Quantitative Biology 2018, 6(1): 68–84
https://doi.org/10.1007/s40484-018-0130-0

RESEARCH ARTICLE

Deciphering the protein-DNA code of bacterial winged helix-turn-helix transcription factors

Adam P. Joyce[1] and James J. Havranek[2,*]

[1] Program in Developmental, Regenerative, and Stem Cell Biology, Washington University in St. Louis, St. Louis, MO 63110, USA
[2] Department of Biochemistry and Molecular Biophysics, Washington University in St. Louis, St. Louis, MO 63110, USA
* Correspondence: havranek@biochem.wustl.edu

Received May 12, 2017; Revised July 14, 2017; Accepted July 24, 2017

***Background*: Sequence-specific binding by transcription factors (TFs) plays a significant role in the selection and regulation of target genes. At the protein:DNA interface, amino acid side-chains construct a diverse physicochemical network of specific and non-specific interactions, and seemingly subtle changes in amino acid identity at certain positions may dramatically impact TF:DNA binding. Variation of these specificity-determining residues (SDRs) is a major mechanism of functional divergence between TFs with strong structural or sequence homology.**

***Methods*: In this study, we employed a combination of high-throughput specificity profiling by SELEX and Spec-seq, structural modeling, and evolutionary analysis to probe the binding preferences of winged helix-turn-helix TFs belonging to the OmpR sub-family in *Escherichia coli*.**

***Results*: We found that *E. coli* OmpR paralogs recognize tandem, variably spaced repeats composed of "GT-A" or "GCT"-containing half-sites. Some divergent sequence preferences observed within the "GT-A" mode correlate with amino acid similarity; conversely, "GCT"-based motifs were observed for a subset of paralogs with low sequence homology. Direct specificity profiling of a subset of OmpR homologues (CpxR, RstA, and OmpR) as well as predicted "SDR-swap" variants revealed that individual SDRs may impact sequence preferences locally through direct contact with DNA bases or distally via the DNA backbone.**

***Conclusions*: Overall, our work provides evidence for a common structural code for sequence-specific wHTH:DNA interactions, and demonstrates that surprisingly modest residue changes can enable recognition of highly divergent sequence motifs. Further examination of SDR predictions will likely reveal additional mechanisms controlling the evolutionary divergence of this important class of transcriptional regulators.**

Keywords: transcription factor; SELEX; winged helix-turn-helix; specificity determinants; two-component signaling

Author summary: Although many transcription factors (TFs) possess high sequence similarity, subtle amino acid variation at DNA-contacting positions can yield substantial (and difficult to predict) alterations to intrinsic recognition potential. In this work, we characterized the natural variation in recognition potential (base preference, monomer spacing, and monomer orientation) within a sub-family of *E. coli* winged helix-turn-helix TFs. Using patterns of amino acid conservation, we further predicted a number of amino acids with likely involvement in specificity determination between these related TFs. Finally, we demonstrated the complex local and global roles of predicted SDRs as well as protein sequence context on sequence-specific binding.

INTRODUCTION

All organisms employ signal transduction systems that receive, interprets, and ultimately determine the appropriate physiological response to environmental conditions, available metabolites, and real-time activity of other cells. Most prokaryotic signaling inputs are transduced via two-component signaling pathways (TCSPs), comprised of a transmembrane sensor histidine kinase (HK) and a cytoplasmic response regulator (RR). Typically,

detection of an appropriate stimulus by extracellular sensory domains of a HK results in the downstream phosphorylation of a highly conserved aspartate residue in the "receiver" domain of the cognate RR, altering its behavior [1]. The high degree of structural homology among two-component proteins operating in the same cellular space raises the potential for cross-talk between pathways, which a number of mechanisms have arisen to suppress [2]. For example, non-activated HKs are known to rapidly dephosphorylate their cognate RRs, thereby reducing spurious RR phosphorylation by activated, non-cognate HKs. This process, termed "kinetic buffering", combined with molecular specificity determinants at the HK:RR and RR:RR interfaces [3,4], ensure that signal is transduced through each invidual pathway with high fidelity.

The majority of bacterial RRs contain one or more "effector" domains that diversify the functional output of TCSPs. Sequence-specific DNA-binding domains (DBDs) are by far the most common effector class, and, as such, the majority of TCSPs control the transcriptional output of condition-specific genes and operons [5]. The winged helix-turn-helix (wHTH), a fold present in approximately 50% of bacterial transcription factors (TFs), is the most prevalent effector found in all bacterial RRs (~30%) [5,6]. The wHTH-containing RRs are collectively sub-classified by structural homology to the prototypical osmolarity response protein OmpR, and share certain functional characteristics with it [7,8]. Conventionally, phospho-activation of an OmpR-family RR shifts the monomer-dimer equilibrium toward a predominantly homodimeric state, which co-orients DBDs and promotes cooperative binding at tandem repeat sequences [3,9,10]. Phosphorylation of OmpR-family RRs may also stimulate the coordinated occupancy of multiple adjacent binding sites, potentially leading to complex regulatory outcomes. For example, OmpR-dependent enhancers at multiple porin genes (e.g., *ompF, ompC*, and *ompS1*) exhibit distinct binding affinity and regulatory activity, and their unique, enhancer-specific sensitivity to OmpR mutations further suggests that the conformation of enhancer-bound OmpR complexes is sequence-dependent [11–13]. Although biochemical studies of isolated RR_{DBD} show little evidence of direct (DBD:DBD) cooperativity [14], flexible interactions between regulatory domains [15] and DNA-mediated allostery [16] could both play a role.

Until recently, the sequence preferences of OmpR family members have primarily been defined using consensus-based models based on a small number of experimentally validated operator sequences (e.g., QseB [17]). Larger collections of binding sites generated using genomic SELEX [18–21], ChIP-seq [22,23], and gene expression analysis [24,25] have greatly increased the quantitative resolution of specificity models. However, the DNA sequences of native genomic binding sites are not selected solely for optimal affinity, so these models may not accurately reflect energetic preferences. For example, OmpR engages in highly cooperative, concerted binding to low-affinity sequence variants on the consensus "TGTAACAAAATGTTTC" to carry out staged and signal-dependent transcriptional processes [20,26]. *E. coli* CpxR, an OmpR paralog, binds targets *in vivo* that are enriched for repeats conforming to the similar consensus "$GTAAAN_5GTAAA$" [25]; likewise, *E. coli* RstA binds two targets — identified through genomic SELEX — containing " TGTAACNANATGTA" sequences [19]. As expected, OmpR, CpxR, and RstA bind many of the same genomic targets, the sequences of which will reflect their synergistic and antagonistic regulatory outcomes [27,28]. *Klebsiella* CpxR has additionally been shown to bind two different classes of target sites unique in sequence, location relative to transcription start sites, and phospho-sensitivity, suggesting that still more complex regulatory constraints influence genomic occupancy [29]. Therefore, specificity models built using native targets for OmpR-family RRs will inaccurately conflate energetic preference with regulatory requirements.

To obtain accurate energetic models of RR:DNA specificity, some OmpR family members have been profiled *in vitro*, and their comparison to specificity models derived from genomic sequences provides insight into the evolution of strategies for TCSP-mediated gene regulation. For example, sequence-specific binding by the *E. coli* OmpR paralog ArcA was determined systematically *in vitro* over all single-base mutations to a native operator sequence, yielding a specific binding model conforming to the consensus "$tGTTAN_6tGTTA$" [30]. ArcA was subsequently shown to target genomic regions containing the identical tandem repeat "$TGTTAN_5TGTTA$" (distributed in phase with the DNA double helix) [22], and further analysis demonstrated that both site affinity and site clustering govern the magnitude and phospho-activation threshold of ArcA-dependent gene expression [31]. In *Burkholdia*, KdpE prefers a similar A/T-rich binding motif "TTTTTANA" *in vitro* with relatively low specificity [32], strongly resembling a binding motif derived from a collection of sites identified through comparative genomics [14]. Despite its relatively low sequence-specificity, an *in vitro* survey of binding sites in the *E. coli* genome revealed only two high-affinity bind-ing sequences, indicating a significant role for cooperative assembly in physiologically relevant target selection [18]. Based on this same study, a set of genomic targets identified for phosphorylated BasR were explored using DNase footprinting, and found to contain "TTAAnnTT" repeats with spacing identical to those for ArcA and KdpE [21]. The roles of potentially subtle base

preferences, sequence context, and protein-protein interactions in site discrimination between these paralogs remain unresolved.

In contrast to the large number of studies detailing site recognition by OmpR-family RRs, few studies have systematically probed the residues responsible for divergent DNA-binding characteristics. Five crystal structures in complex with high-affinity oligonucleotides demonstrate clearly that two regions of the wHTH, the "recognition helix" (RH) and "wing" (W), make up the primary protein:DNA interface [8,10,33–36]. Numerous residues within these domains make diverse specific and non-specific contacts with the DNA helix. For example, the *E. coli* PhoB wing residue Arg^{219} projects into the DNA minor groove, making several contacts with an A-rich tract [10]. This residue was hypothesized to facilitate minor groove compression and support global curvature of the bound DNA toward the protein complex. In *E. coli* OmpR, mutation of the residue Val^{203} to Met inverts the relative affinity of OmpR for two native targets differing by a single base that, in the PhoB crystal structure, forms a specific contact with the analogous residue in the major groove [37]. In *Bacillus* WalR, mutation of multiple different RH residues oriented toward the major groove dramatically reduced DNA-binding affinity, suggesting a central role for specific and/or non-specific interactions in this region. A gain in non-specific binding affinity was observed following the conversion of a DNA-adjacent Asp to Arg, which, presumably, was able to contact the DNA backbone [38]. Taken together, current evidence suggests that residues across the RR:DNA interface (or elsewhere in the protein [15]) can contribute specific and/or non-specific binding affinity, but the role of individual residues in paralog-specific sequence recognition remains poorly understood.

In this study we deeply characterize sequence-specific binding by eleven OmpR paralogs, and we identify residues that confer paralog-specific binding attributes. We find that these *E. coli* response regulators (eRRs) are capable of multiple modes of sequence-specific DNA binding, but as a family conform largely to a "canonical", dimeric model with divergent preferences for the sequence, spacing, and orientation of composite half-sites. We identify a small cluster of co-varying residues in the RH with potential involvement in paralog-specific sequence recognition, and demonstrate that two residues, both of which contact the same DNA base, are partially capable of specificity inter-conversion between OmpR paralogs. These residues appear to carry out distinct roles in DNA sequence and shape recognition, and structural context (i.e., proximal residues) may influence their activity. Overall, our representative specificity profile of the *E. coli* OmpR family suggests that a minority of eRRs are capable of discriminating binding sites *in vivo* solely by sequence preferences. As a result, specificity models derived from physiologically relevant sites may deviate from the energetic optimum at key positions in the binding motif. In physiological terms, this indicates a complex balance between intrinsic affinity, site-specific regulatory functions, and paralog-specificity. The structural conclusions and techniques presented will be useful in future studies of protein:DNA specificity involving OmpR-family RRs and other families of wHTH TFs.

RESULTS

Variation in protein and DNA structure at the wHTH:DNA interface

Using the PFAM database [39], we identified 18 putative eRRs (PF00486) present in the *E. coli* K12 genome, 14 of which fall into the archetypical class of bipartite, signal-activated transcriptional response regulators. They exhibit high similarity at residue positions presented toward the domain core as well as those in contact with the DNA phosphate backbone, implying the preservation of both wHTH fold structure and DNA-binding ability anticipated from prior structural and functional studies of OmpR family proteins (Figure 1A) [40]. The protein:DNA interface spans three structural elements within the wHTH domain (α1 N-terminus, α1; beta strands β6-7, W for "wing"; and α3, RH for "recognition helix"), which contact different regions of the double helix (Figure 1B). The wHTH also contains a transactivation loop (TA), which is involved in gene regulation through interactions with RNA polymerase [41].

At the protein:DNA interface, highly conserved residues often interact non-specifically with DNA or make sequence-specific contacts important for "familial" binding specificity [42], whereas residues conserved within sub-groups are more likely to act as specificity determinants between paralogous proteins [43]. We identified 1,925 high-confidence orthologs (see Methods) for our set of 14 eRRs, which we further subdivided into four distinct lineages (LI–IV) based on full-length amino acid sequence similarity (Supplementary Figure S1). We immediately observed multiple positions in the DNA-contacting sub-domains that exhibited patterns of conservation consistent with paralog-specific functions, especially at the DNA-exposed surface of the RH (Figure 1C). [For consistency, we will hereafter reference these positions by their domain context, numbered position, and residue identity where applicable; e.g., $RH_{(12)}$[R] denotes an arginine residue at position 12 of the recognition helix.] Residue dyads Arg-Asp (LII, LIII, and LIV) and Asn-Glu (LI) were frequently observed at $RH_{(2,5)}$, indicating potential for a coevolutionary relationship. Based on available crystal structures, these residues

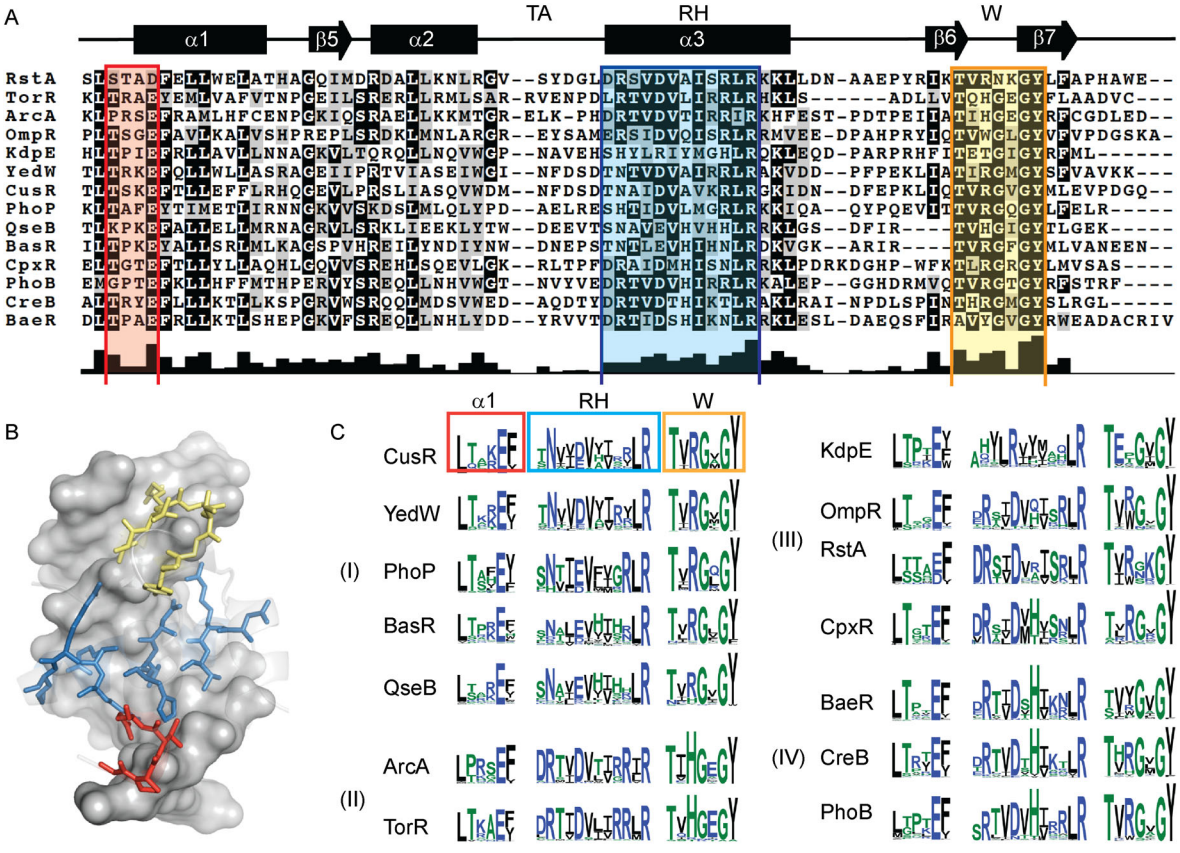

Figure 1. Diverse residue contacts and DNA shape at the protein:DNA interface for OmpR family response regulators. (A) The results of a structural alignment of winged helix-turn-helix domains for *E. coli* K12 OmpR homologues are presented. Highly and moderately conserved residues are highlighted by black and grey boxes, respectively, and a histogram of relative entropy is plotted for each position (bottom) from a sampling of 1000 proteins from OmpR homologues identified previously [5]. Colored regions and labels (recognition helix (RH, blue), transactivation loop (TA), α1 (red), and wing (W, yellow)) indicate structural sub-domains that contact DNA in representative co-crystal structures of OmpR homologues bound to target DNA sequences. (B) DNA-contacting residues in the RH, W, and α1 are shown using the crystal structure of a single PhoB monomer bound to a high-affinity half-site (PDB code: 1GXP [10]). Residues are rendered as sticks and colored in correspondence with the alignment in panel A; DNA is shown in a grey surface representation. (C) Residue conservation within the RH is displayed in sequence logo format, organized into lineages (I–IV) based on amino acid similarity over the full-length protein.

directly interact at the protein:DNA interface, and suggest that $RH_{(5)}$[D/E] serves as a hub of polar interactions between $RH_{(2)}$[R/N], $W_{(7)}$[Y], as well as certain backbone-proximal residues at the N-terminus of α2. Planar residues are frequently observed at $RH_{(7)}$, with a strong His prevalence in LIV proteins and the LIII protein CpxR; non-valine residues at $RH_{(6)}$ appear also to co-occur with $RH_{(7)}$[H]. Residues at $RH_{(9,10)}$ were typically conserved at lower levels than other DNA-contacting residues, but some trends were apparent at the lineage level, such as the preference for $RH_{(9)}$[S] in LIII.

Because multiple residues with paralog- and lineage-specific distributions are positioned to interact primarily with the DNA backbone, we performed a comparative structural analysis to investigate the role of shape-specificity in RR-DNA binding. As previously reported, the DNA minor groove narrows substantially in the spacer region in all five protein:DNA structures, favoring DNA curvature toward the bound face of the dimer [10]. We further observed that the DNA major groove expanded in the region occupied by the RH (Supplementary Figure S2A). We then superimposed RHs to visualize the relative position of the DNA helix, and found that phosphate backbone trajectories diverge strand-specifically directly over each occupied half-site (Supplementary Figure S2B). For each RR, backbone trajectory was similar between upstream and downstream half-sites, suggesting that structural variability is primarily dependent on protein binding, not underlying DNA sequence (Supplementary Figure S3).

Multi-specificity in sequence recognition by eRRs

To systematically explore the intrinsic sequence recognition potential of the OmpR family, we constructed a randomized (20N) library and performed high-throughput SELEX on phosphorylated and non-phosphorylated eRRs. Binding motifs were identified *de novo* using BioProspector [44] for five proteins (KdpE, BasR, QseB, BaeR, and OmpR), revealing two apparent modes of binding (Figure 2). One mode, characterized by a half-site based on the consensus "GT-A", was enriched following selection with CpxR (LIII), OmpR (LIII), and QseB (LI); however, the three "GT-A"-binding eRRs exhibited different responses to chemical phosphorylation. OmpR exhibited near-identical binding specificity in both phosphorylation states to an apparent tandem repeat; however, the proportional representation of sequences in the selected pools was substantially higher under phosphorylating conditions. We can infer from the increased level of enrichment that phospho-OmpR exhibits enhanced affinity toward DNA, most likely through stabilization of the homodimeric complex [13], resulting in a greater number of sequences stably bound (and selected) in each successive round. Phosphorylated and non-phosphorylated CpxR, by contrast, yielded *de novo* binding motifs consistent with direct and inverted repeat architectures, respectively. The recognition of inverted repeats has previously been suggested at genomic binding sites for the copper-responsive RRs, CusR and YedW [47]; however, there is currently no corroborating structural evidence. Additionally, because these particular assays cannot differentiate between complexes of distinct molecular weight, inverted repeat motifs for CpxR, CusR, and YedW are consistent with both (1) an inverted dimer (→←) and (2) a pair of head-to-tail dimers (→→ ←←). Surprisingly, a sequence repeat containing a novel "GCT" core was enriched in selections using QseB, KdpE, and BaeR. Half-site recognition was asymmetric ("ACGCTN$_4$TTGCT"), with preferential specificity toward the upstream and downstream sites in the presence and absence of phosphodonor, respectively.

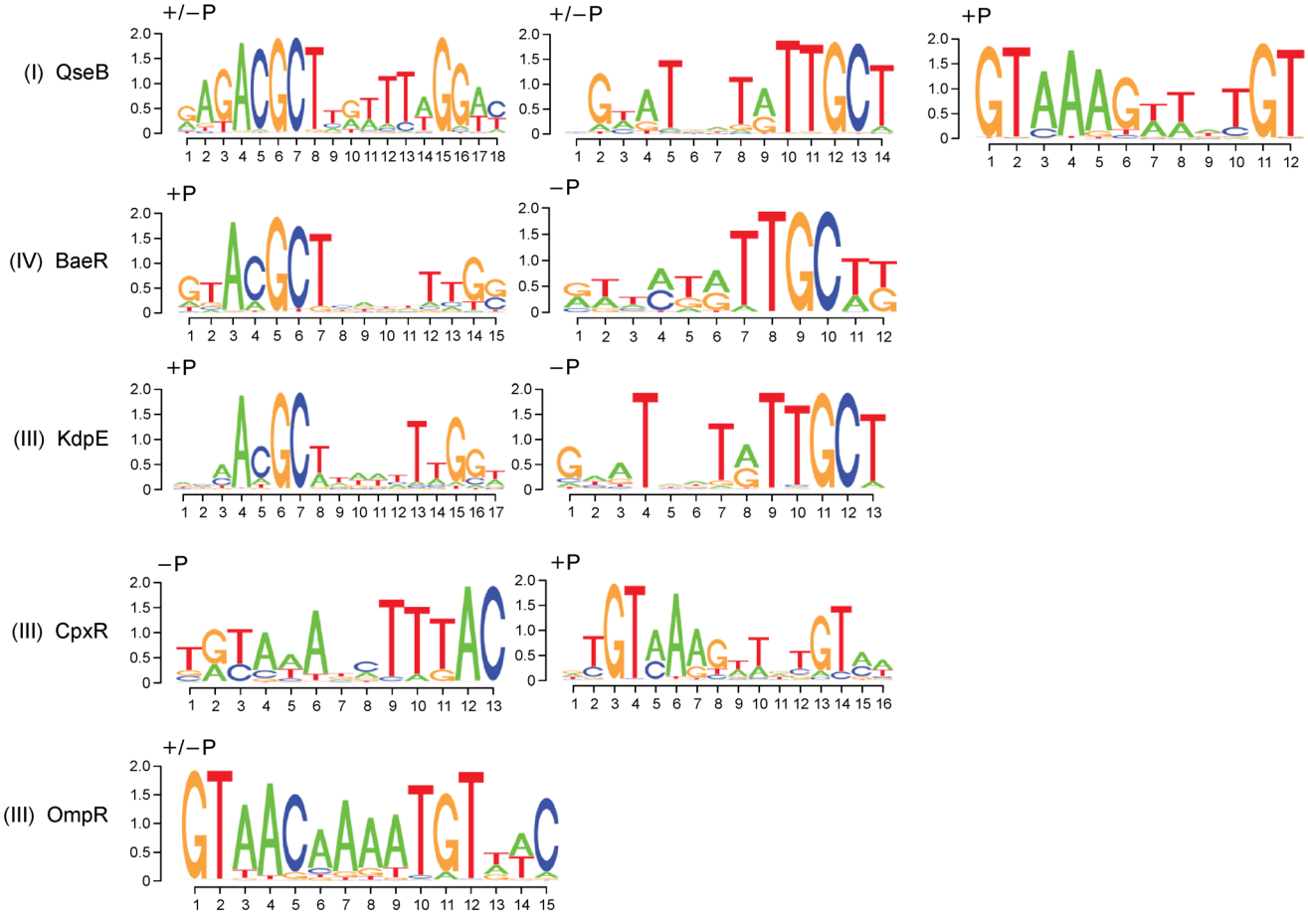

Figure 2. Lineage-independent multi-specificity of eRRs. DNA sequence logos [45,46] are derived from *de novo* motif searching of SELEX pools using BioProspector [44]. Binding motifs were discovered following selection of the indicated eRR in the presence (+P) or absence (−P) of phosphodonor, or are representative of the same motif identified in both conditions independently (+/−P).

eRRs vary in their preference for half-site sequence, spacing, and orientation

The "GT-A" repeat sequences identified through SELEX bore similarity to binding sequences previously derived from *in vitro* and *in vivo* analyses of OmpR ("TGTAA-CAAAATGTTTC") [20], CpxR ("GTAA(N_6)GTAA") [25], RstA ("GTA"/"GTAAC") [19], PhoP ("TGTTta") [48], PhoB ("TGTCA") [23], and ArcA ("TGTTA") [30]. We expected that this could represent a familial mode of binding specific to the OmpR family as described for other TF families [42], but hypothesized that divergence could arise in subtle sequence, spacing, and orientation preferences. We performed three rounds of SELEX on each paralog (+/− phosphodonor) using a partially randomized library flanked on one side by a synthetic half-site, ("AG**GTAA**($N)_{20}$"). Binding motifs (each representing thousands of individual sequences) were identified *de novo* for eight eRRs in their phosphorylated and non-phosphorylated states (Figure 3A). Despite differences in sequence and regulatory function, the OmpR family overall displays a consistent preference for half-sites of the form "$t_{(+1)}$GTnAn$_{(+6)}$" (on the reverse strand, "$n_{(-6)}$TnACa$_{(-1)}$"), hereafter referred to as the "canonical" motif (Figure 3B, left). In general, half-site sequences varied over the profiled TFs, but mainly at the weakly selective fourth and sixth positions. CreB was a notable exception, adopting a preference for a $G_{(+4)}$, highly similar to the "cre tag" sequence previously observed in promoters of several CreBC TCSP targets [49] (Figure 3B, middle); interestingly, PhoB, an LIV protein, exhibited a similar, weaker preference for $G_{(+4)}$.

To investigate spacing and orientation preferences for each eRR, we asked whether "hits" to representative half-site position-weight matrices (PWMs) were over-represented at specific positions (in forward or reverse orientation) within the randomized region for sequences in the selected pools. Overall, the consistent orientation of putative binding events relative to the synthetic half-site suggested that recognition of direct repeats is a familial characteristic, and center-to-center spacing preferences between half-sites ranged only narrowly from 9 bp to 10 bp. A few notable exceptions to this rule include i) strong binding of CusR (L1) to a head-to-head inverted repeat, ii) a lack of spacing preference for the non-phosphorylated form of KdpE, and iii) an atypical pattern of spacing for CpxR. Importantly, these exceptions reflect known binding activities of these factors *in vivo* [14,28,47].

The predominance of canonical and non-canonical binding preferences differed between SELEX platforms for KdpE (for which data were obtained in both experiments), an outcome that was intended from the inclusion of the synthetic "GTAA" half-site. CpxR similarly failed to reproduce its tail-to-tail inverted repeat architecture, but did exhibit non-canonical spacing out of phase with the DNA helix. The loss of non-canonical preferences when using a biased SELEX library is unsurprising, given the reduced search space in which to encounter high-affinity sequences. However, the emergence of an apparently novel CpxR:DNA assembly was unexpected, and was, presumably, in some way related to the inclusion of a synthetic half-site. Given the complex, oligomeric binding properties of CpxR at its regulatory targets [28,29,50], we sought to further characterize the LIII lineage.

LIII specificity determinants include sequence preference and complex assembly

The LIII family members CpxR, OmpR, and RstA target overlapping and/or identical operators in the *E. coli* genome, although they exhibited asymmetric binding preferences and/or complex assembly patterns in a SELEX format. To more deeply characterize subtle specificity determinants within this restricted lineage, we performed Spec-seq, a technique permitting the measurement of relative affinities toward thousands of individual sequences while visualizing protein:DNA assembly in an EMSA format [51,52]. We designed a partially randomized library based on the high-affinity OmpR site ("AATGTAACAAAATGTTTCA"), which was similar to SELEX-derived consensus sequences for CpxR and RstA. Each base pair was biased towards the consensus during synthesis (85% native base, with each alternate base present at 5%) to produce a complex library of targeted variants (~5.4% consensus, 4.7% single-variants, and 1.7% double-variants). This design permitted a broad sampling of sequences with greater similarity to the RstA or CpxR preferred binding sites, albeit at lower frequency in the pool. To generate sequence libraries appropriately targeted to the intended eRR, for each protein we selected high-affinity sequences from the library over three rounds (without replacement) and combined the resulting sequence pools in a 3:2:1 ratio in ascending order of affinity.

Strikingly, we observed distinct banding patterns for all three proteins, indicating alternative mechanisms of assembly and/or complex structure. OmpR formed a single specie that migrated identically in both phosphorylation states, consistent with previous findings *in vivo* and *in vitro* that it minimally requires a homodimer for target recognition (Figure 4D,E, upper). Phospho-CpxR failed to form discrete complexes, but rather shifted the population continuously in proportion to the total protein concentration, producing a prominent "smear" with distinct lower and upper bounds consistent with dimeric and tetrameric assembly states (Figure 4A), upper. Interestingly, non-phosphorylated CpxR also

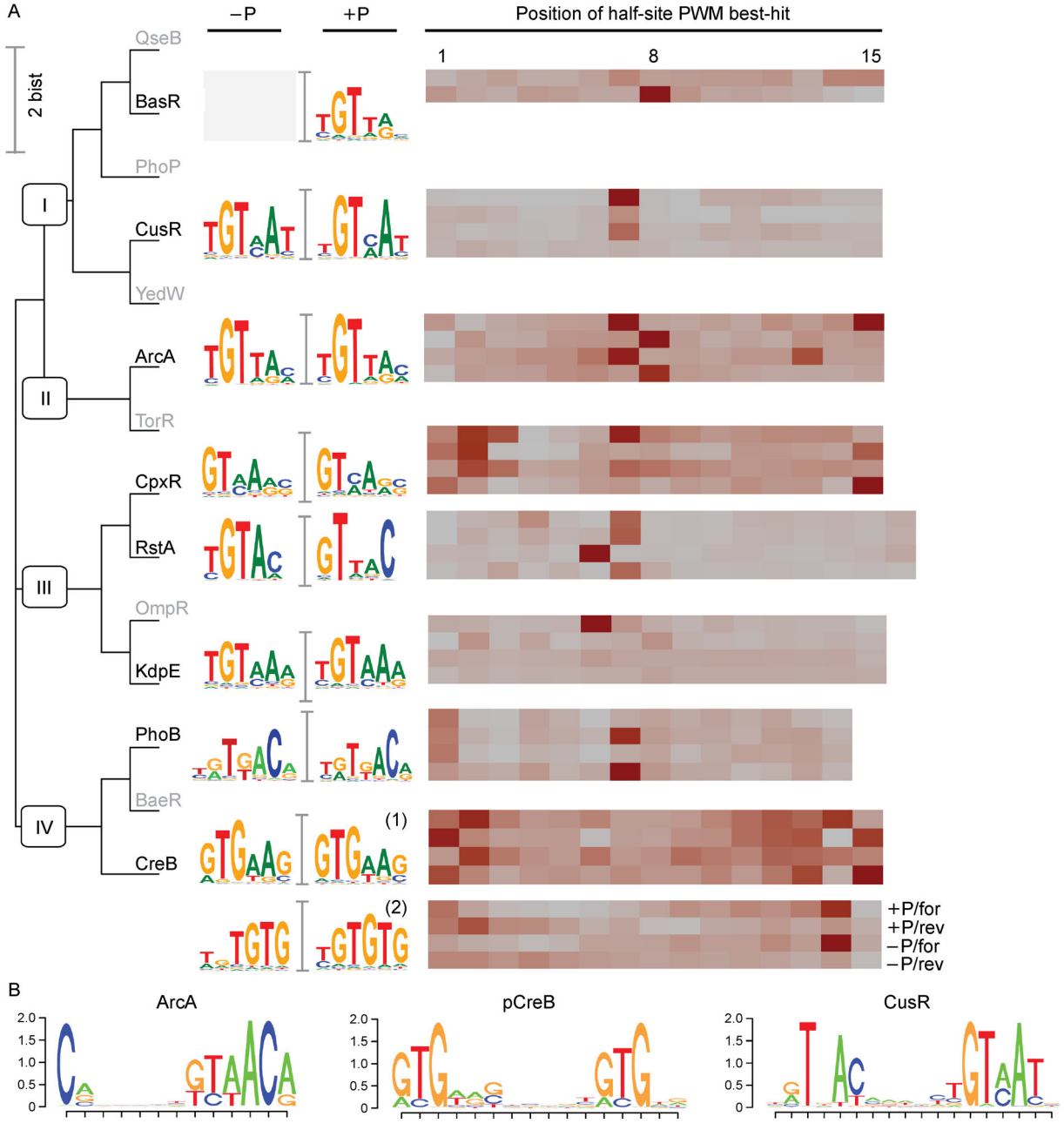

Figure 3. Variation in half-site recognition by OmpR family orthologs. (A) Representative half-site sequence logos derive from selection of the "anchored" GTAA-(20N) degenerate library are shown for nine eRR, with the names of eRR for which logos could not be obtained are in grey. Motifs obtained in the presence or absence of phosphodonor are indicated by +P and –P, respectively. The dendrogram reflects sequence-similarity between full-length consensus sequences derived for each eRR, and distinct lineages (I–IV) refer to Figure S1. For each motif, we display heat maps indicating the distribution of start positions for putative half-site binding events (top-scoring weight matrix hits) within the randomized region of the library; the frequency of putative binding events at each position is scaled from gray (low) to red (high). Note that longer motifs (for example those of PhoB) have fewer potential starting positions in the 20 bp library, giving rise to a heat map with fewer columns. Separate distributions are displayed by row in each heat map for putative binding events by phosphorylated and non-phosphorylated proteins (+P, –P) in either forward or reverse orientation (for, rev) relative to the fixed "GTAA" half-site, as shown for CreB (bottom). (B) Full-length eRR motifs reflect alternative use of the fixed half-site and 20 bp randomized region.

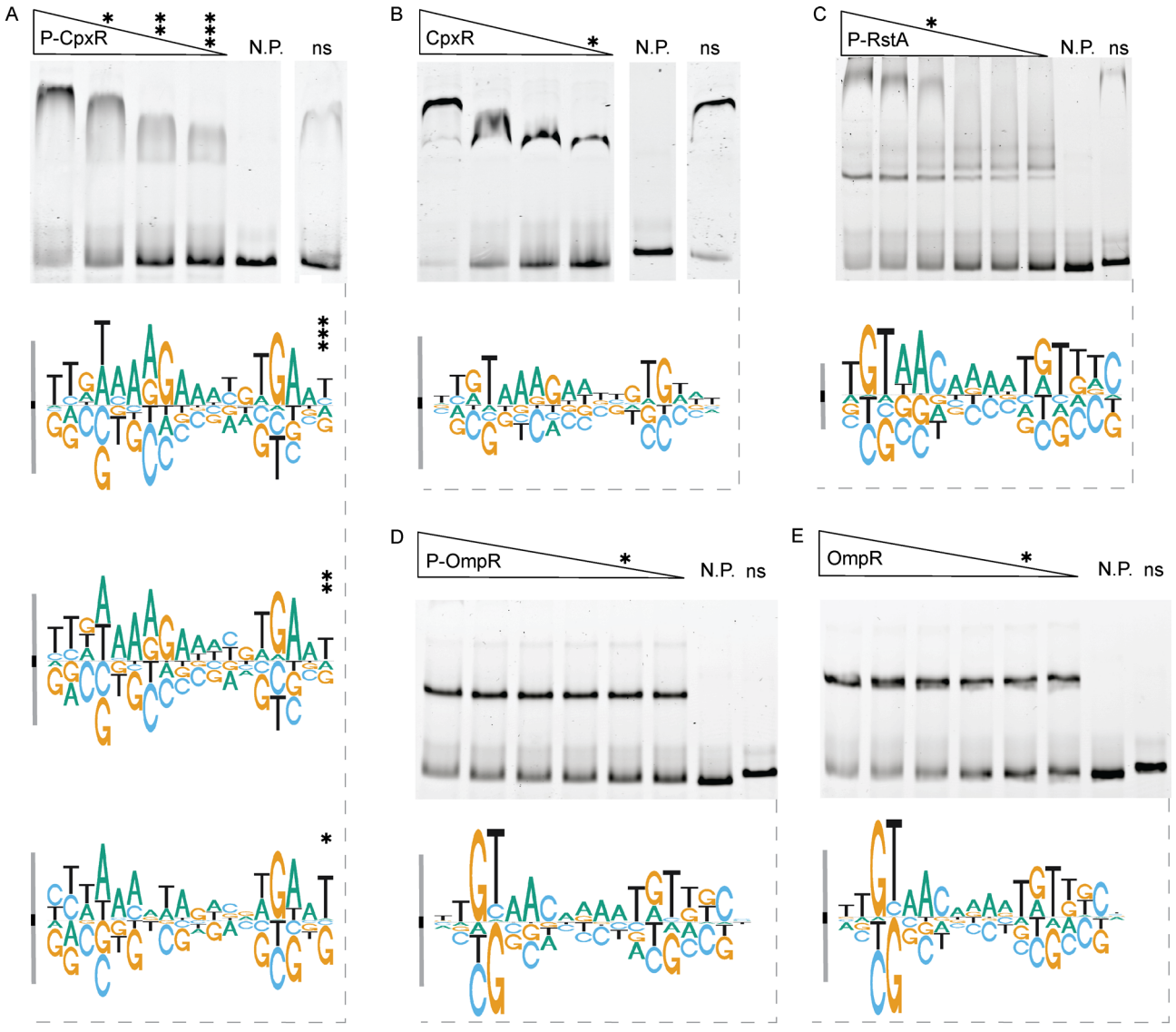

Figure 4. Determination of DNA binding specificity for CpxR, OmpR and RstA. (A) DNA-bound complexes of full-length, strep-tagged phospho-CpxR, (B) CpxR, (C) phospho-RstA, (D) phospho-OmpR, and (E) OmpR were obtained by excising bands separated by gel electrophoresis. Proteins were incubated with pooled, partially-degenerate DNA libraries based on the OmpR consensus binding sequence (See Methods) and run on 8% polyacrylamide (0.8X TBE) gels (N.P., no protein; ns, non-specific library). Bound and unbound DNA bands were extracted from lanes indicated with an asterisk, and relative free energies of binding to different DNA sequences were calculated by Spec-seq [45,53]. Energy logos depicting relative free energies for each nucleotide in the binding site [54] are located below their corresponding gel, with gray bars on the *y*-axis are normalized to a magnitude of +/–1 in units of (–kT) across all logos. In this format, high-amplitude positions contribute strongly to specificity, with bases above and below the center-line increasing and reducing affinity, respectively.

continuously hindered migration, but maintained a discrete band at higher protein concentration (Figure 4B, upper). Phospho-RstA formed a distinct banding pattern composed of three closely separated micro-states (and a fourth high-molecular weight state), which depended on protein concentration (Figure 4C, upper). The upper two micro-states dispersed with increasing phospho-RstA levels, while the lowest band steadily increased in intensity.

The major advantage of Spec-seq over other techniques is the ability to calculate the relative free energies of binding to hundreds of unique sequences in a single

reaction [51,53], in contrast to other techniques that may incur experimental or computational artefacts [55]. Using Spec-seq, we found no evidence of the asymmetric half-site recognition previously observed for RstA using SELEX (Figure 4C, lower). Although a number of bands indicative of high affinity micro-states may represent alternative, asymmetrically bound complexes, these did not yield sufficient material for sequencing. OmpR displayed energetic preferences close to its previously identified binding motif, which was also unchanged by phosphorylation (Figure 4D, E). CpxR, however, produced a recognition model that was strikingly distinct from both OmpR and RstA, and also from its own previously generated by SELEX (Figure 5A, B). Phospho-CpxR half-site preferences were asymmetric, with the downstream monomer adopting a highly specific "$t_{(+1)}g\mathbf{TGA}a_{(+6)}$" while upstream preferences underwent a concentration-dependent shift from the semi-canonical "$t_{(+1)}g\mathbf{TGA}a_{(+6)}$" to a novel "$t_{(+1)}t\mathbf{AAA}n_{(+6)}$" mode. In summary, this analysis demonstrates unique binding properties and sequence recognition among three highly similar LIII eRRs, suggesting complex mechanisms of site discrimination may occur *in vivo*.

Specificity-centric MI subnetwork distinguishes the RH as a SDR hub

Capra and coworkers previously conducted an analysis of mutual information (MI) between cognate HK:RR pairings to predict co-evolving, interfacial SDR pairs responsible for selective phosphotransfer [59]. We hypothesized that this subset of receiver-SDRs would additionally co-vary with determinants of DNA-binding specificity due to the convergent evolutionary pressure to

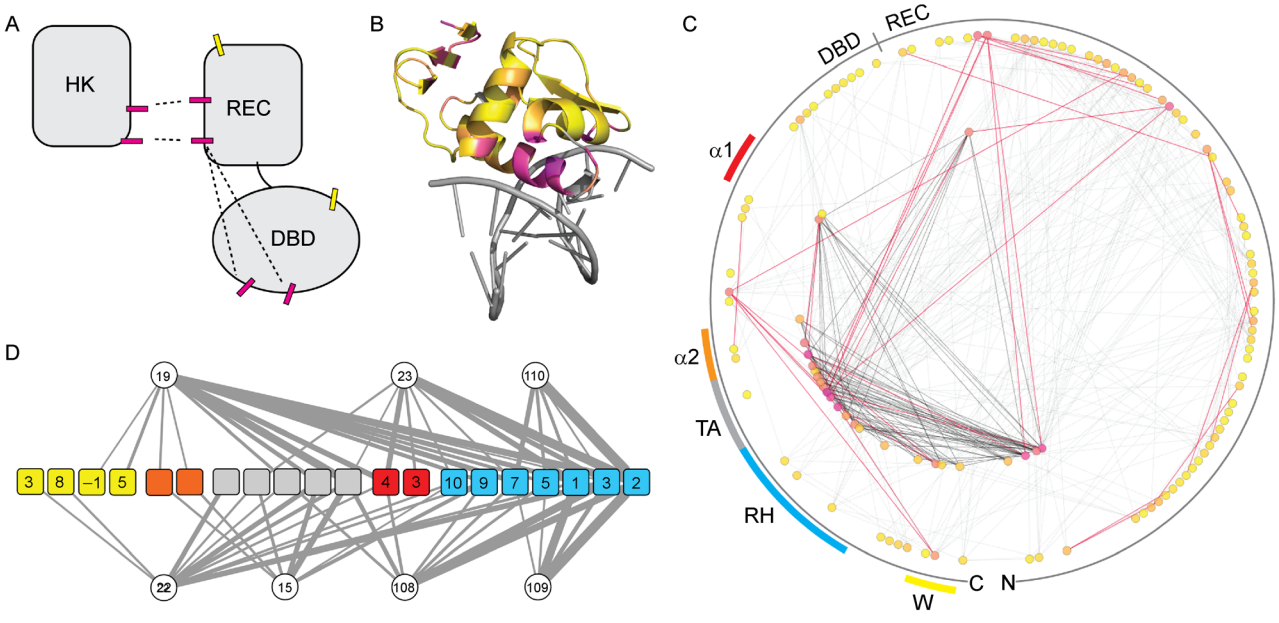

Figure 5. Sequence- and structure-based identification of specificity-determining residues at the protein-DNA interface. (A) Non-physically-interacting SDRs in distant regions of the protein may exhibit covariance due to their shared selective pressure against TCSP crosstalk. (B) Cumulative MI scores [56,57] calculated from a family-wide multiple sequence alignment are projected onto the structure of PhoP (PDB code: 5ED4 [33],) on a scale from minimum MI (yellow) to maximum MI (magenta). (C) MI relationships between positions in full-length RRs are presented as a network [56] created using Cytoscape [58]. Residues are ordered counter-clockwise by primary sequence in a circular layout with the start ("N") and end ("C") positions (at bottom), and nodes are colored according to cumulative MI as in Panel A. Protein sub-regions are color-coded to match the shading in Figure 1A: wing (W, yellow), $\alpha 2$ ($\alpha 2$, orange), trans-activation loop (TA, gray), recognition helix (RH, blue), and $\alpha 1$ ($\alpha 1$, red); the domain border between DBD and receiver domain (REC) is also highlighted. Statistically significant edges are shown only between positions separated > 10 residues, and nodes with < 3 edges were filtered for clarity. Nodes that constitute the first-order network of receiver-SDRs are shifted to an inner, concentric ring. Edges representing MI scores in the top 10% are shown as solid red lines; MI edges occurring within the first-order receiver-SDR contact network are solid black lines; and all other MI edges are represented as grey lines. (D) The first-order MI contact network derived from Panel C is re-displayed without intra-domain edges to highlight covariation arising due to evolutionary convergence. Gray edges represent the MI relationship between receiver-SDRs (white boxes) [59] and individual residues in the structural sub-regions identified in Panel C; line thickness is proportional to MI magnitude. DBD node labels reflect the reference coordinates previously established for residues in the wing, recognition helix, and $\alpha 1$ regions; receiver-SDR labels are numbered according to the sequence presented in Ref. [59].

maintain TCSP-specificity (Figure 5A). As a first step toward building this broader "specificity-centric" network, we generated a MI network using MISTIC [56] based on our alignment of ~2,000 OmpR family protein sequences extracted from a census of two-component RRs spanning 896 bacterial genomes [5]. In this approach, the cumulative MI for each position represents MI contributions summed over all possible pairings over the full protein length, and is thus expected to increase at positions with many significant MI pairings. Residues with high cumulative MI predominantly localized to the HK:RR and DBD:DNA interfaces; however, elevated values were also observed for a cluster of poorly-conserved, core-facing residues in the β1–4 region of the DBD (Figure 5B). To eliminate proximity-based signal (e.g., direct interaction), we next isolated the first-order, cross-domain contact network centered on previously validated receiver-SDRs [59], and observed a strong enrichment for residues in the RH, transactivation loop, and α2 regions (Figure 5C). Notably, β1–4 residues are absent from this sub-network, casting doubt on an active role in the maintenance of TCSP-specificity. Upon removal of intra-domain edges from this subnetwork, receiver-SDRs were clearly most associated with residues in the RH, and less so to other DNA-contacting sub-domains (Figure 5D).

Single SDRs alter different aspects of eRR:DNA specificity

To functionally validate our predicted SDRs, we searched for residues that differ between OmpR and CpxR, potentially accounting for their substantial divergence in sequence recognition. Non-conservative differences were observed only at RH_7 and RH_{10} (Figure 6A). These residue positions are also highly similar between OmpR and RstA, which display near-identical sequence preferences. We homology-modeled the putative SDRs within

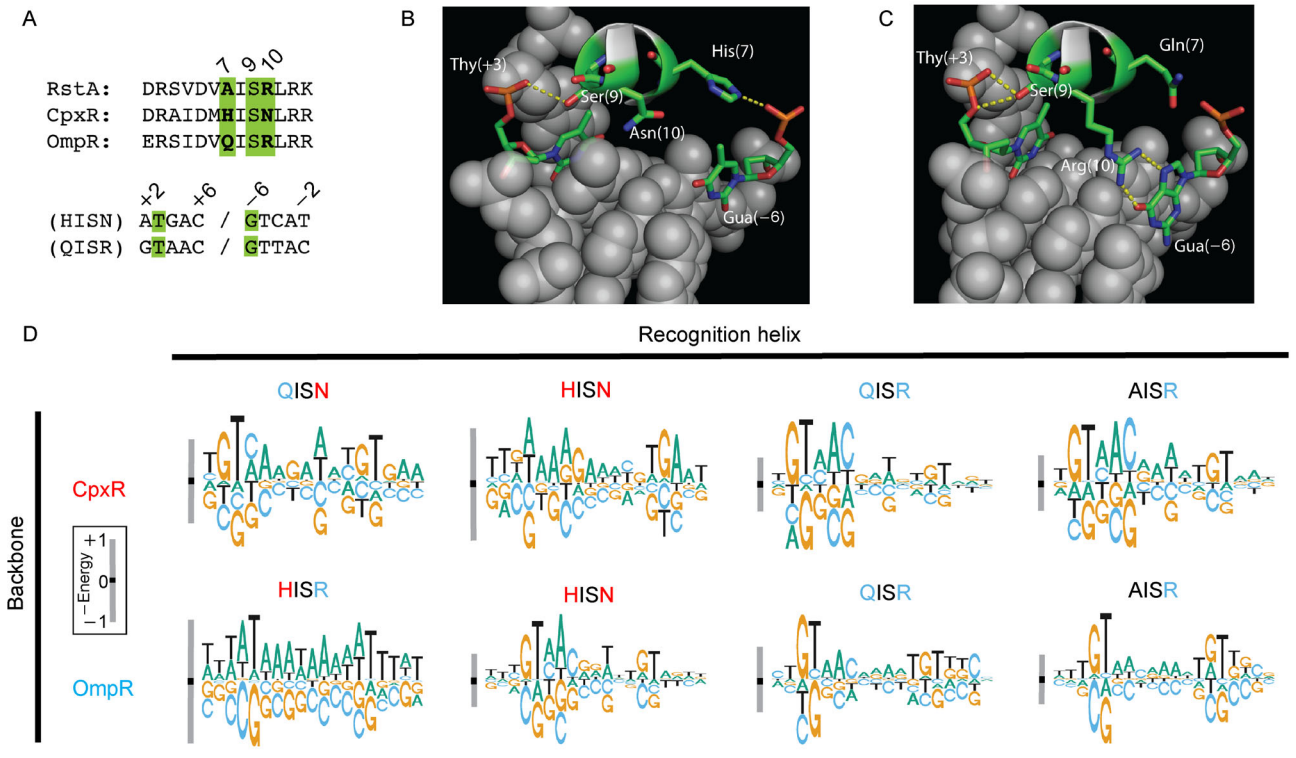

Figure 6. Quantitative determination of binding specificity for recognition helix (RH) variants in the CpxR and OmpR proteins. (A) An alignment is given for the RH regions of OmpR, CpxR and RstA. Green shaded positions match residue and base coloration in panel B; numbering reflects reference positions within the RH and canonical half-site. (B) Homology model of a "HISN"-containing RH based on the crystal structure of the *Klebsiella pneumoniae* PmrA protein (native sequence: HIHN) (PDB code: 4S05 [35]) in complex with an "A_{+2}TGAC" half-site sequence. In panels B and C, hydrogen bonds are represented by yellow dashed lines. (C) Homology model of 'QISR'-containing RH, based on the *K. pneumoniae* RstA protein (native sequence: AISR) (PDB code: 4NHJ [34]) in complex with a "G_{+2}TAAC" half-site sequence (PDB code: 4NHJ [34]). (D) Spec-seq [53] was carried out on phosphorylated RH variants and relative free energies are depicted as energy logos [54]. The background protein context (CpxR or OmpR) for the indicated mutations are shown in the left of the panel, and RH residues (color-coded to their RR of origin) are shown above the corresponding energy logo. As in Figure 4, gray bars on the *y*-axis are scaled across all logos to a magnitude of +/−1 kT.

the context of comparable backbones and DNA half-sites, and found that both RH_7 and RH_{10} were capable of contacting G_{-6} through the backbone and base, respectively, although not in the same structure simultaneously. In a CpxR-like context, RH_7[H] directly contacts the backbone at G_{-6}, whereas RH_{10}[N] makes no specific contacts to the DNA whatsoever (Figure 6B). Conversely, OmpR-RH_7[Q] has no DNA contacts, while RH_{10}[R] specifically recognizes G_{-6} through a bidentate hydrogen bond similar to that observed in the PhoB:DNA structure [10] (Figure 6C). The structures suggest that RH_{10} can specify the characteristic C_{+6} position in the canonical OmpR and RstA binding motifs "GTAAC_{+6}", whereas RH_{10} could mediate large-scale specificity alterations by stabilizing an alternate conformation of the DNA backbone.

To resolve the role of these putative SDRs in the divergent sequence recognition of OmpR and CpxR, we first examined the sequence-specificity and protein:DNA assembly of reciprocal RH_7 mutants. OmpR-RH_7[H] displayed low-specificity binding to A/T-rich sequence, consistent with an intrinsic preference for curved DNA sequences [60] (Figure 6D, bottom). This behavior is most consistent with a role for RH_7[H] in the recognition of DNA shape via the phosphate backbone, as predicted from the structural model. Interestingly, His is also the second-most frequent residue observed at RH_7 in our alignment of naturally occurring OmpR orthologs (Figure 1C), so this particular specificity is likely to have true physiological relevance. In contrast, CpxR-RH_7[Q] adopted canonical sequence preferences toward "G_{+2}T", despite the fact that neither one of these base pairs is in direct contact with RH_7 (Figure 6D, top); identical behavior was observed for a related "AISN" variant (data not shown). No other sequence preferences (including those at positions formerly in the vicinity of RH_7[H]) were significantly affected. Neither mutation altered the assembly behavior of the protein:DNA complex; however, apparent affinity was substantially reduced in the mutants (data not shown). Because the loss and gain of RH_7[H] in two eRR scaffold contexts was concomitant with specificity attributes characteristic of shape recognition, we conclude from these results that RH_7[H] is necessary and sufficient for i) re-shaping the groove structure of the half-site or ii) altering the presentation of RH sidechains to the DNA.

Having observed a nearly reciprocal inter-conversion of specificities between RH_7 mutants, we further aimed to determine whether combined RH_7/RH_{10} double-mutants would complete the re-programming of OmpR-like and CpxR-like binding. CpxR-$RH_{7\text{-}10}$[QISR] performed as predicted, exhibiting a fully OmpR-like specificity profile across both half-sites with just one unique preference at position 9, which occurs in the putative spacer (Figure 6D, top). Surprisingly, this second mutation restored native-like affinity, and the gel migration pattern strongly favored the stable, OmpR-like homodimeric complex (data not shown). This full conversion from the hybrid site observed in the "QISN" form suggests that the specificity-determining activities of RH_7 and RH_{10} are separable to a degree, although the experiment cannot determine whether the two are fully independent. In contrast, OmpR-$RH_{7\text{-}10}$[HISN] exhibits wild-type specificity, with the exception of the ninth base position, where it matches the "QISN/R" mutants in the CpxR scaffold (Figure 6D, bottom). This base is located within the spacer and lacks correlation to any particular residue or scaffold. Finally, we examined the effects of an RstA-like $RH_{7\text{-}10}$[AISR] mutation in both scaffold contexts, and no change in specificity from the "QISR" RHs was observed.

DISCUSSION

In this work, we applied multiple high-throughput techniques to profile specificity determinants in the OmpR sub-family of wHTH TFs, which constitute approximately 30% of downstream effectors in two-component signal transduction systems. Two-component signaling pathways are a critical and predominant sensory modality in bacteria, and are a classic system for the study of functional specificity of paralogous proteins and pathways. This work is notable for several reasons. First, in contrast to the simple mechanisms of DNA binding employed by most of the profiled eukaryotic TF families, OmpR family proteins bind as multimers in response to phosphorylation of a regulatory domain, greatly increasing the complexity of potential sequence interactions. Second, prokaryotic TFs are usually profiled individually with low-throughput methods, whereas we generated high-resolution specificity models from thousands of sequences for a representative majority of the *E. coli* OmpR protein family. Third, we utilized randomized, synthetic binding site libraries that allowed us to challenge OmpR family members with a more complex set of binding partners than they encounter *in vivo*. Our use of these *in vitro* libraries to identify binding motifs, rather than genomic DNA, further ensures that our results are unbiased by native sequence context. Fourth, using a recently developed technique known as Spec-seq, we were able to measure relative affinities toward thousands of sequences directly, while simultaneously visualizing the assembly of distinct protein:DNA complexes in an EMSA format. This approach provided an unprecedented level of insight into the relationship between DNA-binding specificity and protein:DNA assembly, and added to our understanding of TF interactions important for gene regulation. To identify SDRs, we integrated information

from comparative genomics, structural biology, and evolutionary analysis based on amino acid co-variation. We demonstrated the utility of this approach by experimentally confirming the role of two such residue positions in paralog-specific sequence recognition. Overall, this work unites our understanding of the structure, function, and regulatory activities fundamental to the OmpR family.

Based on collections of genomic binding sites, crystal structures, and biochemical assays, the dominant mode for OmpR family proteins involves direct repeats of "TTA"- or "GTA"-containing half-sites with a center-to-center distance of 9–10 bp. Our high-resolution findings largely confirm these overall preferences, but also suggest that the orientation of composite half-sites may vary within the family. Sequences containing inverted repeats were enriched during CusR- and CpxR-SELEX. Although we did not assay these mechanisms of binding directly (e.g., through FRET or cross-linking), an inverted "head-to-head" binding site has been observed for CusR in its native operator [47]; the highly similar eRR YedW also enriched this operator through genomic SELEX [18]. Formally, the solution-state SELEX technique does not resolve complexes differing in size, and therefore the observed architectures may result from symmetrically bound dimers. However, the footprint size of the complex bound to the aforementioned genomic sequence, 28 bp, is insufficient for multimeric binding [47]; neither did we ever observe the formation of high molecular weight protein:DNA complexes in EMSAs performed for CusR and YedW prior to SELEX analysis (data not shown). In summary, we conclude that alternative dimeric architectures are possible, though apparently not widespread, and future work is needed to explore the mechanism of their formation.

Our representative collection of high-resolution binding models provides significant insight into the regulatory logic of OmpR-family RRs. For example, the SELEX-derived binding motif of BasR is the canonical "TGTTA" direct repeat (10 bp spacing), but a previously defined *in vivo* specificity model of BasR:DNA complexes, which conforms to the consensus "cTTAAnnTTnncTTAAnnTT"), diverges substantially from that model at the strongly specified G_{+2} position [21]. It has further been shown that BasR forms multiple distinct complexes with native operator sequences, and it "footprints" very large segments of DNA. From a regulatory perspective, this indicates that BasR-dependent operators are under active selection to maintain multiple low-affinity binding sites. Strangely, upstream and downstream half-sites symmetrically disfavor the G_{+2} position, but, if simple low-affinity binding were the intended result, then mismatches would be spread equally over all positions. In theory, favoring one specific mismatch at a high-specificity position in the canonical binding mode opens a sub-specificity niche that would insulate BasR binding sites from paralogous competitors. In a contrasting strategy, CreB exhibits a wholly unique variation of the canonical motif, matching the "cre-tag" identified in promoters responsive to the CreBC TCSP [49]; similarly, the inverted repeat preference observed for CusR matches precisely to recently characterized binding sites in target promoters [47]. From an evolutionary standpoint, the emergence of novel base and orientation preferences would reduce the risk of cross-talk with other paralogs, thereby reducing the need for an assembly- or affinity-based strategies for operator discrimination.

A striking result of our initial SELEX experiment, which used a fully-randomized 20N library, was the emergence of two binding motifs based on distinct "GTA" and "(AC/TT)GCT" sequences. Interestingly, at least one protein (QseB) appears capable of binding both motifs. Multiple modes of binding specificity within a single structural family have been proposed before, although they have also been shown to arise artefactually from the models used to represent sequence-specific interactions [55]. Nevertheless, bona fide multi-specific binding has been observed for the eukaryotic FOX family of sequence-specific TFs, which notably also contain a DBD belonging to the winged helix-turn-helix class [61]. Furthermore, the two most common binding motifs for the profiled forkhead-domain TFs are of the consensus "GTAAAC" and "ACGC," partial matches for two of the binding motifs shared by OmpR homologues in our SELEX experiment. The alternative "GCT" motif also has a structure somewhat similar to a binding consensus previously generated by *in vitro* SELEX for *Mycobacterium* PhoP (**GCT**GTGA) [62]. Both *Mycobacterium* PhoP and *Klebsiella* PmrA (a BasR homolog) have been crystallized in complex with sequences containing "GCT"-like motifs occupying equivalent positions relative to the protein, with each overlapping at the canonically conserved $T_{(+3)}$ ("GC**T**"/"G**T**$_{(+3)}$-A") [33,35].

In an interesting case, CpxR exhibited multiple motif preferences, partly depending on the experimental approach. Both direct- and inverted-repeat recognition appeared in fully-randomized SELEX, while "seeded" SELEX yielded canonical results (consistent with the previous experiment) and apparent non-canonical spacing. Using Spec-seq, we observed the gradual progression of a canonical binding motif into a non-canonical, multimeric mode with unique sequence preferences (in both a phospho- and concentration-dependent manner). In full context, it is apparent that the complex binding properties observed in our partially randomized SELEX experiments likely resulted from binding to sequences flanking the synthetic half-site, which, when viewed with

flanking bases ("**AG**GTAA"), proves a close match to the upstream, non-canonical half-site observed in Spec-seq. Additionally, it is likely that the increased prevalence of the non-canonical CpxR binding mode during Spec-seq may have resulted from crowding-enhanced stability in the equilibration step or within the gel itself; extensive "smearing" does indicate that this complex is dynamic or unstable. These results illustrate the risk of analyzing TFs (or any proteins) under a single set of conditions, and demonstrate that no DNA library contains innocuous sequence. We are confident in the results reported in this work, but also recommend independent follow-up to any *de novo* binding motifs generated using enrichment-based analytical techniques.

Despite intensive study of the OmpR family, there is very little known about the role of specific residues in specificity determination; moreover, no common model has been proposed that successfully predicts paralog-specific behaviors. Surprisingly, we found that a simple, consensus-based metric of protein similarity captured functionally related protein lineages within the OmpR family. For example, we observed that the LIII RRs OmpR, CpxR, and RstA all preferred direct repeats with 9 bp distance between half-site centers. However, we further observed significant differences between the sequence-specificity and the assembly state of these TFs reflective of their interactions *in vivo* in the regulation of the *csgD* promoter. Both OmpR and RstA positively regulate the expression of *csgD*, and it was shown using DNase I footprinting that they do so through a shared target sequence; CpxR, which is a repressor of *csgD*, occupied a large region of the promoter (*in vitro*) overlapping the RstA/OmpR binding site [28]. Our Spec-seq analysis was consistent with these regulatory relationships, and further revealed that oligomeric binding by phospho-CpxR is supported by a unique binding motif and mode of assembly. Our findings suggest that usage of alternative binding motifs and multimeric complexes may discriminate between activator and repressor functions, and we argue that a more complete *in vitro* characterization of OmpR family members at native operator sequences is necessary to fully understand those mechanisms.

Many aspects of DNA-binding by OmpR-family TFs complicate the identification of residues important for sequence recognition. First, there are ~20 residue positions (per monomer) that have the potential to contact the DNA helix via the phosphate backbone and major or minor grooves. Many exhibit preferences for multimeric binding and/or recognition of curved DNA, so large-scale changes in shape or geometry may play an outsized role in target recognition. Computational metrics are often applied to narrow the field of potential SDRs, often using a statistical proxy (e.g., MI) to infer a coevolutionary relationship between residues in the case of protein interactions [59], or residues and DNA binding site sequences in the case of protein:DNA interactions [63]. It was novel, to our knowledge, to consider that SDRs important for many different, TCSP-specific functions might co-vary in a "specificity-centric" network through a process analogous to convergent evolution. One of the major benefits of this approach was that it required signal between completely non-interacting residue positions, implicitly controlling for statistical correlations due to structural proximity and/or co-evolution. Additionally, the approach required no structural input outside the fact that the two domain interfaces were non-interacting. Two such predictions were able to fundamentally alter or interconvert aspects of sequence-specific binding and protein:DNA assembly between OmpR and CpxR, surprisingly via distinct molecular contacts with the same DNA base. These findings and the techniques described begin to lay the groundwork for a holistic structural understanding of the RR:DNA interface, which could enable the direct prediction of binding specificity from amino acid sequence (e.g., for pathogenically-relevant organisms or mutants).

In summary, we have quantified widespread differences in the recognition of half-site sequence, spacing, and orientation by *E. coli* OmpR family RRs with physiological relevance to gene regulatory activity. Integrating phylogenetic, structural, and functional sources of information, we predicted and tested novel SDRs through the transplantation of sequence-specific binding attributes between paralogs. Our analysis revealed a complex role for SDRs in the establishment not only of sequence preferences, but also protein:DNA assembly. Overall, these results greatly advance our understanding of the wHTH-specific "protein-DNA code," which may be useful to predict the targets of newly discovered OmpR homologues as well as design new regulatory tools for synthetic biology. Further, this work provides a basis for the continued study of two-component system evolution, which will help to decipher the regulation of complex homeostatic, pathogenic, and industrially relevant bacterial processes. Finally, we have provided a substantial resource of both specificity profiles and SDR predictions for continued functional analysis of this important, widespread TF family.

MATERIALS AND METHODS

Cloning, expression, and purification

Coding sequences of 14 response regulators (RRs) of the OmpR sub-family (ArcA, BaeR, BasR, CpxR, CreB, CusR, KdpE, OmpR, PhoB, PhoP, QseB, RstA, TorR, YedW) were amplified directly from *E. coli* MG1655

genomic DNA. Coding sequence for the StrepTagII affinity tag (WSHPQFEK) was added by PCR amplification along with upstream and downstream restriction sites for *Mfe*I and *Xho*I, respectively. Strep-RR fusion protein sequences were sub-cloned into the pET-42a(+) expression vector in-frame with N-terminal GST and 6×His purification tags and a thrombin protease cleavage site, generating triple-tagged constructs. Stock plasmids were stored, purified and handled using standard laboratory techniques.

ArcticExpress (DE3) competent cells (Agilent) were chemically transformed with expression plasmids, and single colonies from selective (Kan) LB-agar plates were used to inoculate 5 mL LB-Kan starter cultures. After 6–8 h growing at 37°C, starter cultures were scaled up to 400 mL expression cultures in triple-baffled 4 L flasks prepared with auto-induction media containing Kanamycin according to the Studier method [64]. Cultures were expanded at 37°C for 3–6 h, then grown several hours past saturation (24–36 h total growth time) at 25°C to achieve maximum protein yield. Bacterial pellets were harvested by centrifugation, sonicated, re-pelleted at high speed to remove cellular debris, and lysate (diluted with 1× PBS to reduce viscosity) was passed through a 0.45 μm syringe-tip filter for clarification. Lysate was passed over a HiTrap GST affinity column (1 mL capacity, GE Healthcare) and eluted under manufacturer-specified buffer conditions. Fusion protein was cleaved with 5 U thrombin protease, and GST-6×His was removed with two rounds of treatment with Ni-NTA resin (Thermo Scientific). Protein samples were cleared completely of resin by passage through 0.22 μm syringe-tip filters. Purity was assessed by both SDS-PAGE and size-exclusion chromatography, and protein concentration was determined by NanoDrop (Thermo Scientific).

Construction of mutants

All mutants were generated by site-directed mutagenesis of wild-type plasmid construct using Gibson Assembly Master Mix (New England Biolabs) [65]. Expression and purification were carried out as described for wild-type proteins.

SELEX and Spec-seq library preparation

DNA libraries were designed to contain flanking sequences to support PCR amplification and direct sequencing on the Illumina platform, and were obtained as a single-stranded, PAGE-purified oligonucleotides from Integrated DNA Technologies. For SELEX library construction, 250 ng single-stranded DNA (ssDNA) were mixed with a reverse primer in two-fold molar excess in 1× NEBuffer 2 (50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9 at 25°C), heated to 85°C and slowly annealed to 30°C. Following the addition of 10 U Klenow Fragment (New England Biolabs) and 1 mM dNTPs, extension reactions were incubated at 37°C for 2 hours, and double-stranded DNA (dsDNA) libraries were subsequently purified using Qiaquick PCR Purification columns (Qiagen) and eluted in Qiagen EB (10 mM Tris-Cl, pH 8.5 at 25°C). Labeled dsDNA libraries for Spec-seq were generated by two-step PCR with Phusion or Q5 High-Fidelity DNA Polymerase (NEB) using FAM-labeled primers and purified as described above.

SELEX

Ammonium phosphoramidate for protein phosphorylation was synthesized in-house according to an established protocol [66]. Strep-tagged proteins were pre-incubated 1 hour at 32°C in binding buffer (10 mM Tris-Cl, 7.5; 200 mM KCl; 20 mM NaCl, 2 mM $MgCl_2$), 2 μg polydI-dC, 0.1 mg/ml BSA, and either 50 mM NH_4Cl or 50 mM ammonium phosphoramidate for non-phosphorylated and phosphorylated conditions, respectively. Pre-incubated protein samples were aliquoted (40 μL final volume) into PCR strip tubes containing 200 ng of the appropriate DNA library, and incubated an additional hour at 32°C (10 mM Tris-Cl, 7.5; 100 mM KCl; 10 mM NaCl, 1 mM $MgCl_2$, 25 mM NH_4Cl or phosphoramidate). Binding reactions were mixed with a washed suspension of Strep-tactin magnetic beads (Qiagen) and placed on ice for 30 minutes; to prevent bead settling, reactions were mixed by gentle pipetting at 10-minute intervals. Beads were pelleted magnetically and supernatant was removed by gentle pipetting. Pellets were washed once (without disturbance) with a single volume of ice-cold binding buffer. Pellets were resuspended in 20 μL elution buffer (Qiagen TE + 150 mM NaCl) and incubated for 20 minutes at 80°C. Eluted DNA was amplified for subsequent selections in a two-step reaction using either Phusion or Q5 High-Fidelity DNA polymerase for 12–18 cycles, and purified using the MinElute PCR purification system (Qiagen).

Spec-seq

Binding reactions were prepared on ice in 12 μL volumes containing 20 ng FAM-labeled dsDNA library in 1× EMSA Buffer (10 mM Tris-Cl, 15 mM KCl, 60 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mg/mL BSA, 5% glycerol, 3% Ficoll, 10 ng/μL salmon sperm DNA, pH 8.3 at 8°C). Reactions were incubated 2 h at 32°C with 25 mM ammonium phosphoramidate or ammonium chloride for binding of phosphorylated and non-phosphorylated response regulators, respectively. Bound and unbound DNA pools were separated by native PAGE (8%

polyacrylamide, 0.8× TBE [72 mM Tris-borate, 1 mM EDTA]) at 8°C. Gels were visualized on a Typhoon FLA 9500 (GE Healthcare) Biomolecular Imager. Bands containing bound and unbound DNA were excised, and DNA was extracted by the crush and soak method [67] in gel diffusion buffer (0.3 M sodium acetate, 1 mM EDTA).

Residue covariation analysis

OmpR family orthologs were identified using a reciprocal best-BLAST hit criterion [68] from a previously curated list of OmpR family members spanning 896 bacterial genomes [5]. To qualify as orthologous, a conservative cutoff of 40% sequence identity was imposed, and redundancy within each group was reduced using CD-HIT (90% identity clustering) prior to alignment using M-coffee [69]. To build a family-wide alignment, we conducted step-wise, progressive profile alignments guided by a preliminary tree based on alignment similarity. A MI network was constructed for this multiple alignment using the MISTIC web interface [56] and visualized using Cytoscape [58].

ABBREVIATIONS

RR,	response regulator
eRR,	*E. coli* OmpR family response regulator
HK,	histidine kinase
TCSP,	two-component signal pathway
DBD,	DNA-binding domain
wHTH,	winged helix-turn-helix
SDR,	specificity-determining residue
α1,	alpha helix 1
W,	wing
RH,	recognition helix
TF,	transcription factor
TA,	trans-activation loop
PWM,	position weight matrix
MI,	mutual information
EMSA,	electrophoretic mobility shift assay
SELEX,	systematic evolution of ligands by exponential enrichment

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-018-0130-0.

ACKNOWLEDGEMENTS

We are grateful for the thoughtful input of former members of the laboratory Benjamin Borgo and Chi Zhang in the initial phases of this work, and for continuing helpful discussions and whose technical expertise greatly added to this work. We thank Dr. Cailin Joyce and Dr. GiNell Elliott for their critical commentary during the preparation of this manuscript. We especially thank Jessica Hoisington-Lopez for her input and heroic patience in our SELEX library design and the development of sequencing approaches. This project was completed with support of NSF Graduate Research Fellowship Award DGE-1143954 (to A.P.J.) and National Institutes of Health Award Number R01GM101602 (to J.J.H.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors report no potential conflicts of interest.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Adam P. Joyce and James J. Havranek declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Stock, A. M., Robinson, V. L. and Goudreau, P. N. (2000) Two-component signal transduction. Annu. Rev. Biochem., 69, 183–215
2. Laub, M. T. and Goulian, M. (2007) Specificity in two-component signal transduction pathways. Annu. Rev. Genet., 41, 121–145
3. Gao, R., Tao, Y. and Stock, A. M. (2008) System-level mapping of *Escherichia coli* response regulator dimerization with FRET hybrids. Mol. Microbiol., 69, 1358–1372
4. 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
5. Galperin, M. Y. (2010) Diversity of structure and function of response regulator output domains. Curr. Opin. Microbiol., 13, 150–159
6. Pérez-Rueda, E., Collado-Vides, J. and Segovia, L. (2004) Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. Comput. Biol. Chem., 28, 341–350
7. Martínez-Hackert, E. and Stock, A. M. (1997) Structural relationships in the OmpR family of winged-helix transcription factors. J. Mol. Biol., 269, 301–312
8. Kenney, L. J. (2002) Structure/function relationships in OmpR and other winged-helix transcription factors. Curr. Opin. Microbiol., 5, 135–141
9. Toro-Roman, A., Wu, T. and Stock, A. M. (2005) A common dimerization interface in bacterial response regulators KdpE and TorR. Protein Sci., 14, 3077–3088
10. Blanco, A. G., Sola, M., Gomis-Rüth, F. X. and Coll, M. (2002) Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator. Structure, 10, 701–713
11. Mattison, K., Oropeza, R., Byers, N. and Kenney, L. J. (2002) A phosphorylation site mutant of OmpR reveals different binding conformations at *ompF* and *ompC*. J. Mol. Biol., 315, 497–511
12. Flores-Valdez, M. A., Fernández-Mora, M., Ares, M. Á., Girón, J. A., Calva, E. and De la Cruz, M. Á. (2014) OmpR phosphorylation regulates *ompS1* expression by differentially controlling the use of

promoters. Microbiology, 160, 733–741
13. Head, C. G., Tardy, A. and Kenney, L. J. (1998) Relative binding affinities of OmpR and OmpR-phosphate at the *ompF* and *ompC* regulatory sites. J. Mol. Biol., 281, 857–870
14. Narayanan, A., Paul, L. N., Tomar, S., Patil, D. N., Kumar, P. and Yernool, D. A. (2012) Structure-function studies of DNA binding domain of response regulator KdpE reveals equal affinity interactions at DNA half-sites. PLoS One, 7, e30102
15. Walthers, D., Tran, V. K. and Kenney, L. J. (2003) Interdomain linkers of homologous response regulators determine their mechanism of action. J. Bacteriol., 185, 317–324
16. Kim, S., Broströmer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y. Q., *et al.* (2013) Probing allostery through DNA. Science, 339, 816–819
17. Clarke, M. B. and Sperandio, V. (2005) Transcriptional regulation of *flhDC* by QseBC and σ^{28} (FliA) in enterohaemorrhagic *Escherichia coli*. Mol. Microbiol., 57, 1734–1749
18. Ishihama, A., Shimada, T. and Yamazaki, Y. (2016) Transcription profile of *Escherichia coli*: genomic SELEX search for regulatory targets of transcription factors. Nucleic Acids Res., 44, 2058–2074
19. Ogasawara, H., Hasegawa, A., Kanda, E., Miki, T., Yamamoto, K. and Ishihama, A. (2007) Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. J. Bacteriol., 189, 4791–4799
20. Shimada, T., Takada, H., Yamamoto, K. and Ishihama, A. (2015) Expanded roles of two-component response regulator OmpR in *Escherichia coli*: genomic SELEX search for novel regulation targets. Genes Cells, 20, 915–931
21. Ogasawara, H., Shinohara, S., Yamamoto, K. and Ishihama, A. (2012) Novel regulation targets of the metal-response BasS-BasR two-component system of *Escherichia coli*. Microbiology, 158, 1482–1492
22. Park, D. M., Akhtar, M. S., Ansari, A. Z., Landick, R. and Kiley, P. J. (2013) The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. PLoS Genet., 9, e1003839
23. Yang, C., Huang, T. W., Wen, S. Y., Chang, C. Y., Tsai, S. F., Wu, W. F. and Chang, C. H. (2012) Genome-wide PhoB binding and gene expression profiles reveal the hierarchical gene regulatory network of phosphate starvation in *Escherichia coli*. PLoS One, 7, e47314
24. Nishino, K., Honda, T. and Yamaguchi, A. (2005) Genome-wide analyses of *Escherichia coli* gene expression responsive to the BaeSR two-component regulatory system. J. Bacteriol., 187, 1763–1772
25. De Wulf, P., McGuire, A. M., Liu, X. and Lin, E. C. (2002) Genome-wide profiling of promoter recognition by the two-component response regulator CpxR-P in *Escherichia coli*. J. Biol. Chem., 277, 26652–26661
26. Harlocker, S. L., Bergstrom, L. and Inouye, M. (1995) Tandem binding of six OmpR proteins to the *ompF* upstream regulatory sequence of *Escherichia coli*. J. Biol. Chem., 270, 26849–26856
27. Batchelor, E., Walthers, D., Kenney, L. J. and Goulian, M. (2005) The *Escherichia coli* CpxA-CpxR envelope stress response system regulates expression of the porins *ompF* and *ompC*. J. Bacteriol., 187, 5723–5731
28. Ogasawara, H., Yamada, K., Kori, A., Yamamoto, K. and Ishihama, A. (2010) Regulation of the *Escherichia coli* csgD promoter: interplay between five transcription factors. Microbiology, 156, 2470–2483
29. Feldheim, Y. S., Zusman, T., Speiser, Y. and Segal, G. (2016) The *Legionella pneumophila* CpxRA two-component regulatory system: new insights into CpxR's function as a dual regulator and its connection to the effectors regulatory network. Mol. Microbiol., 99, 1059–1079
30. Wang, X., Gao, H., Shen, Y., Weinstock, G. M., Zhou, J. and Palzkill, T. (2008) A high-throughput percentage-of-binding strategy to measure binding energies in DNA-protein interactions: application to genome-scale site discovery. Nucleic Acids Res., 36, 4863–4871
31. Park, D. M. and Kiley, P. J. (2014) The influence of repressor DNA binding site architecture on transcriptional control. MBio, 5, e01684–14
32. Nowak-Lovato, K. L., Hickmott, A. J., Maity, T. S., Bulyk, M. L., Dunbar, J. and Hong-Geller, E. (2012) DNA binding site analysis of *Burkholderia thailandensis* response regulators. J. Microbiol. Methods, 90, 46–52
33. He, X., Wang, L. and Wang, S. (2016) Structural basis of DNA sequence recognition by the response regulator PhoP in Mycobacterium tuberculosis. Sci. Rep., 6, 24442
34. Li, Y.-C., Chang, C. K., Chang, C.-F., Cheng, Y.-H., Fang, P.-J., Yu, T., Chen, S.-C., Li, Y.-C., Hsiao, C.-D. and Huang, T. H. (2014) Structural dynamics of the two-component response regulator RstA in recognition of promoter DNA element. Nucleic Acids Res., 42, 8777–8788
35. Lou, Y. C., Weng, T. H., Li, Y. C., Kao, Y. F., Lin, W. F., Peng, H. L., Chou, S. H., Hsiao, C. D. and Chen, C. (2015) Structure and dynamics of polymyxin-resistance-associated response regulator PmrA in complex with promoter DNA. Nat. Commun., 6, 8838
36. Narayanan, A., Kumar, S., Evrard, A. N., Paul, L. N. and Yernool, D. A. (2014) An asymmetric heterodomain interface stabilizes a response regulator-DNA complex. Nat. Commun., 5, 3282
37. Rhee, J. E., Sheng, W., Morgan, L. K., Nolet, R., Liao, X. and Kenney, L. J. (2008) Amino acids important for DNA recognition by the response regulator OmpR. J. Biol. Chem., 283, 8664–8677
38. Dhiman, A., Rahi, A., Gopalani, M., Bajpai, S., Bhatnagar, S. and Bhatnagar, R. (2017) Role of the recognition helix of response regulator WalR from *Bacillus anthracis* in DNA binding and specificity. Int. J. Biol. Macromol., 96, 257–264
39. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., *et al.* (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res., 44, D279–D285
40. Itou, H. and Tanaka, I. (2001) The OmpR-family of proteins: insight into the tertiary structure and functions of two-component regulator proteins. J. Biochem., 129, 343–350
41. Blanco, A. G., Canals, A., Bernués, J., Solà, M. and Coll, M.

(2011) The structure of a transcription activation subcomplex reveals how σ^{70} is recruited to PhoB promoters. EMBO J., 30, 3776–3785

42. Sandelin, A. and Wasserman, W. W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J. Mol. Biol., 338, 207–215
43. Sloutsky, R. and Naegle, K. M. (2016) High-resolution identification of specificity determining positions in the LacI protein family using ensembles of sub-sampled alignments. PLoS One, 11, e0162579
44. Liu, X., Brutlag, D. L. and Liu, J. S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In Biocomputing 2001. pp. 127–138. Singapore: World Scientific Publishing Company
45. Stormo, G. D. (2013) Modeling the specificity of protein-DNA interactions. Quant. Biol., 1, 115–130
46. Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res., 18, 6097–6100
47. Urano, H., Umezawa, Y., Yamamoto, K., Ishihama, A. and Ogasawara, H. (2015) Cooperative regulation of the common target genes between H_2O_2-sensing YedVW and Cu^{2+}-sensing CusSR in *Escherichia coli*. Microbiology, 161, 729–738
48. Harari, O., Park, S. Y., Huang, H., Groisman, E. A. and Zwir, I. (2010) Defining the plasticity of transcription factor binding sites by Deconstructing DNA consensus sequences: the PhoP-binding sites among gamma/enterobacteria. PLoS Comput. Biol., 6, e1000862
49. Cariss, S. J., Tayler, A. E. and Avison, M. B. (2008) Defining the growth conditions and promoter-proximal DNA sequences required for activation of gene expression by CreBC in *Escherichia coli*. J. Bacteriol., 190, 3930–3939
50. Yamamoto, K. and Ishihama, A. (2006) Characterization of copper-inducible promoters regulated by CpxA/CpxR in *Escherichia coli*. Biosci. Biotechnol. Biochem., 70, 1688–1695
51. Stormo, G. D., Zuo, Z. and Chang, Y. K. (2015) Spec-seq: determining protein-DNA-binding specificity by sequencing. Brief. Funct. Genomics, 14, 30–38
52. Zuo, Z., Chang, Y. and Stormo, G. D. (2015) A quantitative understanding of lac repressor's binding specificity and flexibility. Quant. Biol., 3, 69–80
53. Zuo, Z. and Stormo, G. D. (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. Genetics, 198, 1329–1343
54. Foat, B. C., Morozov, A. V. and Bussemaker, H. J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics, 22, e141–e149
55. Zhao, Y. and Stormo, G. D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat. Biotechnol., 29, 480–483
56. Simonetti, F. L., Teppa, E., Chernomoretz, A., Nielsen, M. and Marino Buslje, C. (2013) MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res., 41, W8–W14
57. Marino Buslje, C., Teppa, E., Di Doménico, T., Delfino, J. M. and Nielsen, M. (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. PLoS Comput. Biol., 6, e1000978
58. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13, 2498–2504
59. 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
60. Mizuno, T. (1987) Static bend of DNA helix at the activator recognition site of the *ompF* promoter in *Escherichia coli*. Gene, 54, 57–64
61. Nakagawa, S., Gisselbrecht, S. S., Rogers, J. M., Hartl, D. L. and Bulyk, M. L. (2013) DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc. Natl. Acad. Sci. USA, 110, 12349–12354
62. He, X. and Wang, S. (2014) DNA consensus sequence motif for binding response regulator PhoP, a virulence regulator of *Mycobacterium tuberculosis*. Biochemistry, 53, 8008–8020
63. Korostelev, Y. D., Zharov, I. A., Mironov, A. A., Rakhmaininova, A. B. and Gelfand, M. S. (2016) Identification of position-specific correlations between DNA-binding domains and their binding sites. Application to the MerR family of transcription factors. PLoS One, 11, e0162681
64. Studier, F. W. (2005) Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif., 41, 207–234
65. 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
66. Sheridan, R. C., McCullough, J. F., Wakefield, Z. T., Allcock, H. R. and Walsh, E. J. (2007) Phosphoramidic Acid and its Salts Inorganic Syntheses. Hoboken: John Wiley & Sons, Inc.
67. Sambrook, J., Russell, D.W. (2006) Isolation of DNA fragments from polyacrylamide gels by the crush and soak method. CSH Protoc, 198–202
68. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410
69. Wallace, I. M., O'Sullivan, O., Higgins, D. G. and Notredame, C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res., 34, 1692–1699

