactions for PhoQ*-PhoP* variants, as well as PhoQ* with wild-type PhoP (middle), and wild-type PhoQ with PhoP* (right). These experiments were repeated independently two times with similar results. (b) Quantification of phosphatase activity for PhoQ$_1$*,PhoQ$_5$*,PhoQ$_{12}$*, and PhoQ$_{14}$*, as in Fig. 2.6c. Lines indicate mean ± s.d. from n = 3 biological replicates. (c) Phosphotransfer profiles of wild-type PhoQ (top) and three PhoQ* variants, as in Fig. 2.6d, but for 60 min and 5 min. In each case, the kinase was autophosphorylated and then incubated for 5 or 60 min. individually with each of 27 response regulators from *E. coli*, and its selected PhoP* variant, followed by SDS-PAGE and autoradiography. The position of the autophosphorylated kinase and the approximate positions of any phosphorylated regulators are indicated by arrowheads on the left. These experiments were repeated independently two times with similar results.

Insulation is enforced by stringent phosphatase specificity

To examine the mechanistic basis of insulation, we purified the mutant response regulators and the cytoplasmic domains of the mutant histidine kinases for 7 of the 16 validated, insulated protein pairs. The autokinase, phosphotransfer, and phosphatase activities of a histidine kinase can be assessed using a single assay (Fig. 2.6b) (Skerker et al., 2005; Stock et al., 2000). Each kinase is autophosphorylated with [^{32}P-γ]-ATP and then mixed with a given partner, resulting in phosphotransfer and a phosphorylated response regulator; as unphosphorylated kinase accumulates, it then stimulates the dephosphorylation of a response regulator, leading to depletion of radiolabeled response regulator.

For three of the functional mutant pairs, we observed the same pattern of activities as with wild-type proteins, demonstrating that these PhoQ*-PhoP* pathways harbor functional phosphotransfer and phosphatase activities (Fig. 2.6b, 2.7a). We then tested each mutant protein with a wild-type partner. We found that the wild-type PhoQ could phosphotransfer to the PhoP* variants, but could not efficiently dephosphorylate them. Similarly, we found that the PhoQ* variants could phosphorylate wild-type PhoP, but could not dephosphorylate it. Thus, the orthogonality of these variant pairs with respect to the wild-type system is driven largely by highly-specific phosphatase activity.

For four of the functional, insulated mutant pairs purified, the kinases exhibited no detectable autokinase activity *in vitro*. However, they retained phosphatase activity and each was specific for the selected, cognate PhoP* partner relative to the wild-type PhoP (Fig. 2.6c, 2.7b). Thus, the phosphatase activity of a PhoQ* variant may be sufficient to support a functional and insulated PhoQ*-PhoP* system; phosphoryl donors such as acetyl-phosphate (or other kinases) would drive the phosphorylation of PhoP* in low Mg^{2+}, with PhoQ* phosphatase activity ensuring that PhoP* is not active in high Mg^{2+}.

Orthogonality to endogenous two-component pathways

Next, we assessed the orthogonality of functional variants with respect to the other ~30 two-component signaling pathways in *E. coli*. We purified 27 response regulators from *E. coli* and assayed, in parallel, the ability of $PhoQ_{13}$* and $PhoQ_{15}$* to phosphorylate each regulator and their partners, $PhoP_{13}$* and $PhoP_{15}$*, respectively (Fig. 2.6d). Strikingly, no phosphotransfer was detected from either kinase to any of the endogenous *E. coli* response regulators after 5 minutes except the wild-type PhoP. However, as noted above, the $PhoQP_{13}$* and $PhoQP_{15}$* pairs are insulated from the wild-type PhoQP by the highly-specific phosphatase activity of each kinase. Some phosphotransfer to other response regulators was detected after longer incubations, when many histidine kinases exhibit apparent promiscuity (Skerker et al., 2005), reflecting their homology (Fig. 2.7c). We also examined phosphotransfer from 5 native *E. coli* kinases to 11 different PhoP* variants (Fig. 2.8); in each case, the native kinase preferentially phosphorylated its cognate response regulator. Indeed, PhoP* variants were typically even more insulated than wild-type PhoP was from these non-cognate kinases. Finally, we phosphorylated 12 *E. coli* response regulators and then examined their dephosphorylation by three PhoQ* variants (Fig. 2.9); in each case, PhoQ* robustly dephosphorylated its cognate PhoP*, but not the native response regulators. Taken all together, these results indicate that the functional PhoQ*-PhoP* variants identified are insulated from all native *E. coli* pathways.

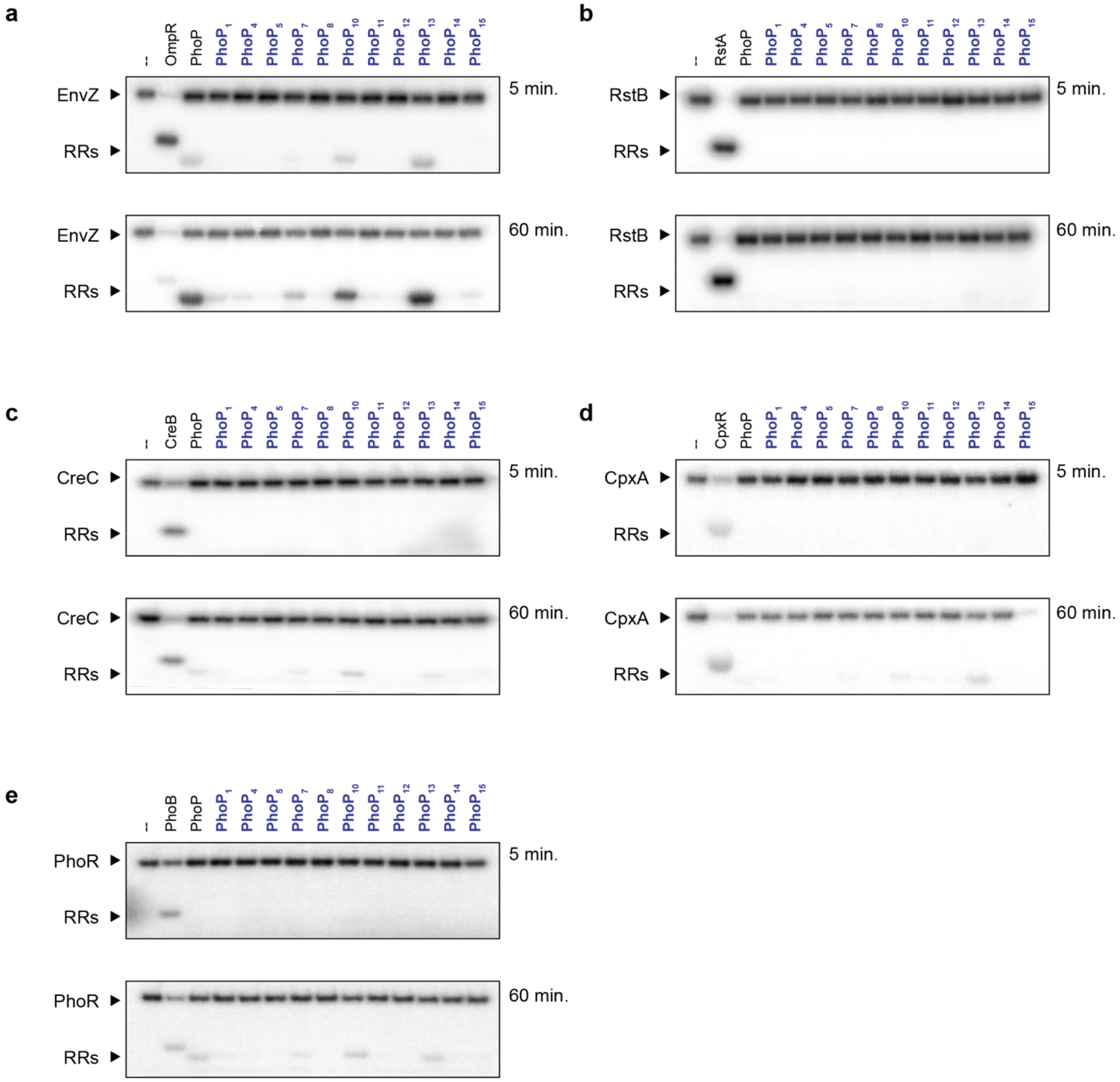


Figure 2.8 Insulation of E. coli histidine kinases from PhoP* variants.

Phosphotransfer profiles of five histidine kinases endogenous to *E. coli*: EnvZ (a) , RstB (b) , CreC (c) , CpxA (d) , and PhoR (e). In each case, the kinase was autophosphorylated, then incubated for 5 or 60 min. with its cognate response regulator, with wild-type PhoP, or with one of eleven PhoP* variants, and analyzed as in Fig. 2.7.

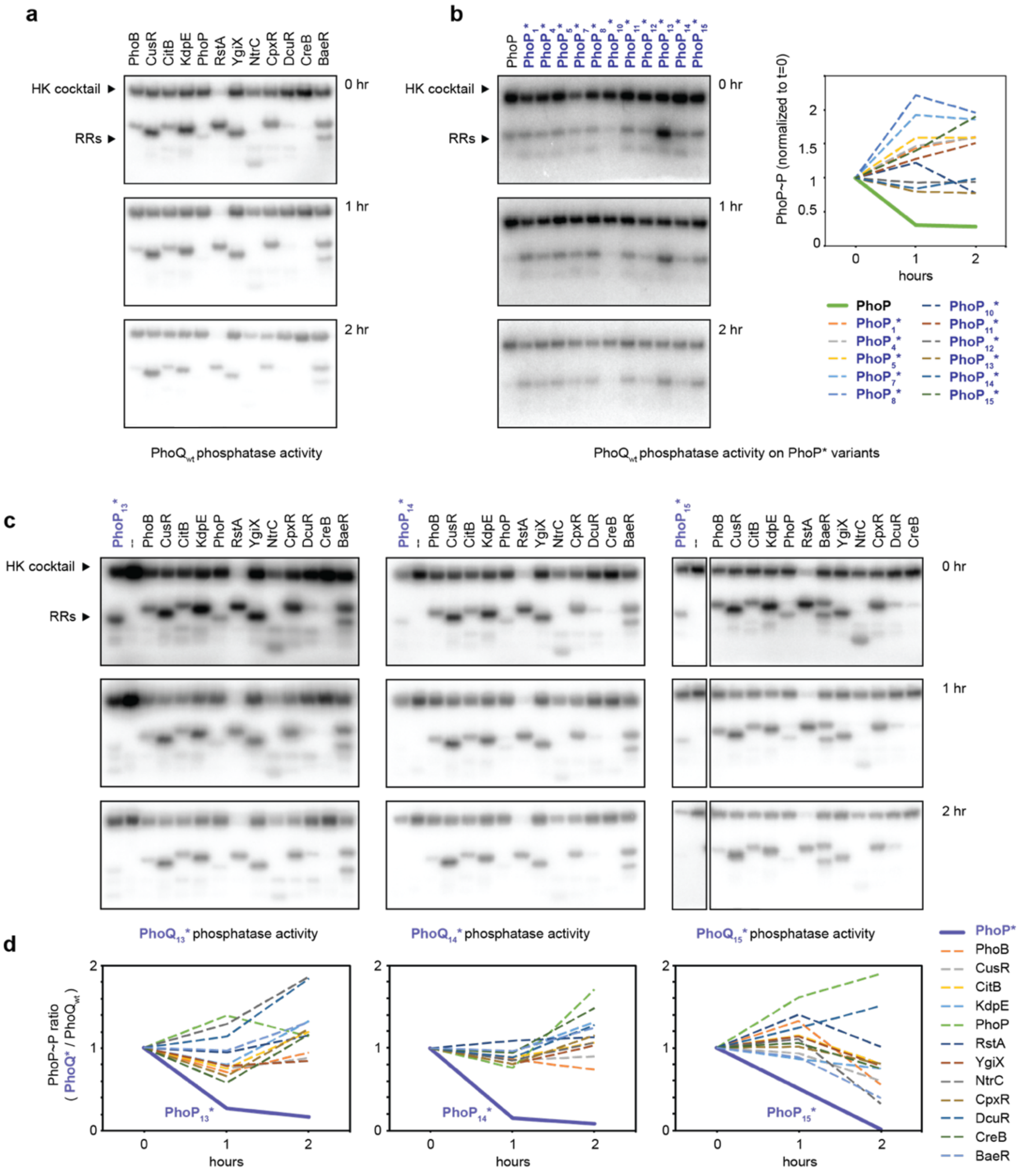


Figure 2.9 Insulation of PhoQ* variants from E. coli response regulators, with respect to phosphatase activity.

(a) Phosphatase activity of PhoQ was assessed by measuring the decay of phosphorylated response regulators. Twelve *E. coli* response regulators were selected for their ability to be stably phosphorylated in vitro by a cocktail of six *E. coli* histidine kinases (CreC, RstA, PhoR, PhoP, EnvZ and CpxA, each at 250 nM). After 2 hours of pre-incubation with radiolabeled ATP and this kinase cocktail, each regulator was combined with 2 mM PhoQ and phosphorylation state of the regulators was measured after 0, 60, and 120 minutes. (b) Phosphatase profiles conducted as in (a) for PhoP* variants. Quantification of wild-type PhoP and PhoP* variant phosphorylation (normalized to t = 0 to display decay) is plotted on the right. (c) Phosphatase profiles conducted as in (a) for PhoQ* variants. (d) Ratio of response regulator phosphorylation between phosphatase profiles with PhoQ* variants (c) and wild-type PhoQ (a,b).

To further test the global insulation of selected PhoQ*-PhoP* variants, we used RNA-Seq to examine gene expression in strains carrying one of six different PhoQ*-PhoP* variant pairs. In each case, cells were grown with excess or limiting extracellular Mg^{2+} to repress or stimulate PhoQ, respectively, before harvesting RNA. Each system produced a similar induction of known PhoP-dependent genes (Fig. 2.10a). To assess whether a variant PhoQ* cross-phosphorylated other response regulators, we took advantage of the fact that, when active, most response regulators autoregulate and promote expression of themselves and their cognate histidine kinase(Capra and Laub, 2012). Notably, none of the six strains tested showed significant induction of other two-component systems relative to a wild-type control (Fig. 2.10b-d). These RNA-Seq analyses further indicate that the PhoQ*-PhoP* systems identified are globally insulated from other pathways. Thus, there are unoccupied regions of sequence space where new systems with novel interaction specificities can be introduced without producing cross-talk to existing systems.

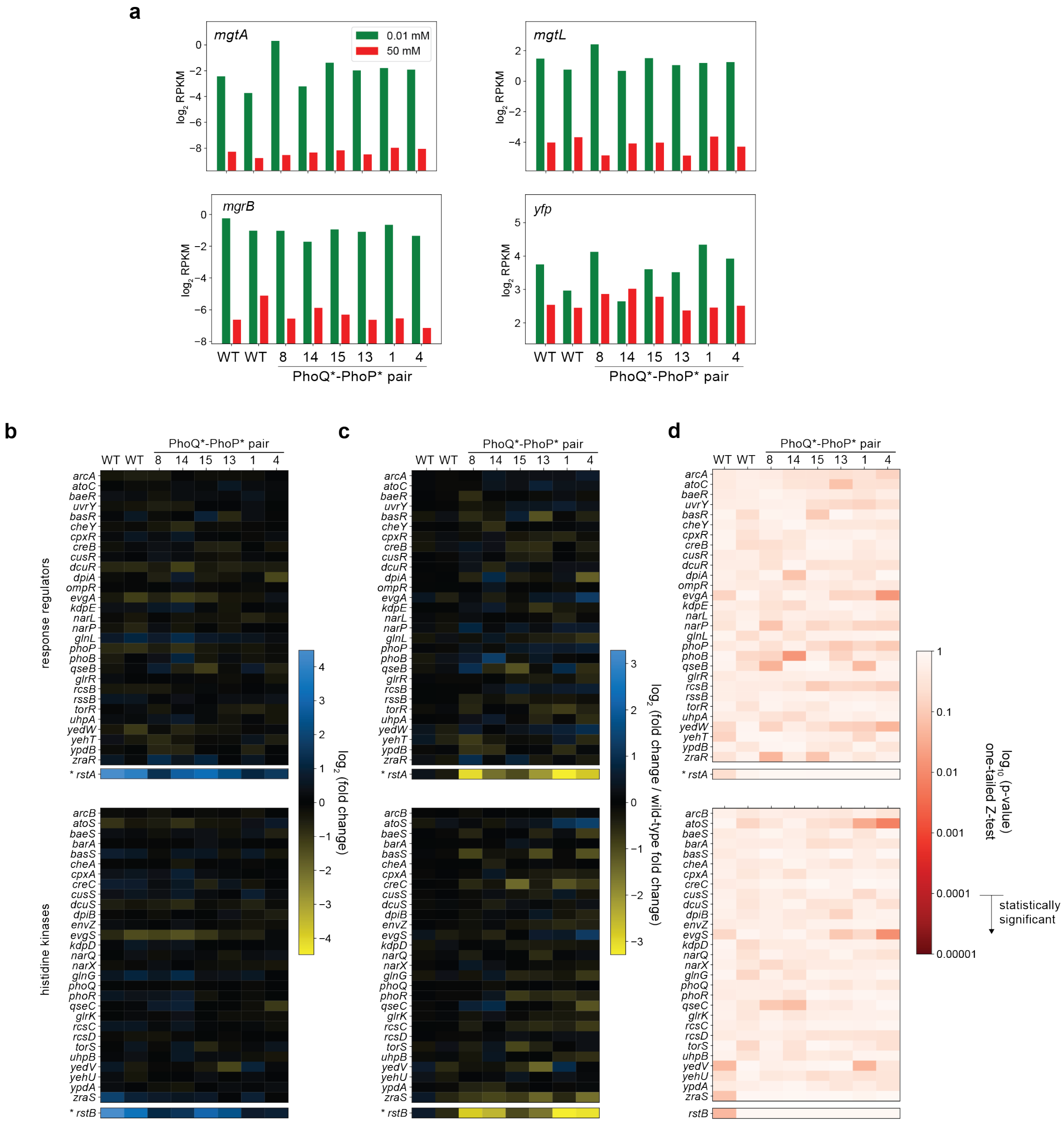


Figure 2.10 RNA-seq analysis of strains harboring PhoQ*-PhoP* variants

(a) RNA-seq analysis of strains containing wild-type PhoQ-PhoP or the indicated variant pair, measured after 30 min. with 10 µM or 50 mM Mg^{2+}. Each strain displays a similar Mg^{2+}-limitation induction of three genes (*mgtA*, *mgtL*, and *mgrB*) in the PhoP regulon, as well as the PhoP-dependent reporter gene *yfp*. (b) The expression change of each response regulator and histidine kinase gene in *E. coli* with colors representing the log2 fold change in response to low extracellular Mg^{2+}. Note that *rstAB* are part of the PhoP regulon and so show changes in transcription following activation of PhoQ and several PhoQ* variants. Otherwise, most two-component pathways show little induction by the wild-type PhoQ-PhoP and the PhoQ*-PhoP* pairs. (c) The same data as in (b) but with fold change of each variant pair normalized to the fold change seen with wild-type PhoQ-PhoP, where the latter is the geometric mean of two wild-type replicates. (d) *p* values of the Z-score calculated for each value in (c). For each gene and each variant, Z-scores represent the deviation of that gene's variant/wild-type ratio when compared to the distribution of every gene's variant/wild-type ratio. Using all *E. coli* genes with reads across multiple samples (n = 3477), *p* values were calculated with a one-tailed Z-test to identify genes induced more strongly with the variant pairs than with the wild-type pair. The statistical significance threshold after correcting for multiple hypothesis testing is indicated on the color legend encoding *p* values. None of the other two-component signaling genes in *E. coli* are significantly induced by the variant PhoQ*-PhoP* pairs tested.

Diversity of specificity amongst PhoQ*-PhoP* variants

We also wanted to test the insulation of our selected, functional PhoQ*-PhoP* variants with respect to each other. To this end, we selected 79 variant pairs (79 unique kinases, 71 unique response regulators) which had high fold-induction values and broad sequence diversity. We then combinatorically combined these PhoQ* and PhoP* variants, producing a library with a theoretical diversity of 5,609. This library was transformed into cells harboring the P_{mgrB}-*yfp* reporter and subjected to Sort-seq, as before (Fig. 2.3b, 2.12a-b), allowing us to infer the fold-induction of each PhoQ*-PhoP* combination. Within the resulting interaction matrix (Fig. 2.11a), 58 variant pairs were orthogonal to the wild-type proteins (Fig. 2.12c).

Design of mutually orthogonal signaling pathways

To isolate orthogonal sets of PhoQ*-PhoP* variants, we searched the 79x71 interaction matrix for sub-matrices in which strong interactions were seen only along the diagonal. With fold-induction thresholds of >15 for cognate pairs and <6 for all other combinations, we isolated more than 2,500 unique sets of 5 orthogonal signaling pairs and dozens of sets of size 6 (Fig. 2.11b-c, 2.12d-e). With a slightly relaxed threshold for non-cognate interactions, we found sets with up to

9 orthogonal protein pairs (Fig. 2.11d). To verify the orthogonal nature of these sets of PhoQ*-PhoP* variants, we cloned and analyzed the 25 individual pairs comprising a specific 5x5 matrix. Flow cytometry analysis showed strong agreement with the fold-induction values inferred by Sort-seq (Fig. 2.11c,e).

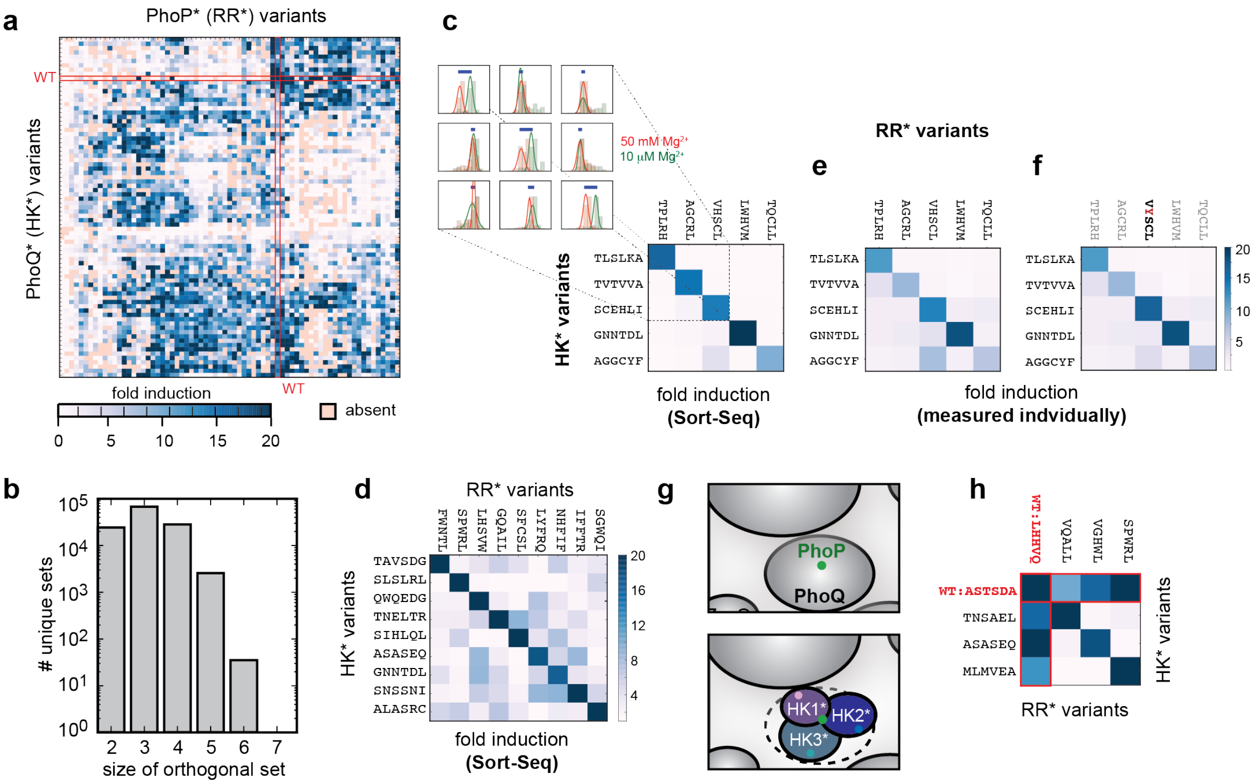
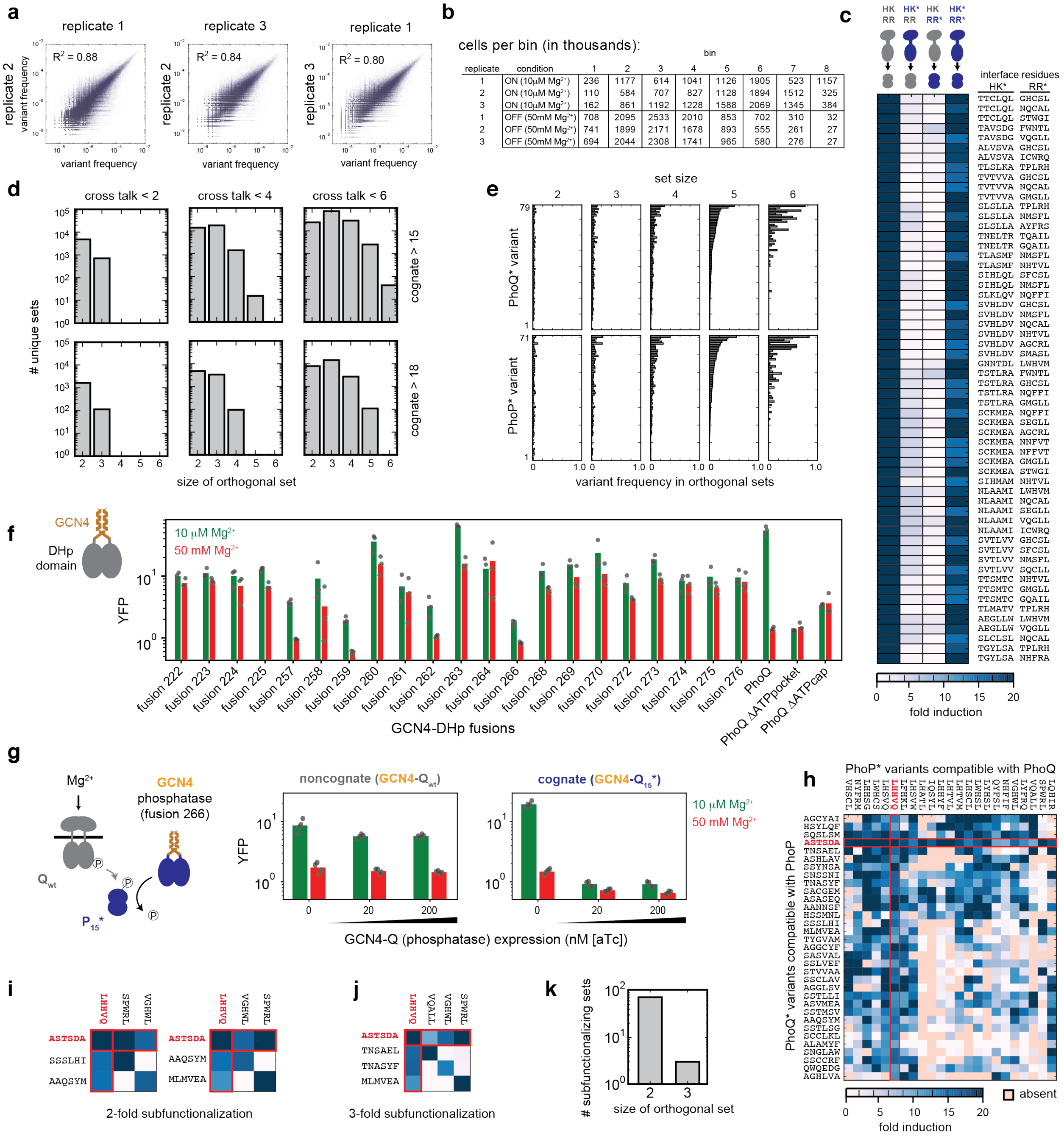


Figure 2.11 Identification of sets of orthogonal signaling proteins.

(a) Fold-induction values for all possible pairings from a set of 79 PhoQ*-PhoP* variants. The matrix was clustered in both dimensions. (b) Number of unique sets of various sizes of orthogonal PhoQ*-PhoP* pairs, with fold-induction values >15 for cognate pairs and <6 for all non-cognate pairs. (c) A set of 5 PhoQ* and PhoP* variants that are functional and mutually orthogonal. (d) A set of 9 PhoQ* and PhoP* variants that are functional and mutually orthogonal with fold-induction values >17 for cognate pairs and <10 for non-cognate pairs. (e) Fold-induction values for the mutant combinations in (c) measured individually. (f) Same as (e) but with a point mutant, PhoP* VHSCL to VYSCL, that reduces cross-talk. (g) Model for subfunctionalization of PhoQ in sequence space. (h) Subfunctionalization of PhoQ specificity. A set of three PhoQ*-PhoP* variants that are mutually insulated, but retain interactions with the parent proteins.



cells per bin (in thousands):

replicate	condition	bin 1	2	3	4	5	6	7	8
1	ON (10μM Mg^{2+})	236	1177	614	1041	1126	1905	523	1157
2	ON (10μM Mg^{2+})	110	584	707	827	1128	1894	1512	325
3	ON (10μM Mg^{2+})	162	861	1192	1228	1588	2069	1345	384
1	OFF (50mM Mg^{2+})	708	2095	2533	2010	853	702	310	32
2	OFF (50mM Mg^{2+})	741	1899	2171	1678	893	555	261	27
3	OFF (50mM Mg^{2+})	694	2044	2308	1741	965	580	276	27

Figure 2.12 Additional characterization of orthogonal sets of PhoQ*-PhoP* variant pairs.

(a) Reproducibility of replicates for the combinatorial library in Fig. 2.11a. Correlations between bin frequencies of individual variant pairs measured in independent replicates. R^2 values indicate the Pearson correlation coefficients, calculated using all data points (n = 210,319 variant pairs). (b) Counts of cells sorted into each bin, for each replicate and growth condition. (c) Functional PhoQ*-PhoP* variants that are orthogonal to wild-type PhoQ and PhoP. The fold-induction values, taken from the matrix in Fig. 2.11a, measured by Sort-seq for the variant pairs indicated in each row, either together (far right column of the heat map) or with a wild-type protein (middle two columns), compared to the wild-type pair (far left column). (d) Number of unique sets of various sizes of orthogonal PhoQ*-PhoP* pairs, for various thresholds of activity and cross-talk. (e) Frequency of each PhoQ* (top) or PhoP* (bottom) variant within the orthogonal sets of various sizes (fold induction > 15, crosstalk < 6). (f) Phosphatase- and kinase-locked variants of PhoQ were identified by fusing the catalytic DHp-CA domains to the leucine zipper GCN4 at different fusion sites (see Methods). (g) Phosphatase-locked PhoQ$_{15}$* is sufficient to suppress nonspecific phosphotransfer from wild-type PhoQ to PhoP$_{15}$*. (h) Heat map as in Fig. 2.11a, but restricted to variants that retain an interaction (fold induction > 10) with wild-type PhoQ and PhoP, which are shown in red. (i-j) Orthogonal sets of PhoQ* and PhoP* variants that, like the set in Fig. 2.11h, comprise exclusively proteins that retain interactions with the parent PhoQ and PhoP. (k) Number of unique sets of various sizes of orthogonal PhoQ*-PhoP* pairs in which all variants retain an interaction (fold induction > 10) with wild-type PhoP and PhoQ.

Notably, the non-cognate pairs in Fig. 2.11 were measured in the absence of each variant's cognate partner. Any weak cross-talk seen should be eliminated by the phosphatase activity of the cognate kinasev(Groban et al., 2009; Siryaporn and Goulian, 2008), if present; this prediction was confirmed for one instance of a PhoP* variant that exhibited modest cross-talk from wild-type PhoQ unless its cognate PhoQ* was also expressed (Fig. 2.12f-g). Thus, the off-diagonal values seen with orthogonal sets in Fig. 2.11b-e represent upper limits on cross-talk. This small degree of cross-talk is also easily reduced further. For example, with the set in Fig. 2.11e, we screened point mutations of PhoP* (VHSCL) for reduced cross-talk, finding that PhoP* (V**Y**SCL) had reduced interaction with the non-cognate kinase PhoQ* (SCEHLI) while maintaining interaction with its cognate kinase PhoQ* (SCEHLI) (Fig. 2.11f).

The 79x71 matrix of interactions (Fig. 2.11a) also offered insight into the ease with which new pathways can arise through sub-functionalization, the partitioning of a niche in sequence space rather than movement into a new region (Fig. 2.11g). For example, we found three pairs of PhoQ*-PhoP* variants that were insulated from each other, but with each exhibiting substantial cross-talk to the parental, wild-type proteins (Fig. 2.11h, 2.12h-k). Thus, these pairs have effectively partitioned the niche of wild-type PhoQ (and PhoP) in sequence space, yielding three insulated pathways. This partitioning is akin to the sub-functionalization of duplicate proteins derived from a promiscuous ancestor, and may be a common mechanism by which insulated paralogs arise during evolution.

Collectively, our results indicate that the sequence space of paralogous two-component signaling pathways in *E. coli* is relatively sparsely occupied, such that new, orthogonal signaling pathways can readily be introduced. The 502 functional variant pairs isolated here have few specificity residues in common with wild-type PhoQ-PhoP, with each other, or with extant two-component signaling proteins (Fig. 2.13a). A force-directed graph based on similarity of the specificity residues highlights the diversity of naturally-occurring interfaces and the variants we isolated (Fig. 2.13b). To estimate how easily a new, insulated pathway can be introduced, we noted that 502 functional pairs came from 10,595 pairs with fitted distributions in low and high Mg^{2+} with ~40% estimated to be insulated from the wild-type PhoQ-PhoP (Fig. 2.6a), suggesting that ~200 (or 1.6%) of the 10,595 are likely functional and insulated. This frequency is an upper bound as the 10,595 pairs with fitted distributions arose from low Mg^{2+} selection (Fig. 2.3e), which enriches ~100-fold for functionality (Fig. 2.4b). Nevertheless, given the size of sequence space, these estimates underscore the relative ease of creating kinase-substrate pairs that are functional and orthogonal to their parent proteins.

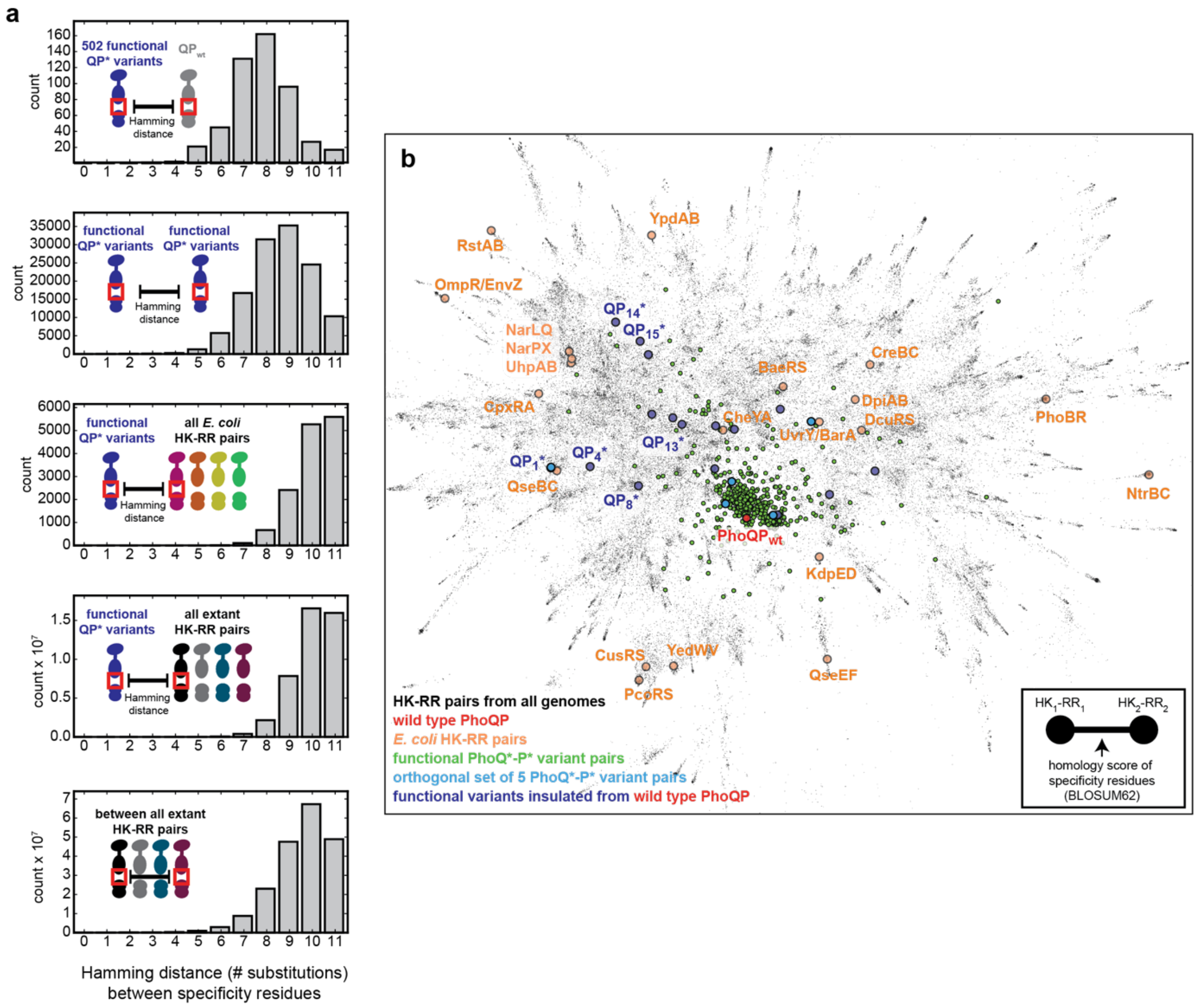


Figure 2.13 Specificity residues of novel PhoQ*-PhoP* pathways are distinct from all extant two-component signaling interfaces.

(a) Hamming distance was calculated between the 11 specificity residues of each PhoQ*-PhoP* variant pair and the 11 specificity residues of wild-type PhoQ-PhoP, the two-component signaling paralogs in *E. coli*, or all extant two-component signaling proteins. When comparing two sets, all distances between all members of both sets are plotted. (b) Force-directed graph representing the sequence space of histidine kinases. Each node represents a single histidine kinase, with the relative positions reflecting the similarity of their interface residues (see Methods). Colored nodes highlight specific sets of kinases as indicated in the legend.

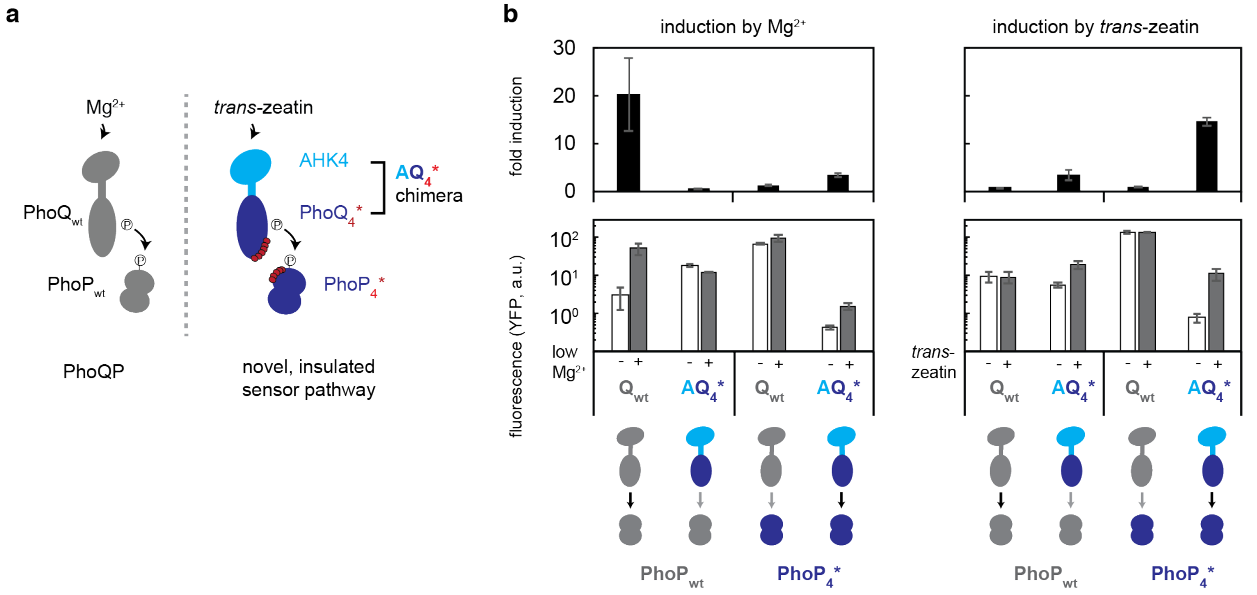


Figure 2.14 Construction of an insulated sensor.

(a) An insulated cytokinin (*trans*-zeatin) sensor constructed by fusing the sensory domain of *Arabidopsis thaliana* AHK4 to PhoQ$_4$*. (b) Chimeric sensor AQ$_4$* is specifically responsive to 1 µM *trans*-zeatin, phosphorylating its cognate mutant PhoP$_4$* to activate a YFP reporter. Wild-type PhoQ responds only to Mg^{2+} and not to the cytokinin. Bars indicate mean from n=3 biological replicates.

Utilizing orthogonal PhoQ*-PhoP* domains to insulate an *Arabidopsis* two-component sensor

Orthogonal signaling pathways will be useful in generating synthetic sensors and novel regulatory systems. As an example, we sought to generate a new pathway in *E. coli* that responds to the cytokinin *trans*-zeatin, a plant hormone. The histidine kinase AHK4 from *Arabidopsis thaliana* senses *trans*-zeatin, but cross-talks extensively with at least one native two-component pathway in *E. coli* (Suzuki et al., 2001; Yamada et al., 2001). To overcome this limitation, we fused the AHK4 sensory domain to the kinase domains of an orthogonal PhoQ* and expressed this construct in *E. coli* along with the cognate PhoP* *(*Fig. 2.14a). This engineered sensor kinase enabled *E. coli* to respond specifically to *trans*-zeatin (Fig. 2.14b) and was insulated from the native two-component pathways in *E. coli*, as measured *in vitro* by phosphotransfer profiling and *in vivo* by RNA-seq (Fig. 2.15). Thus, this chimeric sensor kinase and its cognate PhoP* expand the sensory repertoire of *E. coli* without introducing undesirable cross-talk.

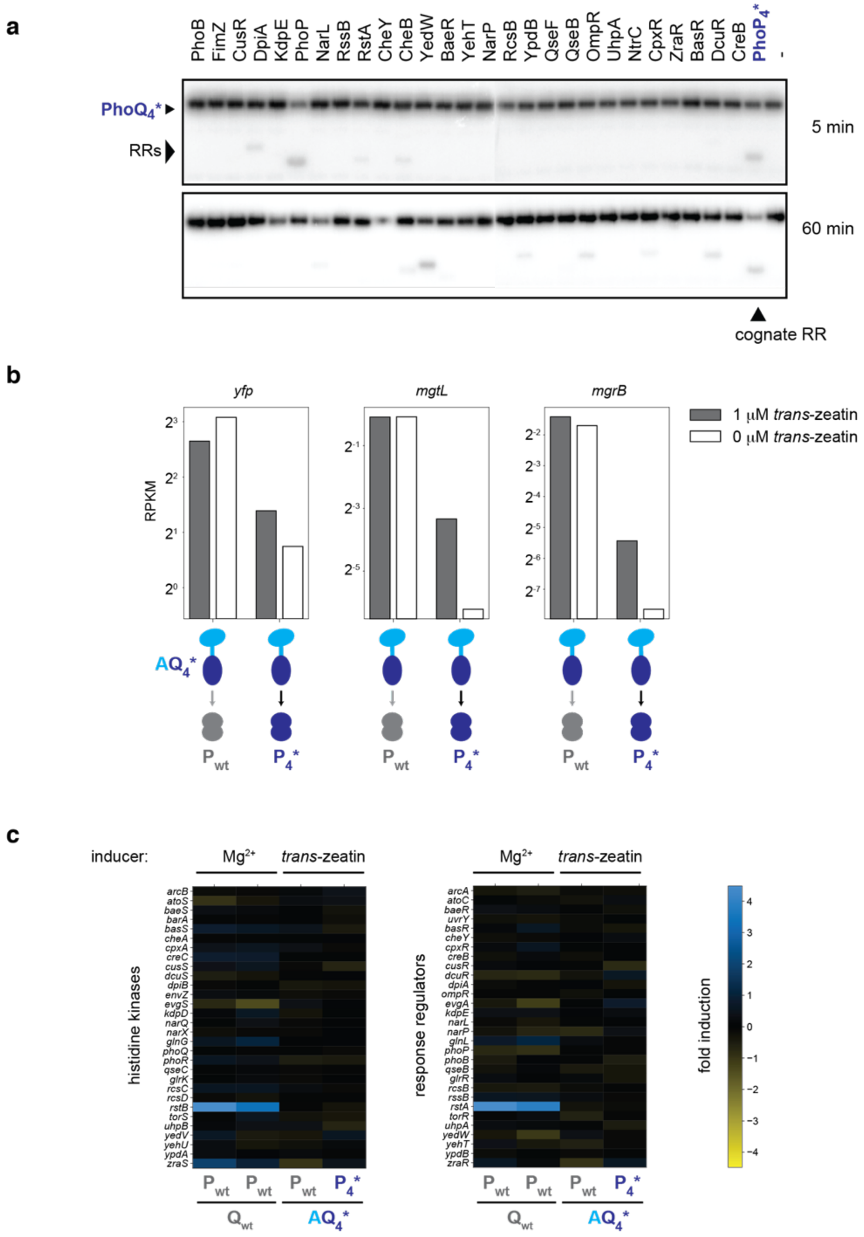


Figure 2.15 Global insulation of the AQ4*-P4* chimeric signaling pathway

(a) Phosphotransfer profile of PhoQ$_4$*. PhoQ$_4$* was autophosphorylated and then incubated for 5 and 60 min. with each of 27 response regulators from *E. coli* and with PhoP$_4$*, as in Fig. 2.7c. This experiment was repeated independently two times with similar results. (b-c) RNA-seq analysis, as in Fig. 2.10b-c, of strains expressing AQ$_4$* and either wild-type PhoP or PhoP$_4$* (measured after 30 min. induction with 0 mM or 1 mM *trans*-zeatin). (b) PhoP regulated genes *yfp*, *mgtL*, and *mgrB* are induced by *trans*-zeatin only when AQ_4* is paired with $PhoP_4$*. (c) The expression change of all response regulators and histidine kinases (fold change in response to 10 mM Mg^{2+} or 1 mM *trans*-zeatin).

Discussion

In sum, our work supports a model in which sequence space is not fully or even densely occupied. The relatively sparse distribution of extant proteins in sequence space presumably reflects their evolutionary history. Duplicated signaling proteins are likely to remain somewhat close in sequence space; they are under pressure immediately post-duplication to change and become insulated, but subsequent movement in sequence space would then arise only from neutral changes. Indeed, a prior study indicated that pathway duplication triggers an initial burst of changes in specificity residues followed by relatively few changes in the two paralogs (Capra et al., 2012). Although duplicated proteins are initially subject to selection against cross-talk with each other, each protein is likely not subject to system-wide negative selection or global optimization, as suggested previously (Stiffler et al., 2007; Zarrinpar et al., 2003).

The relatively sparse distribution of paralogs in sequence space means that new pathways can readily be introduced during evolution by duplication or lateral transfer. Additionally, this sparsity allows the construction of synthetic signaling circuits that operate independent of the host machinery. Synthetic circuits have, to date, been mainly built from nucleic acid components because of their intrinsic modularity and programmability (Nielsen et al., 2016). Protein-based circuits offer faster response times, broader options for circuit connectivity, and richer functionality, but require more complicated programming of protein interactions. Our work enables the design of two-component signaling-based circuits that can be deployed in bacteria or eukaryotes. Additionally, our work highlights the power of using coevolution-guided mutant libraries and deep sequencing to map sequence spaces relevant to protein-protein interactions, enabling the generation of customized signaling circuits insulated from endogenous pathways.

Methods

Bacterial strains and media

Escherichia coli strains were grown in M9 media (1x M9 salts, 100 μM $CaCl_2$, 0.2% glucose, 0.1% casamino acids and $MgSO_4$ at indicated concentrations). When indicated, antibiotics were used at the following concentrations: carbenicillin, 50 μg/mL; kanamycin, 50 μg/mL; spectinomycin, 50 μg/mL, chloramphenicol, 32 μg/mL.

The base strain for all studies was *E. coli* strain TIM171 (MG1655 Δ*phoPQ*Δ*lacZYA* attλ::[P_{mgrB}-*yfp*] attHK::[PtetA-cfp+]) with a ColE1/amp^R plasmid (pCM150) containing P_{mgrB}-*yfp*. All libraries were cloned onto a low copy pSC101/spec^R plasmid (pCM099, a derivative of pLPQ2), where *phoPQ* was driven by a constitutive *lacUV5* promoter. We also introduced a bicistron RBS (BCD18) upstream of *phoPQ(Mutalik et al., 2013)*, which leads to expression of a single transcript encoding a small (17 a.a.) ORF followed by an independent ribosome binding site and then *phoPQ*. This configuration ensures that mutations near the 5' end of the *phoP* coding region do not substantially affect expression by changing interactions between the 5' end of *phoP* and the upstream leader sequence. Expression from the *lacUV5* promoter on a plasmid likely produces more PhoQ-PhoP than natively produced, increasing the chance of cross-talk; thus, the variant pairs identified as orthogonal would perform even better with respect to cross-talk at lower expression levels. Additional characterizations of PhoQ* and PhoP* variants isolated from the library were done with a three-plasmid setup: reporter plasmid pCM150, pCM143 (*lacUV5-BCD18-phoP*, pSC101/spec^R), and pCM149 (*lacUV5-RBS_B0034-phoQ*, p15A/kan^R). Point mutations were introduced using blunt end ligation(Ashenberg et al., 2013) and Gibson assembly(Gibson et al., 2009).

Flow cytometry characterization

To induce PhoPQ, cells were grown to mid-exponential phase (OD_{600} ~ 0.5) in M9 before being washed once with M9 containing 0 mM $MgSO_4$ and diluted 1:100 into M9 containing 10 μM $MgSO_4$ (for ON/induction) or 50 mM $MgSO_4$ (for OFF/repression). Cells were grown for 6 hr, diluted 1:50 into PBS with 0.5 g/L kanamycin, and fluorescence measured on a Miltenyi MACSQuant VYB. An identical procedure was used to induce AHK4-PhoQ fusions, except cells were grown in M9 containing 2 mM $MgSO_4$ and 1 nM aTc (anhydrotetracycline, Sigma) at all times, with the ON condition containing 1 μM *trans*-zeatin (Sigma) and the OFF condition containing no *trans*-zeatin. In each cytometry experiment, three replicates of each sample were induced independently and 20,000 cells were measured per replicate. FlowJo was used to analyze the data, gating on single, live cells and extracting the geometric mean of the YFP distribution. Error bars indicate the standard deviation of the geometric means measured in each replicate.

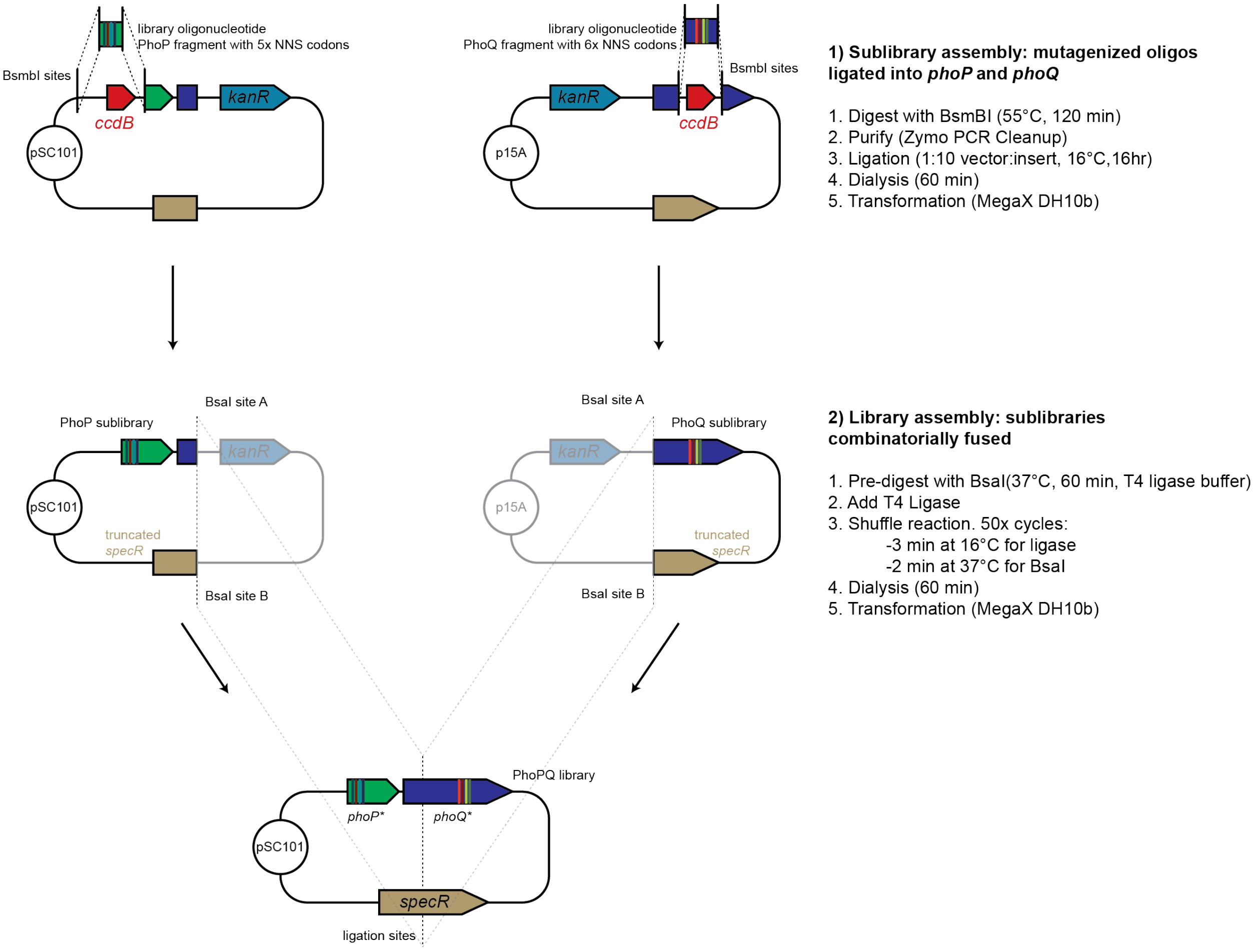
Design and assembly of degenerate PhoQ-PhoP library

The PhoQ-PhoP saturation mutation library was constructed by replacing the targeted residues with NNS codons(Diss and Lehner, 2018; Fowler and Fields, 2014; Starr et al., 2017). The residues targeted were selected, as indicated in the main text, based on amino acid coevolution analyses performed using GREMLIN (Balakrishnan et al., 2011).

The plasmid library was assembled in two general steps: 1) individual PhoP and PhoQ libraries were built in separate vectors and 2) sections of these vectors were combined to produce a new vector containing both PhoP and PhoQ mutants (Fig. 2.16). For the first step, oligonucleotide libraries for the sections of *phoP* and *phoQ* to be mutated were ordered from DNA 2.0. NNS nucleotides replaced codons 12, 14, 15, 18, and 19 in PhoP and codons 284, 288, 289, 292, 302, and 303 in PhoQ. These oligonucleotides were cloned into vectors pCM071 and pCM076 using the Type IIS restriction enzyme BsmBI. A toxic *ccdB* locus on these plasmids, used as a counter selection, was replaced during the process, ensuring a high rate of insertion incorporation. Both insert and vector were digested with BsmBI at 55°C for 2 hr and then purified on a Zymo DNA clean column. 1 pmol of both insert and vector were combined in a 25 μL reaction with 400 units of T4 ligase and incubated at 16 °C for 16 hr. Three ligations of each library were done, to ensure sufficient numbers of transformants. Ligations were dialyzed on Millipore VSWP 0.025 μm membrane filters for 60 min. and then the entire volume was electroporated into 20 μL of Invitrogen MegaX DH10B cells. From three ligations of each library, a total of 2.3×10^8 and 7.4×10^7 transformants were obtained for the PhoQ and PhoP libraries, respectively.

BsaI sites were used to join the two sublibraries into a single plasmid. The fusion points were designed such that faulty assemblies would not be viable: one junction was within the spec^R cassette (both the PhoP and PhoQ sublibraries harbored a kan^R cassette, but only half the spec^R open reading frame) and the other was within PhoQ. 500 fmol of minipreped DNA product (Qiagen) from each of the two libraries were combined in 25 μL T4 ligase buffer and digested with 1 μL BsaI for 1 hr at 37 °C. T4 ligase was then added and the reaction was cycled between 16 °C (3 min) and 37 °C (2 min) for 50 cycles to allow iterative ligation and digestion, running the reaction to completion. Final ligation product was dialyzed on Millipore VSWP 0.025 μm membrane filters for 60 min and the entire volume electroporated into 20 μL of Invitrogen MegaX DH10B cells. In total, 12 ligations and electroporations were done to produce a total of 5.72×10^8 transformants. Transformations were pooled and grown overnight (14 hr) in 100 mL 2xYT +

carbenicillin and spectinomycin. Following assembly, the plasmid library was purified by miniprep (Qiagen), dialyzed, and electroporated into Δ*phoPQ* strain CJM2044 to yield 3.8 x 10^9 transformants.



1) Sublibrary assembly: mutagenized oligos ligated into *phoP* and *phoQ*

1. Digest with BsmBI (55°C, 120 min)
2. Purify (Zymo PCR Cleanup)
3. Ligation (1:10 vector:insert, 16°C,16hr)
4. Dialysis (60 min)
5. Transformation (MegaX DH10b)

2) Library assembly: sublibraries combinatorially fused

1. Pre-digest with BsaI(37°C, 60 min, T4 ligase buffer)
2. Add T4 Ligase
3. Shuffle reaction. 50x cycles:
    - -3 min at 16°C for ligase
    - -2 min at 37°C for BsaI
4. Dialysis (60 min)
5. Transformation (MegaX DH10b)

Figure 2.16 Schematic summary of protocol for constructing the dual PhoQ-PhoP library.

Library selection and Sort-seq

The PhoQ-PhoP library was subjected to an initial selection of Mg^{2+} starvation to enrich for functional variants before performing fluorescence activated cell sorting (FACS). To this end, 6 mL of an overnight culture of the library (in 2xYT) was washed in M9 and diluted to an OD_{600} ~ 0.1 in 100 mL M9 containing 2 mM $MgSO_4$. Three replicates of this culture were made, and carried separately through the subsequent selection, FACS, and deep sequencing. Cells were grown for approximately 2 hours to OD_{600} ~ 0.4. at which point 1.6 mL of culture was washed three times in M9 containing no $MgSO_4$, and used to inoculate 100 mL of M9 containing no $MgSO_4$. After each dilution, the culture was sampled and a dilution series was plated on LB plates to ensure no bottlenecking occurred (CFUs > 1 x 10^9). The cultures in M9 containing no $MgSO_4$ were grown overnight (14 hr), with the OD_{600} increasing from 0.05 to only ~0.07. $MgSO_4$ was then added to bring the concentration to 2 mM, and cells were grown to an OD_{600} of 0.5 in 6 hrs, at which point glycerol stocks were made.

For FACS, 1 mL glycerol stocks were thawed and inoculated into 25 mL of M9. For each library replicate, one frozen stock aliquot was added directly to M9 containing 50 mM $MgSO_4$ (OFF state) and one aliquot was washed three times in M9 containing 0 mM $MgSO_4$ before inoculation into M9 with 10 μM $MgSO_4$ (ON state). To maintain cells in exponential phase, cultures were diluted (1:4 for ON state, 1:10 for OFF state) after 3 hours. After 6 hrs, cells were diluted again (1:5), and chloramphenicol was added to a concentration of 320 μg/mL and cells were placed on ice for sorting. CFP was expressed at a low constitutive level (attHK::[P_{tetA}-*cfp*]), and used to normalize YFP expression. Cells were sorted into ratiometric bins on the diagonal of CFP and YFP expression, to control for extrinsic expression noise in the YFP signal. For each library replicate, both the ON and OFF cultures were sorted into 8 separate bins, generating 48 total bins. Up to 2.5 million cells were sorted into bins per replicate (Fig. 2.4b, 2.12a). Sorted cells were added to 2xYT media containing 2 mM $MgSO_4$, carbenicillin, and spectinomycin, and then grown overnight.

Illumina sample preparation

After FACS, plasmids were purified (Qiagen MiniPrep) from overnight cultures representing each bin from each library replicate. For the two mutagenized regions of the plasmid to be brought into close enough proximity (< 790 bp) for paired-end Illumina sequencing, plasmids were digested with XhoI and then self-ligated (T4 ligase, 4 hr). To isolate only self-ligation products, and not cross-ligation products, ligation reactions were cleaned (Zymo PCR Clean Up) and gel purified to select for the correct size on FlashGels (Lonza). Two PCR reactions were performed, both using KAPA HiFi Hotstart, to add Illumina sequencing adaptors and barcodes. First, ligation reaction products were amplified for 30 cycles (95 °C for 30 s, 65 °C for 15 s, 72 °C for 120 s) with primers CJM642 and CJM643 in an emulsion PCR (Micellula Emulsion PCR) to avoid PCR chimeras. Second, purified PCR product from the first reaction was subjected to a second PCR with barcoding primers for 9 cycles (95 °C for 30 s, 65 °C for 15 s, 72 °C for 60 s). Final products were quantified (NanoDrop), normalized, combined, and sequenced on an Illumina NextSeq. For each bin, 1-33 million reads were collected.

Construction of combinatorial 79 x 71 mutant library

79 pairs of PhoQ*-PhoP* variants were selected that displayed broad sequence diversity and high fold induction in the initial PhoQ*-PhoP* library. These variants were cloned into plasmids pCM143 (PhoP) and pCM149 (PhoQ) using blunt end ligation (Ashenberg et al., 2013) and Gibson assembly (Gibson et al., 2009). Each pCM143 and pCM149 variant was amplified by PCR (KAPA Hifi, 30 cycles) to generate amplicons for Gibson assembly (primers CM937/CM1531 for pCM149 and primers CM938/CM1532 for pCM143). PCR products were cleaned (Zymo PCR Cleanup), quantified by NanoDrop, and combined into an equimolar mix of pCM143 amplicons and an equimolar mix of pCM149 amplicons. The two mixes were combined

in Gibson Assembly master mix (300 fmol of large pCM143 fragment, 900 fmol of smaller pCM149 insert), incubated at 50 °C for 2 hr and heat killed at 79 °C for 20 min. The assembly was dialyzed on Millipore VSWP 0.025 μm membrane filters for 60 min and transformed into electrocompetent CJM2044 cells.

Unlike the treatment of the initial, larger library, this library was not subjected to a low magnesium selection step. Immediately after construction, this library underwent Sort-seq, as described above.

Illumina data processing

The frequency of each mutant in each bin was calculated as a function of the total reads in each bin and the total number of cells sorted into each bin:

$$freq\ (seq, bin, replicate) = \frac{reads_{seq,bin}}{\sum_{sequences \in bin}^{s} reads_{s,bin}} * \frac{cells_{bin,replicate}}{\sum_{bins\,\in\,replicate}^{b} cells_{b,replicate}}$$

All Sort-seq plots display the mean frequencies in each bin across three replicates, with error bars indicating standard deviation. Gaussian functions were fit to each distribution (in $\log_{10}$YFP units), from both the ON and OFF sorts (SciPy optimize package). Variants with fewer than 25 total reads were discarded before fitting. Poor Gaussian fits have high variances on the estimated parameters. The standard deviation error on the estimated log (YFP) mean (σ_{fit}) was used as a metric to filter poorly fit sequences: sequences were removed if $\sigma_{fit,ON} + \sigma_{fit,OFF} > 2$. In total, 10,595 unique variants passed these filters. Fold-induction values were calculated as the ratio of the fit means between the induced and uninduced states: μ_{ON} / μ_{OFF}.

During analysis of the second (79 x 71) library, the fold induction was calculated for the most frequent nucleotide sequence representing each amino-acid sequence. For visualization and orthogonal set design, fold inductions were bounded between 1 and 20. Individually tested mutants generally did not surpass the sensitivity of wild type PhoQ-PhoP, which displayed 20-fold induction during Sort-seq, suggesting that signal above 20-fold may be due to noise. The axes of the 79 x 71 matrix were clustered hierarchically using the WGPMC method (Scipy). During clustering, matrix entries which lacked data were assigned the mean value of all other entries.

Analysis of the sensitivity of Sort-seq quantification to read count

To assess the quality of Sort-seq-based quantification for variants with lower read coverage, variants with high read coverage (2,000-10,000 total reads) were down-sampled to simulate low read coverage and fed through the Sort-seq analysis pipeline (Fig. 2.5). For each of these high-coverage variants, simulated data was produced by down sampling 100 independent times by the factors indicated in Fig. 2.5. Simulated read coverage was generated by sampling (with replacement) from the original reads of each variant up to the desired read coverage. Simulated reads were then subjected to the same Gaussian fitting protocol as before. As in the original analysis, poor fits ($\sigma_{\text{fit,ON}} + \sigma_{\text{fit,OFF}} > 2$) were discarded. All simulated variants were classified as functional or non-functional, based on the fold-induction values of original, high coverage variants they were sampled from (Fig. 2.5c, e). False positive rates (Fig. 2.5d) and false negative rates (Fig. 2.5f) were then calculated as a function of read coverage by computing the fraction of simulated variants that were mis-classified.

Orthogonal set design

Orthogonal sets of PhoQ* and PhoP* variants up to 7 pairs were identified by systematically scanning all PhoQ*-PhoP* permutations within the 79 x 71 matrix. Larger orthogonal sets were identified using a greedy search algorithm. An objective function *Q* was used to quantify the (geometric) average difference between fold-induction values and the maximum cross-talk value for each pair within a given set of variants:

$$Z = \prod_{n=1}^{N} p\,(h_{i_n}, r_{j_n})^{\frac{1}{N}}$$

$$p\,(i,j) = \begin{bmatrix} d_{i,j} & if\ d_{i,j} > 0 \\ exp(d_{i,j}) & if\ d_{i,j} \leq 0 \end{bmatrix}$$

$$d_{i,j} = FI_{i,j} - \max\left(\max_{a\neq i} FI_{a,j}\,, \max_{b\neq j} FI_{i,b}\right)$$

where *p* represents the pair score for each PhoQ*/PhoP* pair *n* within the set and *Z* is the score for the entire set of N pairs. $FI_{i,j}$ indicates fold induction for a given pair PhoQ$_i$*/PhoP$_j$* and *a, b* are other PhoQ*, PhoP* variants, respectively, within the set. Values of *FI* missing Sort-seq data were assumed to be worst-case scenarios: missing fold-induction values of pairs were assumed to be 1, and missing cross-talk values were assumed to be 20.

Using these scoring functions, orthogonal sets were generated *in silico*. Sets were randomly initialized and individual pairs from the set were replaced in a biased, random fashion. Pairs were dropped from the set with a probability proportional their pair score *p*. To choose the replacement pair, difference score $d_{i,j}$ was calculated for the ~6,000 new pair possibilities and one was selected with a probability proportional to the exponent of its score within the current set:

$$P\left(h_i, r_j = new\ pair\right) \propto e^{d_{i,j}}$$

The 5x5 orthogonal set described in Fig. 2.11e was further optimized by testing single point mutants of the single PhoP* variant (VHSCL) that initially displayed cross-talk with a non-cognate

PhoQ* variant. Each of the 5 PhoP* specificity residues was replaced independently with an NNK codon to generate all possible single point mutants (xHSCL, VxSCL, VHxCL, VHSxL, VHSCx where x is any amino acid specified by the NNK codon). These mutants were cotransformed with the cognate PhoQ* variant (SCEHLI) into CJM2044 and grown overnight in M9 media containing 0 mM $MgSO_4$ to remove non-functional variants. After plating the surviving strains on LB agarose plates, 48 clones were tested for Mg^{2+} induction (see below). The pCM143-PhoP* plasmids from the 24 clones with the strongest induction were purified and cotransformed with the non-cognate PhoQ* variant (AGGCYF) into CJM2044. Mg^{2+} induction was measured by cytometry and the 8 clones displaying the highest cognate / non-cognate induction ratio were selected for testing with all five PhoQ* variants. Two of these eight clones were PhoP* (VYSCL), which displayed the highest specificity.

Reconstruction and *in vivo* characterization of individual PhoQ* and PhoP* variants

Variants were cloned into plasmids pCM143 (PhoP) and pCM149 (PhoQ) using blunt end ligation (Ashenberg et al., 2013) and Gibson assembly (Gibson et al., 2009). Combinations of pCM143 and pCM149 plasmids were co-transformed into strain CM2044. Colonies were grown overnight in M9 containing 2 mM $MgSO_4$ and induced, as described above, with M9 containing either 10 μM or 50 mM $MgSO_4$. After 6 hours, cultures were diluted 1:50 into cold PBS containing 0.5 g/L kanamycin, and fluorescence measured on a Miltenyi MACSQuant VYB. Note that the fold-induction values of individually tested variant pairs were generally smaller than those measured by Sort-seq, likely due to differences between the Miltenyi cytometer and BD Aria Sorter; however, the two measurements were highly correlated (Pearson $R^2 = 0.91$, Fig. 2.3c).

Purification of two-component signaling proteins and *in vitro* phosphotransfer assays

Expression and purification of PhoQ* and PhoP* variants, and phosphotransfer experiments were carried out as previously described (Podgornaia and Laub, 2015; Skerker et al., 2005). PhoP* was purified fused to a His_6-Trx tag, and the cytoplasmic region of PhoQ* (residues 238-486) was fused to a His_6-MBP (maltose binding protein) tag. For phosphotransfer reactions, the kinase was autophosphorylated for 1 hr at 30 °C with [γ-^{32}P] ATP (Perkin Elmer) before being combined with PhoP* at a 1:8 ratio (10 μL reactions contained 1 μM PhoQ* and 8 μM PhoP*). Reactions were stopped at appropriate times by adding 4x Laemmli buffer with 8% 2-mercaptoethanol. This process allowed monitoring of both phosphotransfer and phosphatase activities between PhoQ* and PhoP* variants (Fig. 2.6b).

For PhoQ* variants where *in vitro* autophosphorylation was not observed, phosphatase activity was assayed by mixing a given PhoQ* variant with a PhoP* variant that was phosphorylated using wild-type PhoQ. This was achieved by incubating 8 μM PhoP* for 1 hr at 30 °C with [γ-^{32}P] ATP and 1 μM wild-type PhoQ, which promiscuously phosphorylates, but does not dephosphorylate, most PhoP variants. After generating phosphorylated PhoP*, 1 μM of the PhoQ* variant was added and samples taken and reactions stopped as before.

Phosphatase activity of PhoQ* variants with respect to other response regulators was measured with a similar assay. Twelve *E. coli* response regulators were selected for their ability to be stably phosphorylated *in vitro* by a cocktail of six *E. coli* histidine kinases (CreC, RstA, PhoR, PhoP, EnvZ and CpxA, each at 250 nM). After 2 hours of pre-incubation with radiolabeled ATP and this kinase cocktail, each regulator was combined with 2 μM PhoQ or PhoQ* variant. Reactions were stopped at 0, 60 and 120 minutes by adding 4x Laemmli buffer with 8% 2-mercaptoethanol.

Response regulators for phosphotransfer profiles were purified as described above. Each was fused to an N-terminal His_6-Trx tag, expressed in BL21 (DE3) cells and purified on a Ni^{2+}-NTA column(Capra et al., 2012). Conditions for phosphophotransfer profiles were also identical to above conditions; 10 µL reactions containing 1 µM [γ-^{32}P] autophosphorylated PhoQ* and 8 µM response regulator were generated and stopped after 5 or 60 minutes. Gel images were analyzed using quantified with ImageJ.

RNA-seq

Cultures were grown overnight in M9 media containing 2 mM $MgSO_4$ and then diluted 1:25 into fresh M9 containing 2 mM $MgSO_4$ and grown for 2 hours to reach OD_{600} ~ 0.5. For the OFF condition, 1 mL of cells was diluted into 2 mL M9 containing 74 mM $MgSO_4$ for a final concentration of 50 mM $MgSO_4$. For the ON condition, 2 mL of cells were pelleted, washed twice with M9 + 10 µM $MgSO_4$, resuspended in M9 + 10 µM $MgSO_4$, and then 1 mL of cells was diluted into 2 mL M9 + 10µM $MgSO_4$. Induction of AQ4 strains was identical, except cells were induced for 30 minutes in M9 (2 mM $MgSO_4$) containing either 0 µM *trans* zeatin (OFF condition) or 1 µM *trans* zeatin (ON condition).

RNA was harvested as previously described (Culviner and Laub, 2018). After 30 minutes of growth, cells from each condition were harvested by adding 1.8 mL culture to 200 µM cold stop solution (95% ethanol, 5% acid buffered phenol, 4 °C). The mixture was centrifuged for 30 s at 13,000 rpm on a benchtop centrifuge, and the supernatant was removed with the pellet flash frozen in liquid nitrogen and stored at -80 °C. To extract RNA, Trizol (Invitrogen) was heated to 65 °C, added directly to the pellet, and incubated at 65 °C for 10 minutes with shaking at 2000 rpm (Eppendorf Thermomixer). The mixture was frozen at -80 °C for at least 10 minutes. After thawing, cells were centrifuged at 15,000 rpm, 4 °C for 5 minutes, and the supernatant was removed into 400 μL ethanol. The mixture was applied to a DirectZol spin column (Zymo) and centrifuged for 30 s at 13,000 rpm. The columns were washed with DirectZol RNA prewash buffer twice (400 μL) and RNA wash buffer (700 μL) once before eluting in 90 μL DEPC water. 10 μL 10x Turbo DNase buffer and 2 μL Turbo DNase (Invitrogen) were added to the eluant. The mixture was digested at 37 °C for 20 minutes, followed by the addition of 2 μL more DNase and another 20-minute incubation. Total volume was brought to 200 μL with DEPC water and combined with 200 μL acid-phenol:chloroform (IAA, Invitrogen), vortexed and centrifuged for 10 minutes at 21,000 g and 4 °C. The top (aqueous) layer was extracted and ethanol precipitated in 20 μL NaOAc (3M), 2 μL GlycoBlue (Invitrogen) and 600 μL cold ethanol. Precipitation mix was incubated at -80 °C for more than 4 hours before centrifuging for 30 minutes at 21,000 g and 4 °C. The pellet was washed twice with 500 μL cold 70% ethanol, then air dried and resuspended in 50 μL DEPC water. RNA integrity was validated on a 6% TBE-urea acrylamide Novex gel (Invitrogen) and yield was quantified by NanoDrop spectrophotometer. rRNA was removed with the RiboZero rRNA Removal Kit for Bacteria (Illumina). RNA was fragmented and cDNA libraries were prepared at the MIT BioMicro Center sequencing core using the KAPA RNA HyperPrep Kit (Roche) and sequenced on an Illumina HiSeq. Reads were mapped to the *E. coli* genome and plasmids with bowtie2 using default parameters (Langmead and Salzberg, 2012).

To determine whether selected PhoQ*-PhoP* variants interfered with other two-component signaling pathways, we examined whether any other response regulator or histidine kinase genes were upregulated transcriptionally when PhoQ* variants were activated by low Mg^{2+}. We calculated the fold change in expression of each two-component regulatory gene as a ratio of reads in low and high Mg^{2+} (Fig. 2.10b). Note that *rstA* and *rstB* are part of the PhoP regulon and are directly upregulated by wild-type PhoQ-PhoP and most variant pairs. To quantify how these fold changes compared to wild-type PhoQ-PhoP, we calculated the ratio the fold change in each gene to the geometric mean of the fold change in the same gene for the two wild-type replicates (Fig. 2.10c). To assess whether any two-component signaling genes were significantly upregulated, we calculated the Z-score of each ratio of fold changes and conducted a one-tailed test to compute *p* values (Fig. 2.10d). After using a Bonferroni correction for multiple hypothesis testing, no genes were found to be significantly upregulated ($p < 0.05$).

Identification of two-component signaling proteins and generation of force-directed graphs

The RefSeq Prokaryotic Genomes database of 5,506 bacterial genomes (Sept, 2017) was downloaded from NCBI. The database was scanned for histidine kinases and response regulators using *jackhmmer (* (E-value cutoff = 0.01) with all two-component signaling proteins from *E. coli* used as queries. The combined lists of HK and RR hits were aligned with *hmmalign (* to the PFAM hidden Markov models for HisKA and Response_Reg domain families, respectively. Columns in the multiple sequence alignment with greater than 80% gaps were eliminated, and sequences with greater than 50% gaps were discarded. Histidine kinases lacking the catalytic histidine and response regulators lacking the catalytic aspartate were removed. Proteins containing both the HK DHp domain and a RR receiver domain were discarded to avoid ambiguity. HKs and RRs were then labeled as exclusive pairs if they were (i) within 20 genes in the genome, with no other HK or RR genes between, (ii) on the same strand, and (iii) closer to no other potential HK or RR partner (with distance defined as the number of genes between partners). The sequences of paired HKs and RRs were concatenated and the multiple sequence alignment was then reduced to the eleven positions mutated in this study.

The force directed graph was generated using the Gephi network visualization package(Jacomy et al., 2014). To construct a network, the 85,782 co-operonic HK-RR pairs identified by HMMER in bacterial genomes were combined with the functional mutant sequences from the PhoQ-PhoP dual library that had fold-induction values > 18 and the mutant variants within the characterized 5x5 orthogonal set (Fig. 2.11c, e). These sequences were treated as nodes and were connected by edges if the pairwise alignment score for the two sequences' 11 specificity residues (using the BLOSUM62 scoring matrix) exceeded a threshold score of 20. If more than 40 edges were connected to a node, only the top scoring 40 edges were kept. If no edge scoring above 20 connected a node, that node retained its top-scoring edge, despite that edge being below the BLOSUM62 threshold. A final model of ~86,000 nodes and 2.5 million edges was loaded into Gephi and visualized using the Force-Atlas-2 tool (Jacomy et al., 2014).

Construction of AHK4-PhoQ chimera

Chimeric histidine kinases sensors were made using a variation of the PATCHY strategy (primer aided truncation for the creation of hybrid proteins)(Ohlendorf et al., 2016). The N-terminal region of AHK4 (residues 1-475) were cloned downstream of the P_{tac} promoter on a p15A/kan^R vector. This plasmid was amplified via PCR with primers containing SapI sites to allow insertion of the PhoQ kinase domain. Five distinct sets of primers allowed five possible junction sites within AHK4 (residues A466, A468, A469, A472 and A478) with identical GCG overhangs. PhoQ (pCM149) was amplified with 32 distinct primers (also containing SapI sites) to generate 32 C-terminal truncations beginning upstream of the DHp domain (residues 213-224, 257-276). PCR products were gel-purified (Zymo), then combined in a 50 μL ligation reaction containing 400 U of T4 ligase (NEB), 20 U Sap1 (NEB), 100 fmol pooled AHK4 PCR products, and 500 fmol pooled PhoQ PCR products. The reaction was cycled 50 times between 37 °C (2 min) and 16 °C (3 min) to drive assembly to completion, heat killed at 50 °C (20 min) and 80 °C (20 min), and dialyzed on Millipore VSWP 0.025 μm membrane filters (60 min). This small library of 160 possible fusions was transformed into electrocompetent CJM2044 cells harboring PhoP on a plasmid (pCM143).

To enrich for chimeras responsive to *trans*-zeatin we used Mg^{2+} starvation as a selection. An overnight culture of the library in M9 media was induced for 1 hour (M9 containing 21 nM aTc, 1 μM *trans*-zeatin), washed three times in M9 containing no $MgSO_4$ and diluted 1:10 into 100 mL M9 containing no $MgSO_4$, 21 nM aTc, 1 μM *trans*-zeatin. After 4 hours, 500 μL of culture was plated on LB. 96 colonies were picked and screened for *trans*-zeatin-dependent YFP expression.

GCN4-DHp fusions to test phosphatase buffering against crosstalk *in vivo*

To generate cytosolic variants of PhoQ locked in a phosphatase state, we followed a previously described strategy (Wang et al., 2014) and fused GCN4 (MKQLEDKVEELLSKNYHLENEVARL) N-terminal to PhoQ's DHp domain. pCM149 (*lacUV5-phoQ, kanR, p15A*) was amplified with 24 distinct primers (containing SapI sites) to generate 24 C-terminal truncations beginning upstream of the DHp domain (residues 222-225, 257-276). The N-terminus of PhoQ was replaced by GCN4 in each of these plasmids, removing the transmembrane and sensory domains. Each GCN4 fusion plasmid was transformed with pCM143 (*phoP*) and pCM150 (P_{mgrB}-*yfp*) into TIM175 and tested by standard Mg^{2+} induction (see above) for activity. As expected, some variants displayed constitutive high YFP (presumably locked kinase conformation) or constitutive low YFP (presumably locked phosphatase conformation) and stepwise amino acid insertions displayed a periodicity of these phenotypes (Fig. 2.12f). One of these fusions (fusion-266) displayed even lower constitutive YFP values than PhoQ with mutations in the ATP cap (R434M, R439M, Q442M, pCM180) or ATP pocket (N385L, N389L, K392M, Y393F, pCM179)(Marina et al., 2001).

To test the ability of a cognate phosphatase to suppress crosstalk from a non-cognate kinase, we used a three-plasmid setup: pCM874 (reporter plasmid pCM150 with PlacUV5-PhoP$_{15}$* inserted), pCM149 (PlacUV5-PhoQ$_{wt}$) and pCM873 or pCM898 (Ptet-GCN4-fusion266-PhoQ$_{wt}$ or -PhoQ$_{15}$*, respectively). Because PhoP* has been moved from a low-copy to medium-copy plasmid, crosstalk between PhoQwt and PhoP* is likely exacerbated by overexpression, as noted by the high level of induction seen in Fig. 2.12g before aTc is added. However, induction of the GCN4-fusion266-PhoQ$_{15}$* phosphatase effectively eliminates this cross-talk (Fig. 2.12g, right panel). Induction of the non-cognate phosphatase, GCN4-fusion266-PhoQ$_{wt}$*, does not relieve this cross-talk.

Data Availability

Datasets generated during this study have been deposited in GEO. Raw reads and processed Sort-seq analysis of each mutant can be found under the accession numbers GSE120780 (degenerate PhoQ-PhoP library) and GSE120786 (combinatorial library of 79 PhoQ* and 71 PhoP* variants). Raw reads and RPKM for all *Escherichia coli* genes from RNA-seq are deposited with the accession number GSE128611.

Code Availability

Python scripts for analysis available at https://github.com/mcclune/nature2019 .

References

Alm, E., Huang, K., and Arkin, A. (2006). The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol *2*, e143.

Ashenberg, O., Keating, A.E., and Laub, M.T. (2013). Helix bundle loops determine whether histidine kinases autophosphorylate in cis or in trans. Journal of molecular biology *425*, 1198-1209.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061-1078.

Bashor, C.J., Helman, N.C., Yan, S., and Lim, W.A. (2008). Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. Science *319*, 1539-1543.

Boyken, S.E., Chen, Z., Groves, B., Langan, R.A., Oberdorfer, G., Ford, A., Gilmore, J.M., Xu, C., DiMaio, F., Pereira, J.H.*, et al.* (2016). De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. Science *352*, 680-687.

Brentjens, R.J., Davila, M.L., Riviere, I., Park, J., Wang, X., Cowell, L.G., Bartido, S., Stefanski, J., Taylor, C., Olszewska, M.*, et al.* (2013). CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. Sci Transl Med *5*, 177ra138.

Capra, E.J., and Laub, M.T. (2012). Evolution of two-component signal transduction systems. Annu Rev Microbiol *66*, 325-347.

Capra, E.J., Perchuk, B.S., Lubin, E.A., Ashenberg, O., Skerker, J.M., and Laub, M.T. (2010). Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. PLoS Genet *6*, e1001220.

Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. Cell *150*, 222-232.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Creixell, P., Palmeri, A., Miller, C.J., Lou, H.J., Santini, C.C., Nielsen, M., Turk, B.E., and Linding, R. (2015). Unmasking determinants of specificity in the human kinome. Cell *163*, 187-201.

Culviner, P.H., and Laub, M.T. (2018). Global Analysis of the E. coli Toxin MazF Reveals Widespread Cleavage of mRNA and the Inhibition of rRNA Maturation and Ribosome Biogenesis. Mol Cell *70*, 868-880 e810.

Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. Elife *7*.

Dueber, J.E., Mirsky, E.A., and Lim, W.A. (2007). Engineering synthetic signaling proteins with ultrasensitive input/output control. Nat Biotechnol *25*, 660-662.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat Methods *11*, 801-807.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods *6*, 343-345.

Groban, E.S., Clarke, E.J., Salis, H.M., Miller, S.M., and Voigt, C.A. (2009). Kinetic buffering of cross talk between bacterial two-component sensors. J Mol Biol *390*, 380-393.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. Plos One *9*.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Laub, M.T., and Goulian, M. (2007). Specificity in two-component signal transduction pathways. Annu Rev Genet *41*, 121-145.

Marina, A., Mott, C., Auyzenberg, A., Hendrickson, W.A., and Waldburger, C.D. (2001). Structural and mutational analysis of the PhoQ histidine kinase catalytic domain. Insight into the reaction mechanism. J Biol Chem *276*, 41182-41190.

Mimee, M., Citorik, R.J., and Lu, T.K. (2016). Microbiome therapeutics - Advances and challenges. Adv Drug Deliv Rev *105*, 44-54.

Morsut, L., Roybal, K.T., Xiong, X., Gordley, R.M., Coyle, S.M., Thomson, M., and Lim, W.A. (2016). Engineering Customized Cell Sensing and Response Behaviors Using Synthetic Notch Receptors. Cell *164*, 780-791.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.-A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P*., et al.* (2013). Precise and reliable gene expression via standard transcription and translation initiation elements. Nature methods *10*, 354-360.

Nielsen, A.A., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., Ross, D., Densmore, D., and Voigt, C.A. (2016). Genetic circuit design automation. Science *352*, aac7341.

Ohlendorf, R., Schumacher, C.H., Richter, F., and Moglich, A. (2016). Library-Aided Probing of Linker Determinants in Hybrid Photoreceptors. ACS Synth Biol *5*, 1117-1126.

Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science *347*, 673-677.

Reinke, A.W., Grant, R.A., and Keating, A.E. (2010). A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. J Am Chem Soc *132*, 6025-6031.

Riglar, D.T., Giessen, T.W., Baym, M., Kerns, S.J., Niederhuber, M.J., Bronson, R.T., Kotula, J.W., Gerber, G.K., Way, J.C., and Silver, P.A. (2017). Engineered bacteria can function in the mammalian gut long-term as live diagnostics of inflammation. Nat Biotechnol *35*, 653-658.

Riglar, D.T., and Silver, P.A. (2018). Engineering bacteria for diagnostic and therapeutic applications. Nat Rev Microbiol *16*, 214-225.

Rowland, M.A., and Deeds, E.J. (2014). Crosstalk and the evolution of specificity in two-component signaling. Proc Natl Acad Sci U S A *111*, 5550-5555.

Siryaporn, A., and Goulian, M. (2008). Cross-talk suppression between the CpxA-CpxR and EnvZ-OmpR two-component systems in E. coli. Mol Microbiol *70*, 494-506.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Sockolosky, J.T., Trotta, E., Parisi, G., Picton, L., Su, L.L., Le, A.C., Chhabra, A., Silveria, S.L., George, B.M., King, I.C., *et al.* (2018). Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. Science *359*, 1037-1042.

Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories in the sequence space of an ancient protein. Nature *549*, 409-413.

Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. Science *317*, 364-369.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-component signal transduction. Annu Rev Biochem *69*, 183-215.

Suzuki, T., Miwa, K., Ishikawa, K., Yamada, H., Aiba, H., and Mizuno, T. (2001). The Arabidopsis sensor His-kinase, AHk4, can respond to cytokinins. Plant Cell Physiol *42*, 107-113.

Thompson, K.E., Bashor, C.J., Lim, W.A., and Keating, A.E. (2012). SYNZIP protein interaction toolbox: in vitro and in vivo specifications of heterospecific coiled-coil interaction domains. ACS Synth Biol *1*, 118-129.

Wang, B., Zhao, A., Novick, R.P., and Muir, T.W. (2014). Activation and inhibition of the receptor histidine kinase AgrC occurs through opposite helical transduction motions. Mol Cell *53*, 929-940.

Yamada, H., Suzuki, T., Terada, K., Takei, K., Ishikawa, K., Miwa, K., Yamashino, T., and Mizuno, T. (2001). The Arabidopsis AHK4 histidine kinase is a cytokinin-binding receptor that transduces cytokinin signals across the membrane. Plant Cell Physiol *42*, 1017-1023.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676-680.

Chapter 3 – Conclusions and future directions

Conclusions

During my graduate studies I have used two-component signaling pathways to study the systems-wide restrictions on protein-protein specificity. Previous work in the Laub Lab identified that the specificity of the kinase-regulator interaction is determined by a small set of residues that can often be transferred modularly from with complex into another (Capra et al., 2010; Podgornaia et al., 2013; Skerker et al., 2008). A subsequent study identified an evolutionary event where duplication of one two-component pathway forced another to change its specificity and avoid cross talk (Capra et al., 2012). These findings raised a number of questions that motivated the work in this thesis: How much does the systems-wide pressure of cross-talk restrict the evolutionary birth of new paralogous systems? Furthermore, can we use specificity-determining regions of a protein not only to transplant specificity, but to develop new-to-nature specificity endowing proteins with orthogonality to existing homologs?

To address these questions, I sought to sample the sequence space of two-component systems in an unbiased fashion and determine which fraction of the sequence space is both functional and available in *E. coli*. Using residue coevolution models, I constructed a dual library of PhoQ-PhoP mutants that maximized the diversity of interaction specificities while minimizing mutations disruptive to protein fold and function. I developed a method utilizing cell sorting and Illumina sequencing to quantitatively characterize the phenotype of thousands of interface variants in multiple induction conditions. The result was the isolation of hundreds of PhoQ-PhoP variant pairs with novel interfaces and novel specificity. About 40% of these engineered pairs were insulated from both parent proteins. *In vitro* characterization identified that partnering is generally enforced not by kinase specificity, but by stringent phosphatase specificity.

These novel PhoQ*-PhoP* variants display diverse specificities, yet do not interfere with the 27 endogenous two-component pathways in *E. coli*. *In vitro* phosphotransfer profiling assays never identified any PhoQ* variant able to substantially phosphorylate or dephosphorylate any endogenous regulator. Similarly, these *in vitro* assays demonstrated that no PhoP* variant is susceptible to significant phosphorylation by an endogenous kinase. Furthermore, RNAseq experiments showed that replacing wild-type PhoQ-PhoP with an orthogonal PhoQ*-PhoP* pathway caused no significant perturbation of any two-component pathway in *E. coli.*

Despite finding such stringent insulation with all endogenous paralogs, the functional PhoQ*-PhoP* variants exhibit a diversity of specificities. A second, combinatorial library revealed that many variants pairs were insulated from each other. Large sets of mutually orthogonal pairs could be constructed from these data, in which each PhoQ* variant functioned with only a single PhoP* variant.

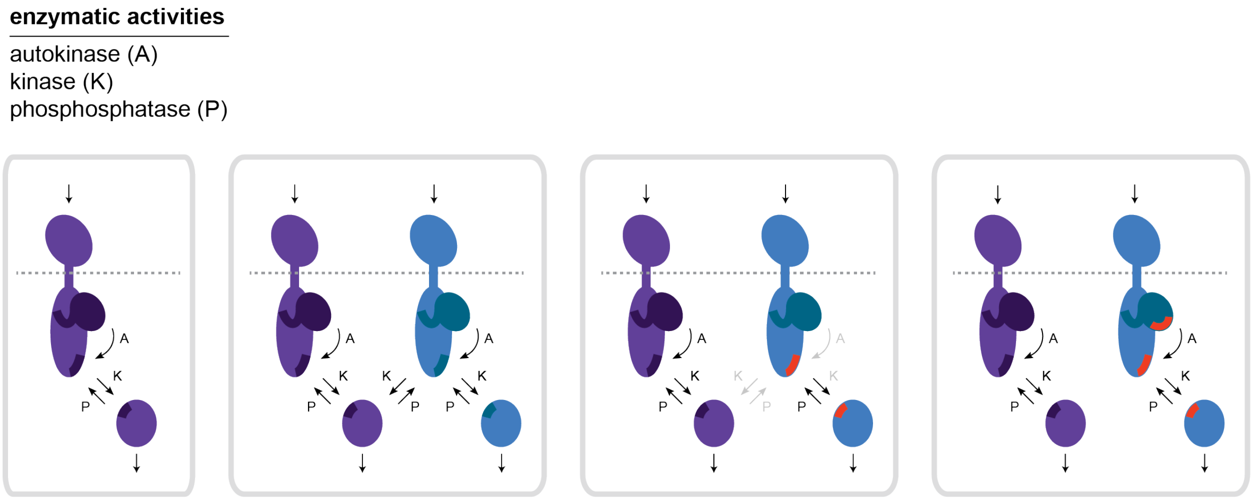
In short, diversification of only a few specificity-determining residues allows histidine kinases and response regulators to access many diverging specificities. These findings show that multiple new pathways can be introduced within the sequence space defined by these residues. New protein complexes can be built such that they are insulated from each other and from all paralogous complexes already present in the host organism. These data suggest that novel specificity can arise both by accessing new regions of sequence space, with interface residues not found in any extant two-component pair, or through subfunctionalization - partitioning the sequence space niche occupied by the parent proteins into multiple, orthogonal components by narrowing specificity.

Future directions

Sufficiency of phosphatase activity

Approximately half of the PhoQ* variants purified in this body of work were unable to autophosphorylate *in vitro*, despite folding and maintaining phosphatase activity. While this loss of activity may be an artifact of truncating the PhoQ to its cytoplasmic domains, it is not unreasonable to imagine that many of these variants function, without any kinase activity, as inducible phosphatases. Data from this work and others has shown that knocking out a histidine kinase can result in high constitutive phosphorylation of the cognate regulator due to other phosphodonors, such as acetyl phosphate (Atkinson et al., 2003) (Podgornaia and Laub, 2015) (Wanner and Wilmes-Riesenberg, 1992). Thus, it is feasible that switching a histidine kinase from a phosphatase state to a state lacking any enzymatic activity, rather than the canonical phosphatase/kinase dichotomy, may be sufficient to achieve signal-dependent phosphorylation of a response regulator. If this is the case, it may explain not only why a fraction of histidine kinases lack the catalytic histidine, but may also represent a minimal, functional two-component sensor that could be an evolutionary precursor to modern histidine kinases. Furthermore, it offers a possible evolutionary intermediate that retains signal responsiveness despite having only one of the three catalytic activities normally found in histidine kinases (Fig. 3.1). Alterations of specificity residues can differentially affect a DHp domain's ability to phosphorylate or dephosphorylate its regulator (Podgornaia and Laub, 2015), as well as its ability to interact with the CA domain during autophosphorylation. When a kinase is under pressure to change specificity, e.g. after a duplication event, dramatic mutations could be tolerated if the phosphatase activity is sufficient for the pathway to continue transmitting information.

Additional interrogation of the three catalytic activities is needed to test this hypothesis. For functional PhoQ* variants that lack *in vitro* autophosphorylation activity, it would be enlightening to explore whether the ATP-hydrolyzing domain, or at least its catalytic residues, are necessary for function *in vivo*. Is there conservation of the CA domain in all histidine kinases? Furthermore, do extant histidine kinases lacking the eponymous histidine or catalytic domain retain phosphatase activity? Is there phylogenetic correlation between changes in specificity residues and loss of catalytic signatures? Pursuing these questions would illuminate whether inducible phosphatases are biologically and evolutionarily relevant for two-component signaling.



single copy (ancestral)

duplication

specificity drift, inducible phosphatase

restoration of kinase activity

Figure 3.1 A model for inducible phosphatases serving as a minimal, functional intermediate during evolutionary rewiring and insulation.

Negative and positive design elements in paralog insulation

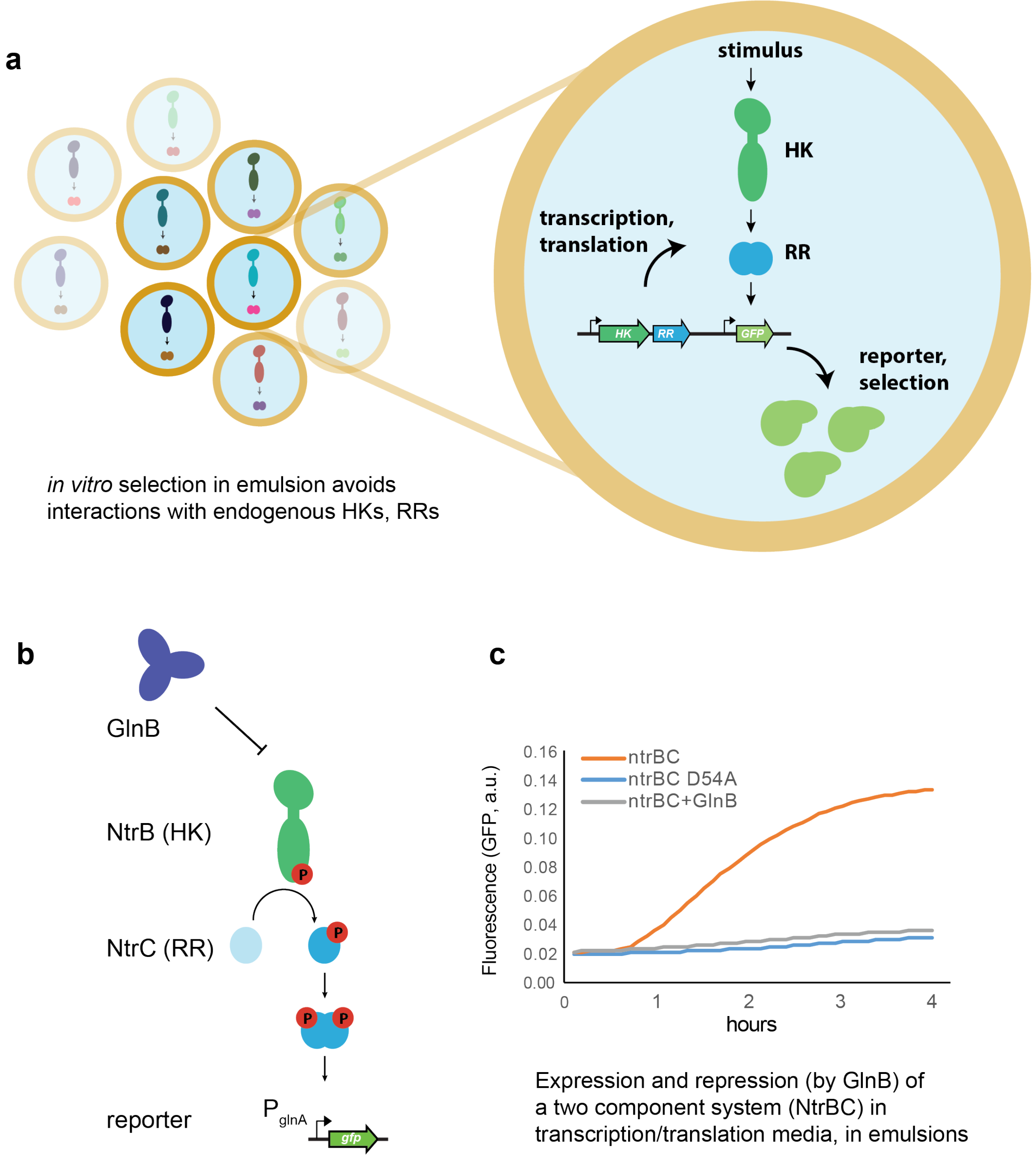
Investigating the occupancy of sequence space is intrinsically tied to the role of negative and positive design elements of protein-protein interactions. At the amino acid level, positive design elements are residues that contribute binding energy for cognate interactions, while negative design elements are residues that inflict energetic costs when contacting non-cognate partners. There is no reason that specificity-determining residues should sort exclusively into these two classifications. However, previous studied have identified either residues or protein loops that act solely as negative design elements to keep paralogs insulated (Lite TL, unpublished data) (Plach et al., 2017) (Richardson and Richardson, 2002). Furthermore, a prevalence of negative design elements would be suggestive of a more crowded sequence space (Zarrinpar et al., 2003). Negative design features narrow the niche of a protein within sequence space, and thus are suggestive that system-wide, negative selection pressure may be necessary to insulate many paralogs. In other words, negative design elements enforce insulation in a context-dependent manner; their presence provides information about how specific an interaction will remain if moved to a new genome.

The orthogonal PhoQ*-PhoP* variants identified in this study did not experience such selection pressure. Thus, they provide a null model for addressing questions about positive and negative design in evolving protein interaction. In the context of two, insulated protein pairs, negative design elements can be experimentally identified as residues that, if mutated, do not change cognate partnering but reduce insulation from noncognates. This precise definition lays a groundwork for systematic mutagenesis studies to compare the role of negative design elements in evolved and engineered sets of mutually orthogonal pairs.

Direct, systematic investigation of paralog interference

Some limitations of my experimental setup restricted what we could learn about paralog interference. Firstly, my PhoQ-PhoP library sparsely sampled ~5e8 of ~2e14 total possible sequences, preventing me from gaining information about possible mutational paths. Secondly, I had no readout or selection for crosstalk between PhoQ-PhoP and *E. coli*'s other two-component pathways. There may have been fitness costs that intrinsically selected against crosstalking variants, but we cannot currently know. However, my experimental approach could be extended in various ways to directly test the role of paralog interference in pathway divergence and insulation maintenance. An interface library of PhoQ-PhoP or other tractable two-component pathway could be subjected to high-throughput selection and characterization in multiple knockout strains, each an endogenous two-component pathway. If a mutant pair arises only when a certain endogenous pathway is first removed, this epistasis is suggestive of interference. Assays such as these would allow systematic measurement of the rate of paralog interference, even if it is rare.

Similarly, a single library could be subjected to functional selection both *in vitro* and *in vivo.* Though these experiments may never be exactly comparable, contrasting these parallel selections would illuminate how often function of a kinase-regulator pair is blocked by the presence of paralogs. Developing *in vitro* selections for two-component function is not trivial. However, cell-free transcription/translation tools have expanded dramatically in recent years, primarily for testing biosynthetic gene clusters and prototyping synthetic gene circuits - applications far more complex than characterizing two-component sensors (Fig. 3.2a).



in vitro selection in emulsion avoids interactions with endogenous HKs, RRs

Expression and repression (by GlnB) of a two component system (NtrBC) in transcription/translation media, in emulsions

Figure 3.2 *In vitro* selections for two-component system function.

(a) A schematic for how two-component libraries could be subjected to functional selections in an *in vitro* system, away from the paralogous proteins of a bacterial cytoplasm. Transcription/translation reagents, such as the commercially available PurExpress®, can translate DNA libraries into proteins and use transcriptional reporters. This can be done in oil-in-water emulsions to isolate individual DNA molecules and separately select different variants. (b) NtrBC is a well characterized two-component system from *E. coli* with no membrane-embedded components, making it ideal for *in vitro* assays. (c) Plasmid DNA containing NtrBC can drive GFP expression in PurExpress®, in oil emulsions, if supplemented with σ_{54}

and IHF. Adding inhibitor GlnB can turn off NtrB. Mutating catalytic residue D54A in NtrC also blocks activity.

This kind of assay is most tractable with histidine kinases that are entirely cytosolic, such as NtrB from *E. coli*. Plasmid containing NtrB, its cognate regulator NtrC, and a fluorescent genetic reporter are sufficient to function in the commercial transcription-translation reagent PurExpress®. NtrB is natively controlled by a repressor protein GlnB, which can be added to PurExpress® to control induction *in vitro* (Fig. 3.2b-c). This system also functions if the PurExpress® solution is emulsified in oil droplets, meaning that DNA molecules can be isolated and library selections are feasible. If a few technical hurtles are overcome, such an experiment could provide incredibly rich data on how protein interactions are constrained by paralogs and other cytosolic components.

A third option for directly measuring paralog interference is running laboratory evolution experiments in which specificity-determining regions of multiple pathways are targeted for accelerated mutation. If the interfaces of kinase-regulator complexes are forced to drift in sequence space, while selection maintains pathway function, one could explicitly determine whether two paralogs influence each other's mutational trajectories. New technologies allowing high-efficiency mutagenesis of distal genomic regions have only recently made these kinds of experiments worth considering. One class of technologies, like the Cas9-error prone polymerase fusion EvolvR, increases the mutation rate at targeted DNA sequences (Halperin et al., 2018). An alternative approach would be to use improved MAGE technologies (Kuznetsov et al., 2017) to introduce randomized codons into different regions on the genome. Approaches such as these would allow us to learn whether paralogs affect each other while drifting simultaneously through sequence space.

Conservation of specificity determining residues

When the residues important for one protein's specificity are identified, it is often assumed that this same network of residues determines the specificity for most paralogs. We do not know the bounds of this assumption. One of the limitations of residue co-variation models for predicting specificity residues is that a large diversity of proteins are required to train a single model. These models therefore cannot be used to determine if different members of a protein family rely on different amino acid positions to encode specificity. High energy contacts could exist in some regions of the interface for some paralogs, and in other regions for others. Subtle but important changes in backbone structure may exist between paralog subfamilies, altering contact points.

Historically, our tool to verify and isolate specificity residues has been low-throughput characterization of interface chimeras, such as the transplantation of residues from one kinase onto another (Skerker et al., 2008). Phosphotransfer assays are clean and conclusive, but are limited by purification bottlenecks and gel lanes (Capra et al., 2010; Podgornaia et al., 2013). The combinatorics of transplanting various permutations of putative specificity residues between different pairs of paralogs quickly breaches this experimental limit. However, these numbers are small compared to degenerate codon scanning. Transplantation libraries, built from the oligonucleotide arrays now broadly used in CRISPR screens, could provide systematic data about which positions are necessary and sufficient for encoding specificity and whether these positions are consistent across the superfamilies of histidine kinases and response regulators. Such information would be foundational for our understanding of two-component system specificity and would be extremely useful for developing better predictive models of binding.

Improved computational models for mapping sequences to specificity

Perhaps the single greatest step forward in our understanding of specificity space could come from better computational models of protein specificity. Previous applications of residue-covariance Pott's models for de-orphaning response regulators and histidine kinases tease at the potential of better *in silico* predictions (Procaccini et al., 2011). An improved ability to predict partnering directly from sequence information could offer an even richer look at how paralogs fit into sequence space than systematic mutational scanning. It would allow unprecedented *in silico* experiments measuring the compatibility of all endogenous paralogs (Bitbol et al., 2016). Additionally, computational screening of mutational variants could tell us if paralog insulation is evolutionarily robust. Thus, better specificity models would both improve our assessment of available sequence space and permit the design of large sets of protein interactions orthogonal to each other and to endogenous homologs (Jenson et al., 2018). Furthermore, modeling improvements would allow us to take advantage of the thousands of sequenced bacterial genomes to observe how large systems of paralogs permit expansion and adapt to evolutionary events like duplication and gene transfer.

Realizing the necessary *in silico* accuracy is not a trivial task, but we may see improvements as additional genomic training data become available and the overlap between machine learning and biology expands. A simple dense neural network can predict the correct kinase or regulator partner seeing only a small number of residues; its accuracy does not improve after using more than ~30 total residues from both kinase and regulator (Fig. 3.3). New model architectures, such as recurrent neural networks, could avoid using sequences in the form of multiple sequence alignments; alignment variability is currently a source of the inconsistency of Pott's models. Another limitation is that most co-variation models train exclusively on extant protein sequences, meaning there is a skew representative of phylogenetic histories and some residues are not represented at some positions. Integrating high throughput library data into training sets may improve these methods (Jenson et al., 2018), analogous to recent trends coupling oligo array libraries and Rosetta modeling (Rocklin et al., 2017). While deep mutational scanning cannot generate the diversity of extant proteins, it can offer datasets where the residue variation is constrained only to specific regions of interest for predicting specificity. The datasets generated in this thesis are heavily skewed towards negatively training data, i.e. mostly nonfunctional variant pairs, but selection-based deep mutational scanning experiments would not have this limitation.

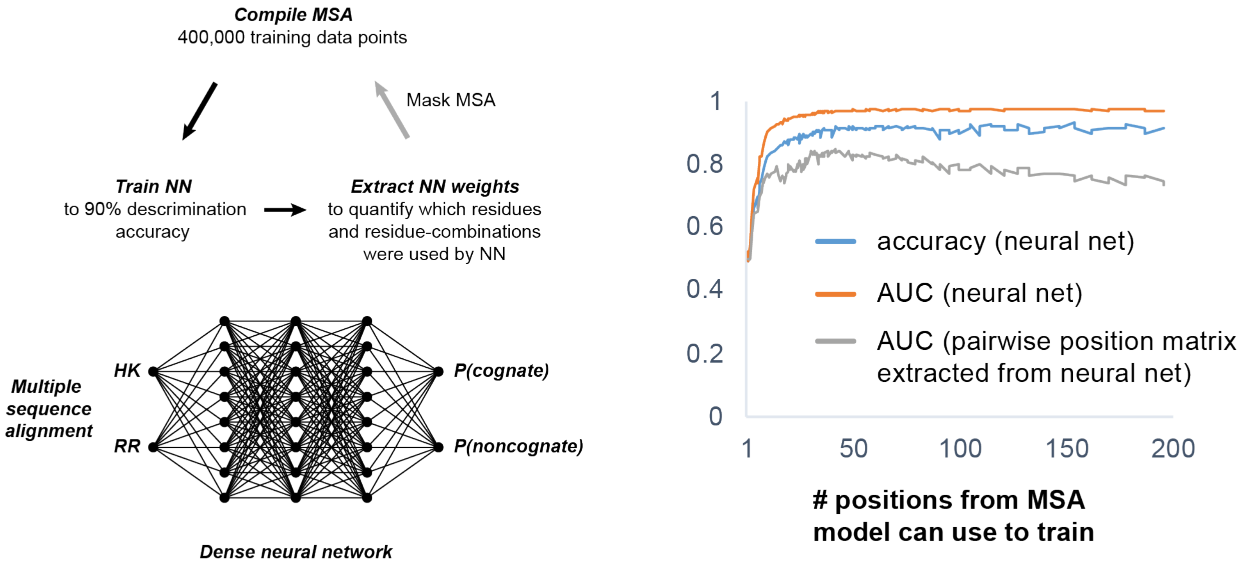


Figure 3.3 A dense neural network learns specificity from very few residues.

A dense neural network was trained on a multiple sequence alignment of HK/RR pairs from the same genomes, to classify pairs as cognate or noncognate. After training to convergence, positions from the MSA were removed if the model weights indicated they were not used. Only ~30 residues, including both kinase and regulator residues, are needed for maximum accuracy.

The problem of paralog specificity prediction is a more tractable computational challenge than generalized protein binding prediction. New techniques offering small improvements may push us over a boundary where we can learn an incredible amount about system-wide paralog interactions and evolution entirely from a genome sequence.

Expansion to other paralogous protein complexes

This work supports a model of two-component system evolution in which sequence space is not densely occupied. It is not known whether this conclusion applies to other paralogous protein complexes. Mutating the co-evolving networks of residues within other complexes would allow an assessment of this phenomenon's generalizability. Furthermore, such efforts would have significant value for cellular engineering. As biological design efforts become more complex, sets of orthogonal components have become increasingly necessary (Nielsen et al., 2013). This technique serves a single-step approach for generating large sets of orthogonal variants from a single interacting set of macromolecules, including proteins, DNAs or RNAs.

References

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell *163*, 594-606.

Alexander, J., Lim, D., Joughin, B.A., Hegemann, B., Hutchins, J.R., Ehrenberger, T., Ivins, F., Sessa, F., Hudecz, O., Nigg, E.A., *et al.* (2011). Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. Sci Signal *4*, ra42.

Alm, E., Huang, K., and Arkin, A. (2006). The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol *2*, e143.

AlQuraishi, M. (2019). AlphaFold at CASP13. Bioinformatics.

An, W., and Chin, J.W. (2009). Synthesis of orthogonal transcription-translation networks. Proc Natl Acad Sci U S A *106*, 8477-8482.

Arabidopsis Interactome Mapping, C. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Ashenberg, O., Keating, A.E., and Laub, M.T. (2013). Helix bundle loops determine whether histidine kinases autophosphorylate in cis or in trans. Journal of molecular biology *425*, 1198-1209.

Atkinson, M.R., Savageau, M.A., Myers, J.T., and Ninfa, A.J. (2003). Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli. Cell *113*, 597-607.

Baker, C.R., Hanson-Smith, V., and Johnson, A.D. (2013). Following gene duplication, paralog interference constrains transcriptional circuit evolution. Science *342*, 104-108.

Baker, C.R., Tuch, B.B., and Johnson, A.D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc Natl Acad Sci U S A *108*, 7493-7498.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061-1078.

Barnea, G., Strapps, W., Herrada, G., Berman, Y., Ong, J., Kloss, B., Axel, R., and Lee, K.J. (2008). The genetic design of signaling cascades to record receptor activation. Proc Natl Acad Sci U S A *105*, 64-69.

Bashor, C.J., Helman, N.C., Yan, S., and Lim, W.A. (2008). Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. Science *319*, 1539-1543.

Bitbol, A.F., Dwyer, R.S., Colwell, L.J., and Wingreen, N.S. (2016). Inferring interaction partners from protein sequences. Proc Natl Acad Sci U S A *113*, 12180-12185.

Bogan, A.A., and Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. J Mol Biol *280*, 1-9.

Boyken, S.E., Chen, Z., Groves, B., Langan, R.A., Oberdorfer, G., Ford, A., Gilmore, J.M., Xu, C., DiMaio, F., Pereira, J.H., *et al.* (2016). De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. Science *352*, 680-687.

Bradley, D., and Beltrao, P. (2019). Evolution of protein kinase substrate recognition at the active site. PLoS Biol *17*, e3000341.

Brentjens, R.J., Davila, M.L., Riviere, I., Park, J., Wang, X., Cowell, L.G., Bartido, S., Stefanski, J., Taylor, C., Olszewska, M., *et al.* (2013). CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. Sci Transl Med *5*, 177ra138.

Capra, E.J., and Laub, M.T. (2012). Evolution of two-component signal transduction systems. Annu Rev Microbiol *66*, 325-347.

Capra, E.J., Perchuk, B.S., Lubin, E.A., Ashenberg, O., Skerker, J.M., and Laub, M.T. (2010). Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. PLoS Genet *6*, e1001220.

Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. Cell *150*, 222-232.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Cheng, R.R., Haglund, E., Tiee, N.S., Morcos, F., Levine, H., Adams, J.A., Jennings, P.A., and Onuchic, J.N. (2018). Designing bacterial signaling interactions with coevolutionary landscapes. PLoS One *13*, e0201734.

Cheng, R.R., Morcos, F., Levine, H., and Onuchic, J.N. (2014). Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Natl Acad Sci U S A *111*, E563-571.

Cheung, J., and Hendrickson, W.A. (2010). Sensor domains of two-component regulatory systems. Curr Opin Microbiol *13*, 116-123.

Chevalier, A., Silva, D.A., Rocklin, G.J., Hicks, D.R., Vergara, R., Murapa, P., Bernard, S.M., Zhang, L., Lam, K.H., Yao, G., *et al.* (2017). Massively parallel de novo protein design for targeted therapeutics. Nature *550*, 74-79.

Chow, D.C., Rice, K., Huang, W., Atmar, R.L., and Palzkill, T. (2016). Engineering Specificity from Broad to Narrow: Design of a beta-Lactamase Inhibitory Protein (BLIP) Variant That Exclusively Binds and Detects KPC beta-Lactamase. ACS Infect Dis *2*, 969-979.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. Science *267*, 383-386.

Cong, Q., Anishchenko, I., Ovchinnikov, S., and Baker, D. (2019). Protein interaction networks revealed by proteome coevolution. Science *365*, 185-189.

Creixell, P., Palmeri, A., Miller, C.J., Lou, H.J., Santini, C.C., Nielsen, M., Turk, B.E., and Linding, R. (2015a). Unmasking determinants of specificity in the human kinome. Cell *163*, 187-201.

Creixell, P., Schoof, E.M., Simpson, C.D., Longden, J., Miller, C.J., Lou, H.J., Perryman, L., Cox, T.R., Zivanovic, N., Palmeri, A.*, et al.* (2015b). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. Cell *163*, 202-217.

Cuff, A., Redfern, O.C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A.*, et al.* (2009). The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. Structure *17*, 1051-1062.

Culviner, P.H., and Laub, M.T. (2018). Global Analysis of the E. coli Toxin MazF Reveals Widespread Cleavage of mRNA and the Inhibition of rRNA Maturation and Ribosome Biogenesis. Mol Cell *70*, 868-880 e810.

Daeffler, K.N., Galley, J.D., Sheth, R.U., Ortiz-Velez, L.C., Bibb, C.O., Shroyer, N.F., Britton, R.A., and Tabor, J.J. (2017). Engineering bacterial thiosulfate and tetrathionate sensors for detecting gut inflammation. Mol Syst Biol *13*, 923.

Dimas, R.P., Jiang, X.L., Alberto de la Paz, J., Morcos, F., and Chan, C.T.Y. (2019). Engineering repressors with coevolutionary cues facilitates toggle switches with a master reset. Nucleic Acids Res *47*, 5449-5463.

Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. Elife *7*.

Dueber, J.E., Mirsky, E.A., and Lim, W.A. (2007). Engineering synthetic signaling proteins with ultrasensitive input/output control. Nat Biotechnol *25*, 660-662.

Dutta, S., Gulla, S., Chen, T.S., Fire, E., Grant, R.A., and Keating, A.E. (2010). Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. J Mol Biol *398*, 747-762.

Elcock, A.H. (2010). Models of macromolecular crowding effects and the need for quantitative comparisons with experiment. Curr Opin Struct Biol *20*, 196-206.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat Methods *11*, 801-807.

Gao, R., and Stock, A.M. (2010). Molecular strategies for phosphorylation-mediated regulation of response regulator activity. Curr Opin Microbiol *13*, 160-167.

Gao, R., and Stock, A.M. (2015). Temporal hierarchy of gene expression mediated by transcription factor binding affinity and activation dynamics. MBio *6*, e00686-00615.

Gee, S.H., Quenneville, S., Lombardo, C.R., and Chabot, J. (2000). Single-amino acid substitutions alter the specificity and affinity of PDZ domains for their ligands. Biochemistry *39*, 14638-14646.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods *6*, 343-345.

Gooderham, W.J., Bains, M., McPhee, J.B., Wiegand, I., and Hancock, R.E. (2008). Induction by cationic antimicrobial peptides and involvement in intrinsic polymyxin and antimicrobial peptide resistance, biofilm formation, and swarming motility of PsrA in Pseudomonas aeruginosa. J Bacteriol *190*, 5624-5634.

Gray, V.E., Hause, R.J., Luebeck, J., Shendure, J., and Fowler, D.M. (2018). Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Syst *6*, 116-124 e113.

Gretes, M., Lim, D.C., de Castro, L., Jensen, S.E., Kang, S.G., Lee, K.J., and Strynadka, N.C. (2009). Insights into positive and negative requirements for protein-protein interactions by crystallographic analysis of the beta-lactamase inhibitory proteins BLIP, BLIP-I, and BLP. J Mol Biol *389*, 289-305.

Groban, E.S., Clarke, E.J., Salis, H.M., Miller, S.M., and Voigt, C.A. (2009). Kinetic buffering of cross talk between bacterial two-component sensors. J Mol Biol *390*, 380-393.

Gushchin, I., Melnikov, I., Polovinkin, V., Ishchenko, A., Yuzhakova, A., Buslaev, P., Bourenkov, G., Grudinin, S., Round, E., Balandin, T.*, et al.* (2017). Mechanism of transmembrane signaling by sensor histidine kinases. Science *356*.

Halperin, S.O., Tou, C.J., Wong, E.B., Modavi, C., Schaffer, D.V., and Dueber, J.E. (2018). CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. Nature *560*, 248-252.

Hogan, G.J., Brown, P.O., and Herschlag, D. (2015). Evolutionary Conservation and Diversification of Puf RNA Binding Proteins and Their mRNA Targets. PLoS Biol *13*, e1002307.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. Cell *149*, 1607-1621.

Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Scharfe, C.P., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. Nat Biotechnol *35*, 128-135.

Howard, P.L., Chia, M.C., Del Rizzo, S., Liu, F.F., and Pawson, T. (2003). Redirecting tyrosine kinase signaling to an apoptotic caspase pathway through chimeric adaptor proteins. Proc Natl Acad Sci U S A *100*, 11267-11272.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L.*, et al.* (2009). InterPro: the integrative protein signature database. Nucleic Acids Res *37*, D211-215.

Ingham, R.J., Colwill, K., Howard, C., Dettwiler, S., Lim, C.S., Yu, J., Hersi, K., Raaijmakers, J., Gish, G., Mbamalu, G.*, et al.* (2005). WW domains provide a platform for the assembly of multiprotein networks. Mol Cell Biol *25*, 7092-7106.

Itzkovitz, S., Tlusty, T., and Alon, U. (2006). Coding limits on the number of transcription factors. BMC Genomics *7*, 239.

Jacob-Dubuisson, F., Mechaly, A., Betton, J.M., and Antoine, R. (2018). Structural insights into the signalling mechanisms of two-component systems. Nat Rev Microbiol *16*, 585-593.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. Plos One *9*.

Janin, J., Rodier, F., Chakrabarti, P., and Bahadur, R.P. (2007). Macromolecular recognition in the Protein Data Bank. Acta Crystallogr D Biol Crystallogr *63*, 1-8.

Jenson, J.M., Xue, V., Stretz, L., Mandal, T., Reich, L.L., and Keating, A.E. (2018). Peptide design by optimization on a data-parameterized protein interaction landscape. Proc Natl Acad Sci U S A *115*, E10342-E10351.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Jones, D.T., Buchan, D.W., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184-190.

Jones, P.T., Dear, P.H., Foote, J., Neuberger, M.S., and Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature *321*, 522-525.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A *110*, 15674-15679.

Kastritis, P.L., and Bonvin, A.M. (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res *9*, 2216-2225.

Keskin, O., Ma, B., and Nussinov, R. (2005). Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol *345*, 1281-1294.

Khafizov, K., Madrid-Aliste, C., Almo, S.C., and Fiser, A. (2014). Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A *111*, 3733-3738.

Korber, B.T., Farber, R.M., Wolpert, D.H., and Lapedes, A.S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A *90*, 7176-7180.

Kuznetsov, G., Goodman, D.B., Filsinger, G.T., Landon, M., Rohland, N., Aach, J., Lajoie, M.J., and Church, G.M. (2017). Optimizing complex phenotypes through model-guided multiplex genome engineering. Genome Biol *18*, 100.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Latysheva, N.S., and Babu, M.M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res *44*, 4487-4503.

Laub, M.T., and Goulian, M. (2007). Specificity in two-component signal transduction pathways. Annu Rev Genet *41*, 121-145.

Lee, B., Schramm, A., Jagadeesan, S., and Higgs, P.I. (2010). Two-component systems and regulation of developmental progression in Myxococcus xanthus. Methods Enzymol *471*, 253-278.

Levin, K.B., Dym, O., Albeck, S., Magdassi, S., Keeble, A.H., Kleanthous, C., and Tawfik, D.S. (2009). Following evolutionary paths to protein-protein interactions with high affinity and selectivity. Nat Struct Mol Biol *16*, 1049-1055.

Levskaya, A., Chevalier, A.A., Tabor, J.J., Simpson, Z.B., Lavery, L.A., Levy, M., Davidson, E.A., Scouras, A., Ellington, A.D., Marcotte, E.M*., et al.* (2005). Synthetic biology: engineering Escherichia coli to see light. Nature *438*, 441-442.

Li, L., Shakhnovich, E.I., and Mirny, L.A. (2003). Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. Proc Natl Acad Sci U S A *100*, 4463-4468.

Lim, W.A. (2010). Designing customized cell signalling circuits. Nat Rev Mol Cell Biol *11*, 393-403.

Liu, C.C., Jewett, M.C., Chin, J.W., and Voigt, C.A. (2018a). Toward an orthogonal central dogma. Nat Chem Biol *14*, 103-106.

Liu, X., Fan, K., and Wang, W. (2004). The number of protein folds and their distribution over families in nature. Proteins *54*, 491-499.

Liu, Y., Palmedo, P., Ye, Q., Berger, B., and Peng, J. (2018b). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. Cell Syst *6*, 65-74 e63.

Long, M., VanKuren, N.W., Chen, S., and Vibranovski, M.D. (2013). New gene evolution: little did we know. Annu Rev Genet *47*, 307-333.

Marina, A., Mott, C., Auyzenberg, A., Hendrickson, W.A., and Waldburger, C.D. (2001). Structural and mutational analysis of the PhoQ histidine kinase catalytic domain. Insight into the reaction mechanism. J Biol Chem *276*, 41182-41190.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

Martin, J. (2010). Beauty is in the eye of the beholder: proteins can recognize binding sites of homologous proteins in more than one way. PLoS Comput Biol *6*, e1000821.

McKeown, A.N., Bridgham, J.T., Anderson, D.W., Murphy, M.N., Ortlund, E.A., and Thornton, J.W. (2014). Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. Cell *159*, 58-68.

Meenan, N.A., Sharma, A., Fleishman, S.J., Macdonald, C.J., Morel, B., Boetzel, R., Moore, G.R., Baker, D., and Kleanthous, C. (2010). The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc Natl Acad Sci U S A *107*, 10080-10085.

Melero, C., Ollikainen, N., Harwood, I., Karpiak, J., and Kortemme, T. (2014). Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. Proc Natl Acad Sci U S A *111*, 15426-15431.

Mimee, M., Citorik, R.J., and Lu, T.K. (2016). Microbiome therapeutics - Advances and challenges. Adv Drug Deliv Rev *105*, 44-54.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A *108*, E1293-1301.

Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2007). Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins *68*, 803-812.

Morsut, L., Roybal, K.T., Xiong, X., Gordley, R.M., Coyle, S.M., Thomson, M., and Lim, W.A. (2016). Engineering Customized Cell Sensing and Response Behaviors Using Synthetic Notch Receptors. Cell *164*, 780-791.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.-A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P.*, et al.* (2013). Precise and reliable gene expression via standard transcription and translation initiation elements. Nature methods *10*, 354-360.

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2017). Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. Proc Natl Acad Sci U S A *114*, 11703-11708.

Netzer, R., Listov, D., Lipsh, R., Dym, O., Albeck, S., Knop, O., Kleanthous, C., and Fleishman, S.J. (2018). Ultrahigh specificity in a network of computationally designed protein-interaction pairs. Nat Commun *9*, 5286.

Nicoludis, J.M., Lau, S.Y., Scharfe, C.P., Marks, D.S., Weihofen, W.A., and Gaudet, R. (2015). Structure and Sequence Analyses of Clustered Protocadherins Reveal Antiparallel Interactions that Mediate Homophilic Specificity. Structure *23*, 2087-2098.

Nielsen, A.A., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., Ross, D., Densmore, D., and Voigt, C.A. (2016). Genetic circuit design automation. Science *352*, aac7341.

Nielsen, A.A., Segall-Shapiro, T.H., and Voigt, C.A. (2013). Advances in genetic circuit design: novel biochemistries, deep part mining, and precision gene expression. Curr Opin Chem Biol *17*, 878-892.

Nocedal, I., and Johnson, A.D. (2015). How Transcription Networks Evolve and Produce Biological Novelty. Cold Spring Harb Symp Quant Biol *80*, 265-274.

Nocedal, I., Mancera, E., and Johnson, A.D. (2017). Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. Elife *6*.

O'Maille, P.E., Malone, A., Dellas, N., Andes Hess, B., Jr., Smentek, L., Sheehan, I., Greenhagen, B.T., Chappell, J., Manning, G., and Noel, J.P. (2008). Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. Nat Chem Biol *4*, 617-623.

Ohlendorf, R., Schumacher, C.H., Richter, F., and Moglich, A. (2016). Library-Aided Probing of Linker Determinants in Hybrid Photoreceptors. ACS Synth Biol *5*, 1117-1126.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife *3*, e02030.

Ovchinnikov, S., Kim, D.E., Wang, R.Y., Liu, Y., DiMaio, F., and Baker, D. (2016). Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. Proteins *84 Suppl 1*, 67-75.

Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F., and Baker, D. (2018). Protein structure prediction using Rosetta in CASP12. Proteins *86 Suppl 1*, 113-121.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. Science *355*, 294-298.

Park, S.H., Zarrinpar, A., and Lim, W.A. (2003). Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. Science *299*, 1061-1064.

Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science *300*, 445-452.

Perez, J.C., Fordyce, P.M., Lohse, M.B., Hanson-Smith, V., DeRisi, J.L., and Johnson, A.D. (2014). How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes. Genes Dev *28*, 1272-1277.

Plach, M.G., Semmelmann, F., Busch, F., Busch, M., Heizinger, L., Wysocki, V.H., Merkl, R., and Sterner, R. (2017). Evolutionary diversification of protein-protein interactions by interface add-ons. Proc Natl Acad Sci U S A *114*, E8333-E8342.

Podgornaia, A.I., Casino, P., Marina, A., and Laub, M.T. (2013). Structural basis of a rationally rewired protein-protein interface critical to bacterial signaling. Structure *21*, 1636-1647.

Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science *347*, 673-677.

Pougach, K., Voet, A., Kondrashov, F.A., Voordeckers, K., Christiaens, J.F., Baying, B., Benes, V., Sakai, R., Aerts, J., Zhu, B.*, et al.* (2014). Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nat Commun *5*, 4868.

Procaccini, A., Lunt, B., Szurmant, H., Hwa, T., and Weigt, M. (2011). Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. PLoS One *6*, e19729.

Ravikumar, A., Arrieta, A., and Liu, C.C. (2014). An orthogonal DNA replication system in yeast. Nat Chem Biol *10*, 175-177.

Ravikumar, A., Arzumanyan, G.A., Obadi, M.K.A., Javanpour, A.A., and Liu, C.C. (2018). Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. Cell *175*, 1946-1957 e1913.

Reinke, A.W., Baek, J., Ashenberg, O., and Keating, A.E. (2013). Networks of bZIP protein-protein interactions diversified over a billion years of evolution. Science *340*, 730-734.

Reinke, A.W., Grant, R.A., and Keating, A.E. (2010). A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. J Am Chem Soc *132*, 6025-6031.

Richardson, J.S., and Richardson, D.C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci U S A *99*, 2754-2759.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat Methods *15*, 816-822.

Riglar, D.T., Giessen, T.W., Baym, M., Kerns, S.J., Niederhuber, M.J., Bronson, R.T., Kotula, J.W., Gerber, G.K., Way, J.C., and Silver, P.A. (2017). Engineered bacteria can function in the mammalian gut long-term as live diagnostics of inflammation. Nat Biotechnol *35*, 653-658.

Riglar, D.T., and Silver, P.A. (2018). Engineering bacteria for diagnostic and therapeutic applications. Nat Rev Microbiol *16*, 214-225.

Rocklin, G.J., Chidyausiku, T.M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V.K., Chevalier, A*., et al.* (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. Science *357*, 168-175.

Rollins, N.J., Brock, K.P., Poelwijk, F.J., Stiffler, M.A., Gauthier, N.P., Sander, C., and Marks, D.S. (2019). Inferring protein 3D structure from deep mutation scans. Nat Genet *51*, 1170-1176.

Rowland, M.A., and Deeds, E.J. (2014). Crosstalk and the evolution of specificity in two-component signaling. Proc Natl Acad Sci U S A *111*, 5550-5555.

Saita, E., Abriata, L.A., Tsai, Y.T., Trajtenberg, F., Lemmin, T., Buschiazzo, A., Dal Peraro, M., de Mendoza, D., and Albanesi, D. (2015). A coiled coil switch mediates cold sensing by the thermosensory protein DesK. Mol Microbiol *98*, 258-271.

Sammond, D.W., Eletr, Z.M., Purbeck, C., and Kuhlman, B. (2010). Computational design of second-site suppressor mutations at protein-protein interfaces. Proteins *78*, 1055-1065.

Schmidl, S.R., Ekness, F., Sofjan, K., Daeffler, K.N., Brink, K.R., Landry, B.P., Gerhardt, K.P., Dyulgyarov, N., Sheth, R.U., and Tabor, J.J. (2019). Rewiring bacterial two-component systems by modular DNA-binding domain swapping. Nat Chem Biol *15*, 690-698.

Schmidl, S.R., Sheth, R.U., Wu, A., and Tabor, J.J. (2014). Refactoring and optimization of light-switchable Escherichia coli two-component systems. ACS Synth Biol *3*, 820-831.

Schmiedel, J.M., and Lehner, B. (2019). Determining protein structures using deep mutagenesis. Nat Genet *51*, 1177-1186.

Schreiber, G., and Keating, A.E. (2011). Protein binding specificity versus promiscuity. Curr Opin Struct Biol *21*, 50-61.

Segall-Shapiro, T.H., Meyer, A.J., Ellington, A.D., Sontag, E.D., and Voigt, C.A. (2014). A 'resource allocator' for transcription based on a highly fragmented T7 RNA polymerase. Mol Syst Biol *10*, 742.

Shrestha, R., Garrett, S.C., Almo, S.C., and Fiser, A. (2019). Computational Redesign of PD-1 Interface for PD-L1 Ligand Selectivity. Structure *27*, 829-836 e823.

Sibener, L.V., Fernandes, R.A., Kolawole, E.M., Carbone, C.B., Liu, F., McAffee, D., Birnbaum, M.E., Yang, X., Su, L.F., Yu, W*., et al.* (2018). Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding. Cell *174*, 672-687 e627.

Siryaporn, A., and Goulian, M. (2008). Cross-talk suppression between the CpxA-CpxR and EnvZ-OmpR two-component systems in E. coli. Mol Microbiol *70*, 494-506.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Sockolosky, J.T., Trotta, E., Parisi, G., Picton, L., Su, L.L., Le, A.C., Chhabra, A., Silveria, S.L., George, B.M., King, I.C.*, et al.* (2018). Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. Science *359*, 1037-1042.

Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories in the sequence space of an ancient protein. Nature *549*, 409-413.

Stein, R.R., Marks, D.S., and Sander, C. (2015). Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. PLoS Comput Biol *11*, e1004182.

Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. Science *317*, 364-369.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-component signal transduction. Annu Rev Biochem *69*, 183-215.

Sunden, F., Peck, A., Salzman, J., Ressl, S., and Herschlag, D. (2015). Extensive site-directed mutagenesis reveals interconnected functional units in the alkaline phosphatase active site. Elife *4*.

Suzuki, T., Miwa, K., Ishikawa, K., Yamada, H., Aiba, H., and Mizuno, T. (2001). The Arabidopsis sensor His-kinase, AHk4, can respond to cytokinins. Plant Cell Physiol *42*, 107-113.

Tabor, J.J., Levskaya, A., and Voigt, C.A. (2011). Multichromatic control of gene expression in Escherichia coli. J Mol Biol *405*, 315-324.

Tan, C.S., Pasculescu, A., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2009). Positive selection of tyrosine loss in metazoan evolution. Science *325*, 1686-1688.

Thompson, K.E., Bashor, C.J., Lim, W.A., and Keating, A.E. (2012). SYNZIP protein interaction toolbox: in vitro and in vivo specifications of heterospecific coiled-coil interaction domains. ACS Synth Biol *1*, 118-129.

Trajtenberg, F., Imelio, J.A., Machado, M.R., Larrieux, N., Marti, M.A., Obal, G., Mechaly, A.E., and Buschiazzo, A. (2016). Regulation of signaling directionality revealed by 3D snapshots of a kinase:regulator complex in action. Elife *5*.

UniProt, C. (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Wang, B., Zhao, A., Novick, R.P., and Muir, T.W. (2014). Activation and inhibition of the receptor histidine kinase AgrC occurs through opposite helical transduction motions. Mol Cell *53*, 929-940.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol *13*, e1005324.

Wanner, B.L., and Wilmes-Riesenberg, M.R. (1992). Involvement of phosphotransacetylase, acetate kinase, and acetyl phosphate synthesis in control of the phosphate regulon in Escherichia coli. J Bacteriol *174*, 2124-2130.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A *106*, 67-72.

Whitaker, W.R., Davis, S.A., Arkin, A.P., and Dueber, J.E. (2012). Engineering robust control of two-component system phosphotransfer using modular scaffolds. Proc Natl Acad Sci U S A *109*, 18090-18095.

Willett, J.W., Tiwari, N., Muller, S., Hummels, K.R., Houtman, J.C., Fuentes, E.J., and Kirby, J.R. (2013). Specificity residues determine binding affinity for two-component signal transduction systems. MBio *4*, e00420-00413.

Wojtowicz, W.M., Flanagan, J.J., Millard, S.S., Zipursky, S.L., and Clemens, J.C. (2004). Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. Cell *118*, 619-633.

Wojtowicz, W.M., Wu, W., Andre, I., Qian, B., Baker, D., and Zipursky, S.L. (2007). A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. Cell *130*, 1134-1145.

Yamada, H., Suzuki, T., Terada, K., Takei, K., Ishikawa, K., Miwa, K., Yamashino, T., and Mizuno, T. (2001). The Arabidopsis AHK4 histidine kinase is a cytokinin-binding receptor that transduces cytokinin signals across the membrane. Plant Cell Physiol *42*, 1017-1023.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., *et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676-680.

Zhang, Z., and Palzkill, T. (2004). Dissecting the protein-protein interface between beta-lactamase inhibitory protein and class A beta-lactamases. J Biol Chem *279*, 42860-42866.

Zimmerman, S.B., and Trach, S.O. (1991). Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. J Mol Biol *222*, 599-620.