Persistence and plasticity in bacterial gene regulation

Leo A. Baumgart[1], Ji Eun Lee[1], Asaf Salamov[1], David J. Dilworth[1], Hyunsoo Na[1], Matthew Mingay[1], Matthew J. Blow[1,2], Yu Zhang[1], Yuko Yoshinaga[1], Chris G. Daum[1], Ronan C. O'Malley[1,2]

[1]U.S. Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720;
[2]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Correspondence: romalley@lbl.gov

Abstract

Organisms orchestrate cellular functions through transcription factor (TF) interactions with their target genes, though these regulatory relationships are unknown in most species. Here we report a high-throughput approach for characterizing TF-target gene interactions across species, and its application to 354 TFs across 48 bacteria, generating 17,000 genome-wide binding maps. This dataset revealed themes of ancient conservation and rapid evolution of regulatory modules. We observed rewiring, where the TF’s sensing and regulatory role is maintained while the arrangement and identity of target genes diverges, in some cases encoding entirely new functions. We further integrated phenotypic information to define new functional regulatory modules and pathways. Finally, we identified 242 new TF DNA binding motifs, including a 70% increase of known *Escherichia coli* motifs and the first annotation in *Pseudomonas simiae*, revealing deep conservation in bacterial promoter architecture. Our method provides a versatile tool for functional characterization of genetic pathways across all organisms.

Introduction

Transcription factors (TFs) are the primary regulators of gene expression. They modulate the rate of RNA expression via direct binding at specific genomic sites near their target genes, and coordinate genome-wide transcriptional programs that allow cells to adapt to dynamic conditions. Understanding the interactions between TFs, their binding sites, and the collection of target genes they regulate is key to our ability to model transcriptional programs and ultimately engineer them. However, large-scale decoding of these interactions is currently limited to a small set of model organisms, in part because of the limitations posed by existing technologies. In vivo methods such as ChIP-seq[1–4] can capture TF binding in a physiologically relevant state, but are difficult to scale up to match the hundreds to thousands of TFs found in a single organism. In contrast, in vitro methods such as protein binding microarrays (PBMs)[5] and systematic evolution of ligands by exponential enrichment (SELEX)[6–8] can be leveraged at large scales. However, most in vitro methods rely on indirect characterization of binding sites by identifying TF binding motifs using synthetic short DNA sequence pools, followed by scanning for these motifs in the reference genome to predict TF binding sites. As a result, these in vitro assays are unable to capture effects of native genomic context including DNA shape, chemical modifications, and conserved local cis-element architecture that can have a large impact on TF binding specificity.

DNA affinity purification sequencing (DAP-seq), the method we developed in 2016,[9] combines advantages of both in vivo and in vitro assays. Like ChIP-seq, DAP-seq directly measures TF binding in native local genomic sequence contexts, however DAP-seq can be easily scaled up to comprehensively assay all TFs within a species, as demonstrated previously with *Arabidopsis* (see **Supplementary Notes**). To achieve this, DAP-seq leverages in vitro expressed and affinity-purified TFs to capture binding events within fragmented native genomic DNA (gDNA), followed by high-throughput sequencing.[10] DAP-seq has proven to be an effective method to study TF binding sites in a variety of model organisms[11–13] and the resulting large-scale datasets have been central to a variety of approaches for understanding gene regulation.[14–16]

One limitation to DAP-seq, as well as all other existing TF binding assays, is the significant upfront investment required to purify each TF of interest. This is the major bottleneck for all high-throughput TF DNA binding techniques, and the primary restriction on the total number of TFs that can be assayed. In the original DAP-seq method, TF proteins are expressed in vitro from *E. coli* plasmid templates, which allows fusion of the TF coding sequence with an affinity tag that is required for the pulldown of the expressed TF and the DNA sequences it binds to. This hinders widespread application to non-model organisms for which pre-existing TF plasmid collections are not available, particularly in microbial studies where short generation times and high mutation rates have generated a vast diversity of TFs. In addition, the original DAP-seq method only enables mapping gDNA binding properties in a single genome at a time. The relationships between TFs, their binding sites, and target genes are known to be conserved sometimes over incredibly long periods of time, and have been shown to be a predictor of conserved biological functions.[17,18] Therefore, a broader understanding of how TF binding sites and target genes evolve across phylogenetically relevant sets of species will be of great value to reveal the conservation, evolution, and the function of TF-target gene pathways, of which our current understanding is very limited. Beyond the specific biological insights gained from this dataset, the methods presented here are broadly applicable and will enable future surveys of diverse genomes and annotation of intergenic sequences.

Results

Streamlined protein expression from gDNA or cDNA

To address the bottleneck imposed by plasmid-based protein expression we developed biotin-DAP-seq, a streamlined clone-free workflow where tagged TF proteins are expressed from templates that are PCR amplified directly from gDNA or cDNA (**Figure 1**). First, we designed primers flanking the TF of interest. The primers contained a T7 promoter and other required components for expression with a commercial in vitro coupled transcription and translation mix but did not contain an affinity tag sequence. Instead, a biotin tag was introduced directly during translation by spiking in a tRNA loaded with biotinylated lysine.[19] This resulted in incorporation of biotin tags at a random subset of lysine codons within the protein sequence. This biotin-tag allowed for downstream affinity capture of TFs along with bound DNA sequences, using streptavidin-coated magnetic beads.

We tested this new streamlined TF expression approach using a set of 216 known *Escherichia coli* TFs. We detected one or more putative binding sites in at least one of two trials for 125 TFs (58%, **Table S1**, see also **Supplementary Notes**). We examined the dataset for known *E. coli* TF binding sites represented in RegulonDB[20] and found at least one published site in our dataset for 113 TFs (90%), at least half of all known sites for 64 TFs (57%), and all sites for 40 TFs (32%) (**Figure S1**). These findings are consistent with other in vitro techniques, where a subset of binding sites may not be detectable due to low affinity, requirement for a cofactor, or binding limited to specific in vivo conditions.[21,22] Despite these limitations, in vitro techniques have been successfully applied in recent years to discover many new functional binding sites that were not reflected in *E. coli* databases.[23–26] Our results demonstrate that biotin DAP-seq is a high throughput approach that allows detection of known functional TF binding sites, as well as discovery of additional putative binding sites. We also demonstrated that this method can be applied to study eukaryotic TFs containing intronic sequences, by amplifying the complete coding sequences directly from cDNA (**Figure S2**). By eliminating the need for plasmid construction, we reduced the time required to produce a species-wide DAP-seq dataset from months to days, and the total reagent cost by more than half.

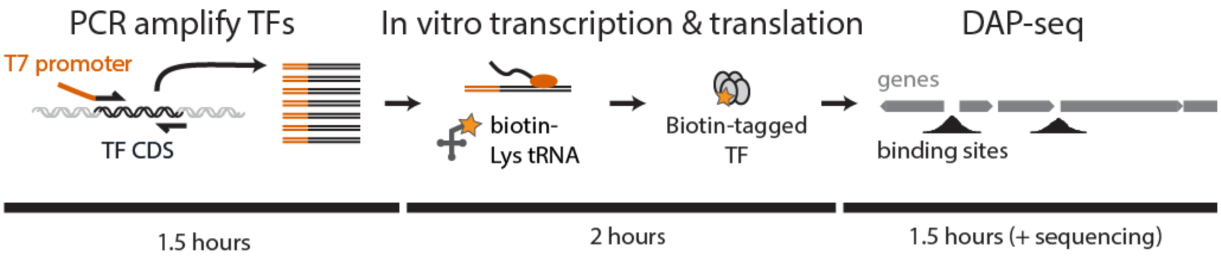


Figure 1. Streamlined protein expression directly from PCR products amplified from gDNA or cDNA circumvents the need for plasmid construction. Addition of tRNA loaded with biotinylated lysine results in incorporation of biotin tags at a random subset of positions that encode a lysine within the TF amino acid sequence. With this protocol an experiment can be completed in approximately 5 hours.

The streamlined biotin-DAP-seq is particularly suited to studying non-model organisms. We demonstrated this by mapping TF binding sites in *Pseudomonas simiae*,[27] an emerging model for plant-commensal microbes that currently has no available TF binding site annotations.[28,29] We compiled a comprehensive set of 567 putative *P. simiae* TFs by combining three different predicted gene annotations from GenBank[30], RefSeq[31], and IMG,[32] of which 138 (24%) were successful in two replicate experiments. The lower overall success rate compared to well characterized *E. coli* TFs is not surprising, as we screened any gene with predicted DNA-binding activity, many of which may not be functional TFs.

Multiplexed TF mapping: multiDAP

In parallel, we developed multiDAP, a method that allows mapping TF binding sites in multiple genomes simultaneously. By using a pool of gDNA samples from different species we were able to directly map TF binding sites across a diverse array of organisms. For our TF set we used a total of 354 TFs (**Table S2**), including the 138 *P. simiae* TFs from the preceding screen, and all 216 *E. coli* TFs regardless of previous success or failure. (**Figure 2a**). Next, we prepared gDNA fragment libraries from 48 bacterial genomes (**Table S3**), each marked with a unique molecular barcode (**Figure 2b**). We selected the set of 48 bacterial species to cover a large evolutionary distance, with a larger proportion of close relatives of *E. coli* and *P. simiae* for higher resolution of local conservation and variation. We pooled all 48 barcoded gDNA fragment libraries and distributed this pool equally to each well of a microtiter plate containing bead-immobilized TF proteins. After several washing steps, the bound gDNA fraction was eluted and amplified by PCR using a set of uniquely barcoded primers to mark the identity of each well and the corresponding TF. At this point samples were pooled together for sequencing.

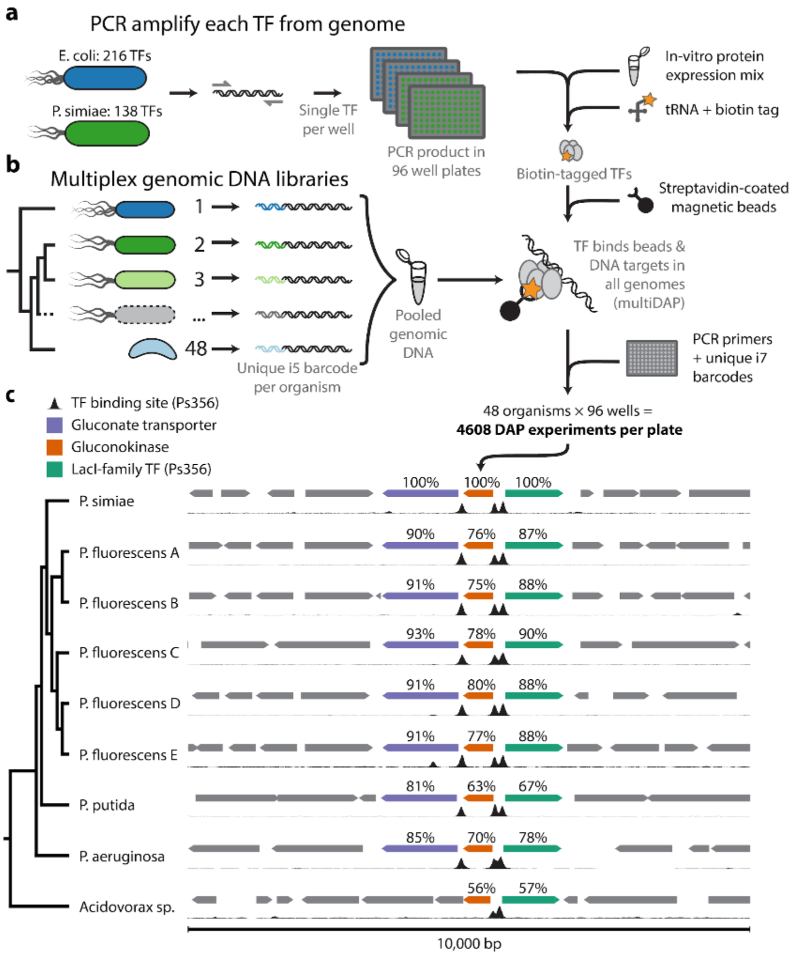


Figure 2. Overview of multiDAP experimental setup and example of resulting data. (a) TFs were PCR amplified directly from *E. coli* or *P. simiae* genomic DNA and used as templates for in vitro protein expression, with biotin tags incorporated. (b) Genomic DNA fragment libraries were prepared from 48 bacterial species, each with a unique molecular i5 barcode. All 48 libraries were pooled and distributed to TF protein plates. TFs bound to target DNA fragments and streptavidin-coated magnetic beads. After washing, remaining bound DNA fragments were PCR amplified using a unique i7 barcoded primer for each well before pooling and sequencing. (c) After demultiplexing of sequence reads, alignment to genome revealed TF binding sites in each species evident as peaks in coverage plots. Coverage plots for a *P. simiae* TF Ps356 are shown in black across a 10,000 base pair window. Genes predicted to be regulated by Ps356 in each species are colored by predicted gene function and percentages indicate BLAST amino acid identity to *P. simiae* orthologs.

Based on the combination of molecular barcodes from each sequencing read, the dataset was computationally de-multiplexed to yield the equivalent of one DAP-seq dataset per TF per organism. After alignment to the corresponding genomes, regions that contain TF binding sites were apparent as peaks, resulting from the pileup of DNA fragments that are bound by the TF. By mapping the binding of the 354 TFs from *E. coli* and *P. simiae* across the set of 48 bacterial genomes, we produced a combinatorial dataset equivalent to 17,000 DAP-seq experiments. This dataset allowed direct comparison across divergent bacterial species to reveal conserved patterns of TF binding at orthologous genes (**Figure 2c**). We recovered binding information for 113 of the 138 (82%) *P. simiae* TFs, and 107 of 216 (50%) of the *E. coli* TFs in a single pass experiment.

We next investigated reproducibility and data quality produced by multiDAP, for which we selected 92 *E. coli* TFs for further benchmarking. PCR primers targeting these 92 TFs, along with 4 negative control samples, were arrayed into a 96 well plate and then tested in three independent multiDAP experiments using the set of 48 genomes. Comparisons between these triplicate experiments demonstrate that multiDAP yields a unique binding signature for each TF, which is highly reproducible across independent experiments (median correlation of 94% between replicates, **Figure S3**). We observed that most TFs show strong affinity to a few genomic sites, and some also reproducibly bind weakly to many additional sites (**Figure S4** and **Table S4**). For some TFs, only the strongest binding sites have known biological functions, however for other TFs even the weakest detected binding sites correspond to known annotated sites (**Figure S4**). Recent studies in *E. coli* have revealed that the true number of regulatory targets for some TFs has been historically underestimated,[21] suggesting that some of the weaker binding sites we observed also represent real unknown regulatory interactions. One approach towards identifying functionally important binding sites is to assess the degree to which these sites have been conserved across evolutionary distances.

Evolutionary conservation of TF targets

Using the multiDAP dataset, we quantified TF target conservation across the 48 bacterial strains and species. Since transcription factors often bind in the promoter region directly upstream of genes,[33] we assigned each peak to the adjacent downstream predicted operon(s) to a set of genes that we predict may be regulated by each TF in each of the organisms. We then calculated a target gene similarity score by comparing the sets of target genes across organisms. We first grouped all protein-coding genes from all 48 species into groups of putative orthologs (orthogroups).[34] Next, we quantified TF target conservation by comparing the set of orthogroups targeted in the species from where the TF itself originated (either *E. coli* or *P. simiae*) with those targeted in each of the remaining 47 organisms. To limit the number of spurious matches based on weak binding sites, we only considered the top ten target operons for each species. The results of this analysis give a global view of TF target gene similarity in divergent bacteria for both TFs from *E. coli* (**Figure 3**) and TFs from *P. simiae* (**Figure S5**). Weak matches do not necessarily imply lack of functional conservation and could be the result of divergent TF motif-specificities between species. However, strong matches suggest conserved gene regulation by the corresponding TF ortholog and may serve as attractive choices for future studies and in vivo characterization.

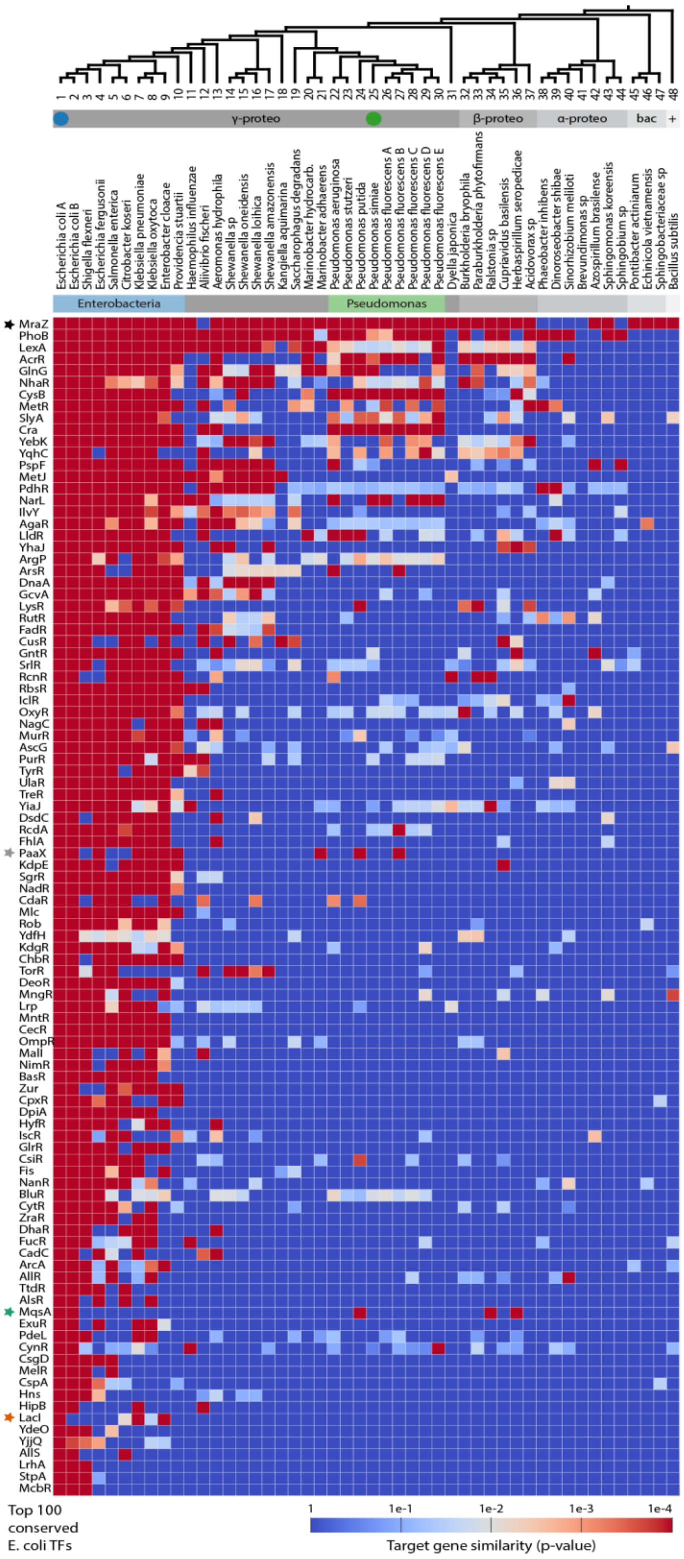


Figure 3. Quantification of TF target gene conservation across species reveals global patterns of conservation and evolution. Gray-shaded vertical bars mark phylogenetic clades (top to bottom): Gammaproteobacteria, Betaproteobacteria, Alphaproteobacteria, Bacteroidetes, Gram-positive. For each *E. coli* TF (rows), the set of target genes in *E. coli* was compared to the set of target genes in each of the other 47 species (columns, labeled 1-48). The two species used in this study as a source of TFs (*E. coli* and *P. simiae*) are indicated in the phylogenetic tree as colored dots (blue and green, respectively). TFs marked with colored stars are discussed in text. Target gene similarity was quantified as the number of matching orthologs appearing in pairs of target gene sets. P-values were determined by comparing to a mock set of target genes randomly selected from each genome for 10,000 iterations. Blue-to-red shades correspond to significance, with darkest red representing the most significant degree of conservation (p-value <= 1e-4). See also Figure S5 for the corresponding analysis results using *P. simiae* TFs.

While some TFs and their targets appear to be confined to a small subset of species, others are highly conserved across large evolutionary distances. As expected, there is a general trend where many TF-target relationships from *E. coli* are well conserved within the closely related Enterobacteria clade. This also holds true for TFs from *P. simiae* within the Pseudomonas clade. One striking feature is the high degree of conservation of some TF targets across clades that diverged long ago. The most highly conserved TF targets are those of the MraZ transcriptional repressor from *E. coli*, which regulates its own expression as well as genes involved in cell division and cell wall synthesis (**Figure 3**, top row black star), This is consistent with previous studies that have shown MraZ to be a highly conserved regulator in bacteria.[35,36] Remarkably, our results indicate that the underlying DNA binding sequence is so well conserved that the *E. coli* MraZ protein is able to bind specifically to the promoter of the *mraZ* ortholog in *Bacillus subtilis*, a Gram-positive bacterium which diverged approximately 2 billion years ago.[37] We found several additional TFs with apparent conservation far beyond the *E. coli* clade, many of which are known to be involved in processes central to bacterial survival and replication, including PhoB (inorganic phosphate metabolism), LexA (response to DNA damage), AcrR (multidrug resistance), and GlnG/NtrC (nitrogen metabolism).

In contrast to these highly conserved features, we also observed evidence of regulatory changes at the sub-species level. To test the ability to accurately discriminate small genetic differences in gene regulation, we included two very closely related strains of *E. coli* (**Figure 3**, species #1 and #2). As expected, we found that the target genes of almost all *E. coli* TFs are conserved between the two strains, with a single notable exception for the LacI repressor protein (**Figure 3**, orange star). This is consistent with the deletion of the LacI binding site upstream of the *lacZ* gene, which is among the small set of documented genetic differences between these two strains of *E. coli*.[38]

Other features appear to be selectively conserved even in closely related species yet are scattered across larger evolutionary distances. For example, the *E. coli* MqsA regulator of the *mqsA*/*mqsR* toxin/antitoxin system is found sporadically throughout the phylum Proteobacteria (**Figure 3**, green star). Similarly, in the case of the *E. coli* TF PaaX, binding sites upstream of genes involved in phenylacetic acid utilization are conserved in seven of the ten Enterobacteria, but also in a subset of the more distantly related Pseudomonas and Marinobacter genuses (**Figure 3**, gray star).

Operon shuffling and genetic rewiring

While this global analysis gives general insights into conserved binding features, closer inspection of individual TF target operons offers specific examples of genetic rearrangements and TF target evolution. We examined the dataset for TFs that target orthologous multi-gene operons in at least two species and found 100 conserved *E. coli* TFs and 95 from *P. simiae*. The apparent rates of TF target evolution span a large spectrum, ranging from highly stable features to those that have diverged rapidly. For example, the strongly conserved *E. coli* autoregulator MraZ shows a rigid operon structure, with only small differences in gene content and operon arrangement in even the most distantly related species (**Figure 4a**). In contrast, an example of strong divergence is seen in the *E. coli* arsenic resistance regulator, ArsR (**Figure 4b**). In *E. coli*, ArsR acts as an arsenic sensor and regulates a set of genes that detoxify arsenic through the combined action of an arsenic exporter (ArsB) and a reductase (ArsC).[39] However in four species of Shewanella the ArsR target operon does not contain any orthologs of these genes, except ArsR itself. Instead they encode distinct arsenic resistance proteins and a predicted glyceraldehyde-3-phosphate dehydrogenase, which is known to be involved in alternative pathways for arsenic detoxification.[40]

We next examined the dataset for evidence of TFs that may have diverged to take on entirely new functions. We found that an ancestor of the *E. coli* TF AscG may have evolved to become PtxS in *Pseudomonas aeruginosa*, acquiring distinct ligand binding functions while maintaining nearly identical DNA binding motifs (**Figure S6**). We identified 13 additional *E. coli* TFs that may have been similarly repurposed (**Figure S7**).

We also demonstrate how multiDAP can be used to discover evidence of conserved metabolic functions within groups of non-model organisms. TF Ps293 from *P. simiae* targets operons predicted to be involved in sugar and dipetide transport and metabolism (**Figure 4c**). The presence of conserved TF binding features in 12 species spread across the phylum Proteobacteria implies that these sites are under strong positive selection, and that all these species likely encode a version of this TF, regulon, and the associated metabolic functions.

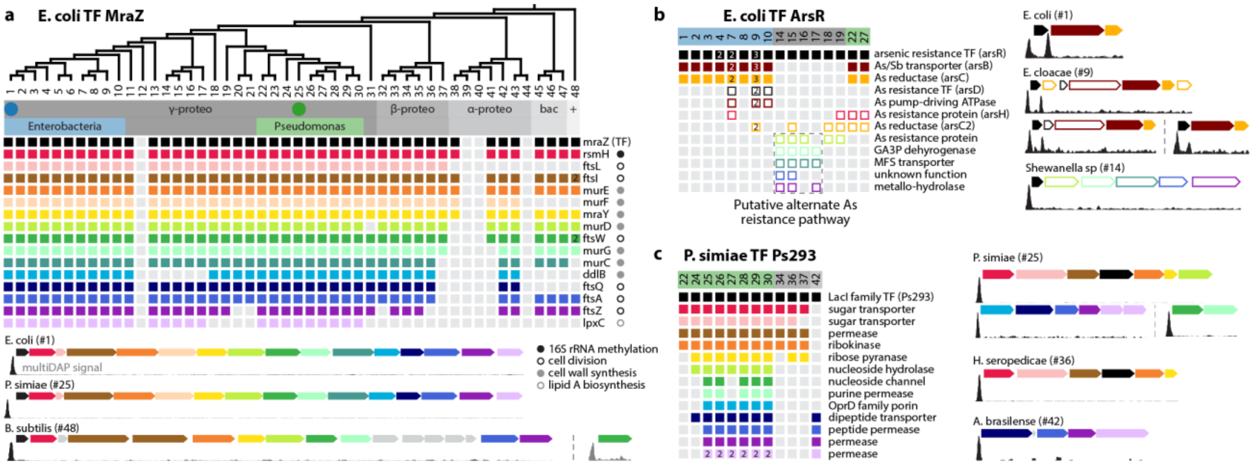


Figure 4. Selected TF target operons compared across species, exemplifying degrees of conversation and instances of rewiring. Predicted orthologs are color coded, with functional predictions from RefSeq annotations. Solid colored genes are present in species from where TF originated (*E. coli* or *P. simiae*), outlined colored genes are absent. Numbers within colored squares indicate multi-copy genes. For each panel, three examples of conserved target operons are shown in more detail, with multiDAP genome coverage signal tracks in black shown below each. (a) The TF MraZ from *E. coli* regulates genes involved in cell division and cell wall synthesis and has highly conserved binding sites and operon structure across divergent clades spanning 2B years. (b) The TF ArsR from *E. coli* is an arsenic sensor that regulates genes for arsenic resistance and is conserved in 16 species from the class Gammaproteobacteria, although the number of copies per genome varies. The ArsR TF is conserved as the first gene in the operon in all cases. In some Shewanella species the regulons appear to have been re-wired to control a distinct arsenic resistance pathway, while other species mix-and-match elements from both pathways. (b) A TF from *P. simiae* (nicknamed Ps293) involved in sugar and dipeptide transport and metabolism is conserved in several species across the phylum Proteobacteria. In *P. simiae* and close relatives it targets multiple distinct operons, different subsets of which are conserved across larger genetic distances.

Functional annotation of TFs and regulons

While multiDAP allows for identification of putative conserved regulons and their functions, additional experimental evidence is required to validate these predictions. One high-throughput approach was described in a study designed to measure the fitness costs of gene knockouts on a range of conditional challenges, including a set of carbon and nitrogen sources.[29] This approach was applied to diverse bacterial species, many of which overlap with our set of 48 species. To investigate how these datasets can complement each other we initially identified a simple and well-characterized example from *E. coli*, FucR. In response to environmental sources of fucose, *E. coli* FucR activates genes involved in fucose import and degradation, as well as the expression of FucR itself.[41] Disruption of *fucR* or other genes in the *fuc* operon resulted in a growth deficit in *E. coli* when grown on fucose. Similarly, in *Klebsiella oxytoca* when the ortholog of *fucR* or genes in its operon were knocked out, a fucose-dependent growth defect was observed. In both *E. coli* and *K. oxytoca* the binding sites predicted by the *E. coli* FucR multiDAP experiment correctly identified the TF and target genes for fucose sensing and metabolism (**Figure 5a**).

Applying the same analysis to the non-model species *P. simiae*, we found that our earlier prediction that *P. simiae* TF Ps293 functions as a regulator of several distinct metabolic pathways (see **Figure 4c**) is also supported by phenotypic measurements (**Figure S8**). In another case, we predict that the *P. simiae* TF Ps109 regulates genes at two different promoters located at distant sites on the chromosome. While the TF knockout confers a growth advantage when 2'-deoxyinosine is the only carbon source, knockouts of genes in both target operons show a growth disadvantage. The multiDAP binding information allows bundling of these individual knockout phenotypes to establish a functional regulatory model, with TF Ps109 acting as a transcriptional repressor at two distant operons involved in 2'-deoxyinosine utilization (**Figure 5b**).

A third example, TF Ps17, shows a conserved functional regulon involved in succinate utilization (**Figure 5c)**. The existing annotations for both the transcription factor (Fis family transcriptional regulator) and target gene (C4-dicarboxylate transporter) are likely too generic to have informed a clear relationship between these TFs and their targets. This demonstrates the value of multiDAP in identifying new regulatory modules, and how some of these predictions can be successfully extrapolated to additional species.

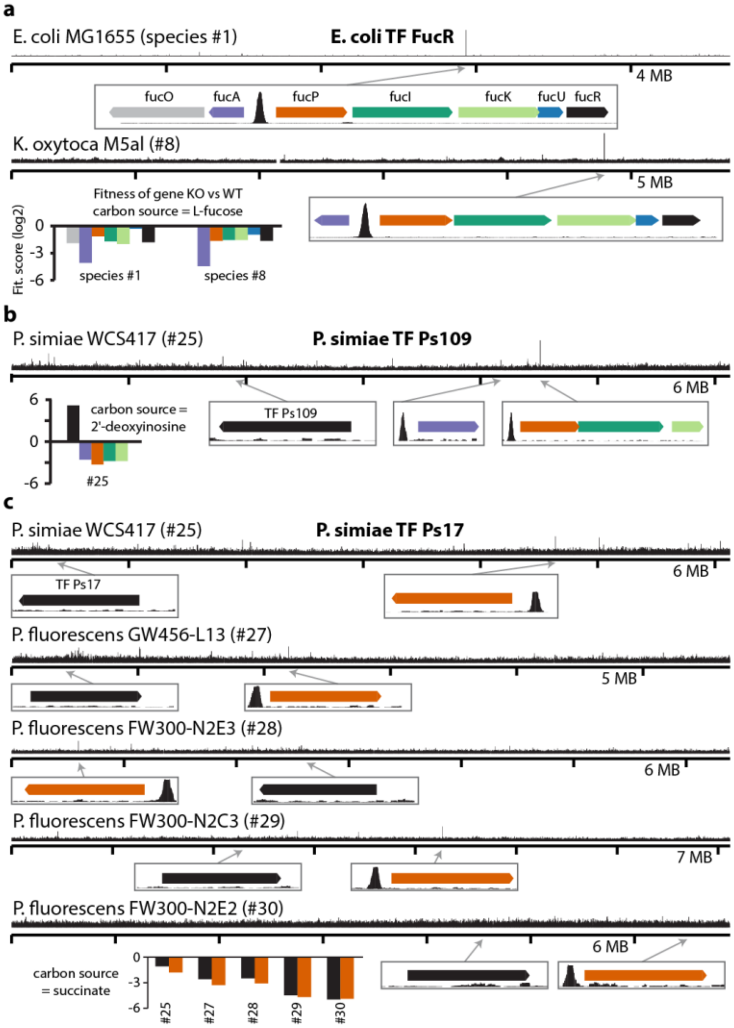


Figure 5. Combining multiDAP with phenotypic measurements enables establishment of functional regulons for a TF across multiple distant operons and multiple species. (a) The *E. coli* autoregulator FucR is a known transcriptional activator which acts on an operon involved in fucose utilization. MultiDAP accurately predicts the target genes in *E. coli*, which is further supported by an observed fitness disadvantage conferred by gene knockouts in the corresponding operons. Both multiDAP and phenotypic measurement also support conservation of the FucR regulon in *Klebsiella oxytoca*. (b) In the non-model organism *P. simiae*, multiDAP allows bundling of TFs and targets located in distant regions of the genome. TF Ps109 and genes found in two distant target operons play a role in 2'-deoxyinosine utilization, as evidenced by phenotypic measurements. The TF gene knockout confers a positive growth impact while the target gene knockouts display a negative impact, suggesting that the TF functions as a transcriptional repressor at the target promoters. (c) MultiDAP results for TF Ps17 reveal a conserved TF and distant target gene found in 5 of the tested Pseudomonas species. Phenotype data shows a positive correlation between the phenotype of the TF and target gene when succinate is supplied as the sole carbon source, suggesting that Ps17 functions as a transcriptional activator and is involved in succinate utilization in all 5 species.

Motifs and promoter architecture

One challenge when studying bacterial transcription factors is that many TFs only bind strongly to few sites in an entire genome, which can make it difficult to confidently identify a DNA binding motif. However, by assaying 48 microbial genomes in a single multiDAP experiment, the total number of examples of binding sequences for each TF is multiplied. Importantly, for the purposes of motif discovery, all detected binding sites are useful, including those that may not be biologically relevant. We were able to call a motif for 124 TFs from *E. coli*, 66 of which are not represented in RegulonDB (**Figure S9**).[20] We used the remaining 58 that are in RegulonDB to validate our motif calls, and found good agreement (50 matches of 58 motifs (86%), each with p-value < 0.01, **Figure S10**). For the 8 remaining motifs, some show matching subsequences flanked by differing bases. We also found good agreement between our motif for the *E. coli* TF YiaJ and the motif published in a recent study which established this TF's function in plant breakdown product utilization (**Figure S11**).[26] The union between 93 known motifs and those newly reported here now provides a total of 158 *E. coli* motifs, a 70% increase over what was previously known for this model bacterium. As no published motifs exist for *P. simiae*, all 118 reported here are new motifs. These results demonstrate that multiDAP experiments offer an expedient and cost-effective method for generating high quality TF binding motifs.

We then applied these motifs to explore conservation of TF binding site architecture in promoters. We mapped motifs back to promoter sequences to identify the precise location and orientation of binding. Autoregulating TFs serve as a particularly tractable set, because there is less ambiguity in identifying the corresponding promoters to compare from each genome. We observed a variety of patterns, some of which are well conserved across divergent species. For some TFs such as MraZ, we consistently find closely spaced clusters of motifs, where the motif orientation is always the same and spacing between individual motifs within a cluster is always exactly 10 base pairs (**Figure 6a**). Another common pattern is exemplified by TF Ps408, which is limited to species in the Pseudomonas clade, and always appears as a doublet with two strong motifs and a 21 base pair gap in the middle **(Figure 6b**). Yet others, such as LysR from *E. coli* have a single strong motif located close to the beginning of the coding sequence **(Figure 6c**). Conserved promoter architecture likely reflects attributes of the TF proteins themselves, including size and shape, ability to form multimers, and protein-protein interactions with other TFs and sigma factors.

Beyond revealing conserved promoter architecture in known gene targets, TF binding sequence motifs can also aid in identifying previously unknown regulatory targets. We expanded our analysis beyond the 48 bacterial species by searching for TF orthologs in all metagenome assembled genomes in the Integrated Microbial Genomes (IMG) database.[32] We identified approximately 1.25M possible TF orthologs, of which >170k showed evidence of conserved autoregulation where TF motifs are enriched in their respective promoters (**Figure 6d**). The presence of motifs in gene promoters may be useful to provide supporting evidence for predictions of protein function, even in species beyond those tested directly.

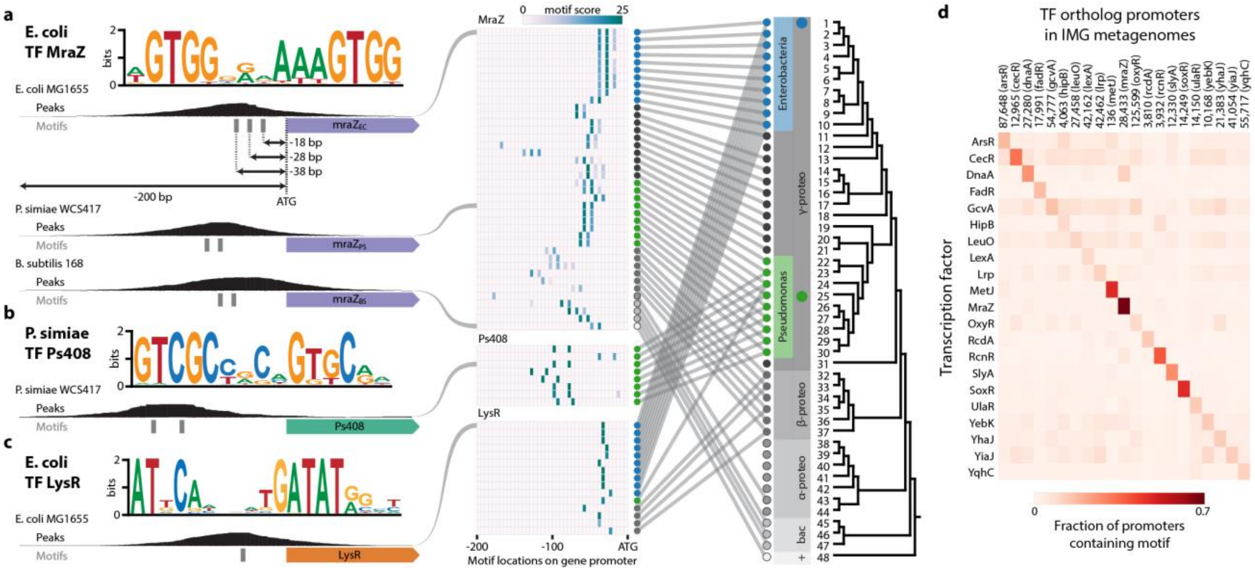


Figure 6. Autoregulator TF binding sequence motifs mapped to promoter sequences reveal conserved promoter architecture. Logo representation of motifs is shown at top, coverage plots are in black, and motif locations are gray bars below. (a) For the *E. coli* TF MraZ, the location of motifs relative to the gene start varies by species, but adjacent motifs are always spaced exactly 10 bases apart even in highly divergent species. Detailed view is shown for *E. coli*, *P. simiae*, and *B. subtilis* (left), and heatmap view is shown for all species (right). (b) A TF from *P. simiae* (nicknamed Ps408) is limited to the genus Pseudomonas where it is found as a pair of motifs spaced 21 bases apart. (c) The *E. coli* TF LysR binds to a single conserved motif located close to the gene start across various species within the phylum Proteobacteria. (d) Autoregulator TF motifs are specifically enriched in the corresponding 100 bp promoter sequences of orthologs throughout metagenomic samples in the Integrated Microbial Genomes (IMG) database.

Discussion

In non-model organisms and metagenomes, a genome sequence provides a wealth of information about gene content and allows prediction of many gene functions based on similarity to known proteins. However, the function of intergenic sequences remains difficult to predict. In this work, we used multiDAP to identify TF binding sites in 48 diverse bacterial species and identified 242 TF binding motifs, most of which have not been described. The resulting dataset illustrates examples of remarkably conserved regulons and promoter architecture, but also reveals patterns of genetic divergence including TF rewiring and repurposing. We speculate that while some of the abundant weak binding sites that we observed for most TFs may not have direct regulatory roles, they could serve as evolutionary raw material for building connections to new gene targets.

Beyond serving as a starting point for future characterization, these results also provide a blueprint for further multiDAP experiments. We showcase specific examples of how these new methods enable discovery and annotation of gene regulatory modules and demonstrate their utility for defining high quality TF binding motifs. These motifs can be valuable in studying promoter architecture, functionally annotating metagenomic sequences, and designing novel synthetic promoters with desired regulatory properties. The new biotin DAP-seq approach facilitates rapid and inexpensive production of tagged TF proteins. MultiDAP allows studying many genomes simultaneously, thereby enriching the biological information extracted from each experiment. These two new techniques can be applied independently or in conjunction for large-scale studies, to begin mapping transcriptional regulatory networks and annotating functional gene regulatory modules across all kingdoms of life.

Methods

Fragment library construction

Genomic DNA from each organism was first sheared using ultrasonic shearing (Covaris LE220-plus) using the following settings: peak power = 450W, duty factor = 30%, cycles/burst = 200. DNA was sheared to an average size of 75 bp in Tris-HCl buffer (pH=8) and applied in multiple cycles of 30 minutes each for a total of 60-90 minutes, allowing time for the water bath to cool between cycles such that the maximum temperature of the samples did not exceed 15°C. Fragment size can be adjusted to meet the desired resolution (see **Figure S12**). After shearing, up to 1 μg of each genomic DNA sample was used to prepare fragment libraries using the KAPA HyperPrep kit and standard manufacturer’s protocol. During the adapter ligation step, custom annealed Y-adapters were introduced at a concentration of 15pM (5 μL adapters in a reaction volume of 110 μL, final adapter concentration = 0.7 pM). These custom adapters were prepared by annealing a full-length i5 index adapter with an index-less stub i7 adapter. Ligated libraries were amplified for 8-10 cycles using primers P1 and P2stub. Strains used in this work and oligonucleotide and barcodes are detailed in the **Table S3** and **Table S5**.

For experiments using *Arabidopsis thaliana Col-0*, the gDNA libraries were constructed from gDNA sheared to an average fragment size of 150 bp.

TF PCR amplification

Primers specific to each transcription factor were designed against the first and last 20-24 bases of the corresponding coding sequence. All non-standard start codons were switched to ATG. In each forward primer, a 5’ constant region was introduced immediately upstream of the sequence annealing to the start of the coding sequence, containing a T7 polymerase promoter and Kozak sequence. In each reverse primer, a 5’ sequence of 30xT was introduced to mimic a poly-A tail and facilitate protein expression in eukaryotic in vitro systems. Primers were arrayed with a single primer pair in each well of a 96-well microtiter plates and used to amplify transcription factor coding sequences directly from the genomic DNA using KAPA HiFi 2x PCR master mix with the following conditions: 10 ng gDNA per well, annealing temperature = 60 °C, 2 minute extension time at 72 °C, total reaction volume = 50 uL, for 24 cycles. PCR products were checked for amplification specificity using an Agilent 2200 TapeStation or Agilent Fragment Analyzer instruments. PCR products were purified using Omega Mag-Bind TotalPure NGS SPRI beads and eluted in 12 µL Tris-HCl buffer pH=8. PCR products were quantified using a fluorescent dye and Synergy plate reader, and then normalized to 100 ng/µL.

For experiments using *Arabidopsis thaliana Col-0*, the gDNA PCR template was substituted with cDNA, generated from RNA extracted from 7-day old seedlings using SuperScript II reverse transcriptase.

In-vitro protein synthesis

TF proteins were expressed in vitro in 96-well microtiter plates using Promega TnT T7 Quick for PCR DNA following the manufacturer's protocol. For each 50 µL reaction, we used 5 µL of purified TF PCR product for a total of 500 ng template. We have obtained good results across a range of concentrations and recommend using 100-1000 ng of each purified PCR product. Also see **Supplementary Notes** for more information on in vitro protein expression and required protein amounts. Negative control wells were included containing mock PCR product, where the PCR was performed with water in place of primers. To produce biotin-tagged TF proteins that can later be purified using streptavidin-coated beads, we also spiked in 4 µL of Promega Transcend tRNA to each 50 µL reaction (see **Figures S13** and **S14**). After combining all components at 4°C, the mixture was incubated at 30°C overnight (12-18 hours).

MultiDAP assay

ThermoFisher Dynabeads MyOne Streptavidin T1 were pelleted on a magnetic rack, washed 4x in PBS pH=7.4 + 0.1% v/v Tween20 and resuspended in an equal volume of this buffer. For each reaction, the following were combined in a master mix (volumes given are per well/reaction): 15 µL resuspended beads, 1 µg salmon sperm DNA, 1 ng amplified DNA fragment library from each organism, and topped off with PBS pH=7.4 + 0.1% v/v Tween20 to a final volume of 50 µL. Master mix volume was scaled up for 384 samples. Subsequent steps were carried out in 96-well plates using a Hamilton Vantage liquid handler.

The bead + library master mix was aliquoted into each well of a 96-well plate, and then transferred into the plates containing the expressed TF proteins. Plates were sealed and incubated for 1 hour at room temperature, with gentle agitation on a rotator.

After incubation, plates were gently centrifuged and unsealed. Next, beads were pelleted on a magnetic rack and washed 4x with PBS pH=7.4 + 0.1% v/v Tween20, with the beads fully resuspended in each wash by pipetting. On the last wash, beads were moved to a fresh 96-well plate. Beads were once again pelleted and wash buffer removed and discarded, after which beads were resuspended in 10 μL i7 index primers (see supplementary table S?) diluted to a final concentration of 1 uM each in Tris-HCl pH=8. An additional 10 μL KAPA HiFi 2x PCR master mix was added to each well. Plates were sealed, vortexed, centrifuged, and placed directly onto a thermocycler running the following program: an initial elution/denaturation step of 98°C for 10 min, followed by 10 cycles of 98°C for 10 sec, 60°C for 30 sec, and 72°C for 30 sec, with a final extension time of 72°C for 1 min then a hold at to 4°C. Finished PCRs were pooled across each 96-well plate, using 10 μL from each well and purified using a 1.4x Ampure bead ratio, followed by elution in 30 μL Tris-HCl pH=8. We found that including an additional gel purification step is critical to remove primer and adapter dimer carry-over, which eliminates issues related to index hopping on Illumina sequencers. The gel purification was carried out using a 2% agarose gel run at 90V in TBE buffer, followed but cutting out the gel piece between approximately 150 bp-300 bp for gel extraction. These gel-cut sizes correspond to 75 bp insert sizes and may need to be adjusted for different insert lengths, while ensuring that anything smaller than 140 bp is removed entirely as this is the fraction that contains contaminating primers, primer dimers, and any un-ligated adapters.

For experiments using *Arabidopsis thaliana Col-0*, the following modifications were made: (1) to account for the larger genome size, 50 ng gDNA library was used in each reaction, and (2) 5 μg salmon sperm DNA was used in each reaction.

Sequencing

Pooled sequencing libraries were quantified by qPCR and sequenced on NovaSeq 6000 S4 Flowcell to target ~1 M paired-end fragments for each of the 36,846 barcode pairs (384 sampless/wells x 48 genomes). Libraries were de-multiplexed, adapter trimmed, and quality filtered using BBTools.[42]

For experiments using *Arabidopsis thaliana Col-0*, to account for the larger genome size each library was sequenced to a depth of 10 M paired-end fragments.

MultiDAP primary sequence data analysis

Analysis scripts are described in brief below. Code is available upon request. Each library was subsampled to at most 1 M fragments, aligned against the corresponding reference genome using Bowtie2[43] and quality filtered with samtools.[44] Coverage plots were generated using deeptools.[45] Peaks were called using MACS2.[46] Target gene assignment for each peak was done using the reference annotation (gff3) and bedtools.[47]

In initial experiments we observed evidence of index hopping, which resulted in cases of leak-through of signal between i7 barcodes. We were able filter this noise from existing datasets using a custom script to identify and filter out leak-through signal between libraries that had been loaded on the same NovaSeq flow cell lane. For subsequent experiments (including all benchmarking experiments), the issue was addressed experimentally by including the stringent gel size selection step prior to sequencing.

Benchmarking

For validation of biotin-DAP-seq, predicted target genes were compared to the RegulonDB database. Each relevant peak region was manually inspected in a genome browser to verify data quality and confirm accurate peak assignment.

For determination of multiDAP reproducibility and TF signal uniqueness in the multiDAP benchmarking experiments, mapped reads were used to generate bigWig coverage files with deeptools and normalized by reads per genomic content (1x coverage across the genome). Pearson correlation of signal across the entire *E. coli* genome was computed using deeptools multiBigwigSummary followed by plotCorrelation with a bin size of 25 bp.

Comparison of gene targets across species

Gene orthology and phylogeny was assigned using Orthofinder2[28]. Phylogenetic trees were visualized using iTOL.[48] For TF target gene comparisons, we used a custom python script: We only considered intergenic peaks and limited the analysis to at most the top 10 target promoters in each organism. We also filtered for peaks with a fold-change >= 5, p-score >= 60, and located < 500 bases from the start codon, where p-score is the value assigned by macs2 equal to -log10(peak p-value). To avoid excessive matches based on very weak binding sites, we filtered out any peaks with a fold-change of less than 5% of the tallest intergenic peak in the same library. We also exclude any TFs that did not perform well in the assay, by examining the corresponding peaks in the species from which they originate *(E. coli* or *P. simiae*). We defined good performance as having at least one intergenic peak with a fold-change >= 15 and p-score

>= 180. For comparison of target gene similarity between species, we only considered matches in organisms that have at least one putative ortholog of the TF in their own genome. Since in some cases a single organism contributes multiple genes to a single orthogroup, we adjusted for the uniqueness of each target gene by weighting each match based on the number of genes that the given species contributes to the corresponding orthogroup. For each target gene set we then calculated a p-score by running the same comparative analysis on an equal sized set of randomly selected genes from each species for 10,000 iterations.

Comparison to phenotype datasets

We downloaded the phenotype dataset for each relevant species from the Fitness Browser website (http://fit.genomics.lbl.gov/). We only considered phenotype measurements that were scored as both significant and specific ("specific phenotypes"). From these datasets, we identified cases where the same conditional challenge yielded a specific phenotype assignment for both the TF itself and the TF target gene(s) that were predicted by multiDAP.

Motif calling and promoter analysis

Motifs were called using MEME.[49] The input sequences used were those flanking the summit position +/- 30 bases. For each TF, we only use the top 30 summits (scored by fold-change over background) in the dataset. Significant motifs (E-value < 0.05) were manually inspected for quality to exclude motifs that were not found enriched near the center of strong peaks and those that had low total information content. Motifs were mapped against promoter sequences using FIMO[50] with default options, and only motifs with scores > 0 were considered.

We used Tomtom[51] to compare the *E. coli* motifs from this study to the motifs published in RegulonDB.[20] Motifs were considered to be in agreement if their comparison produced a score with p-value < 0.01.

We used 113,676 annotated metagenomic datasets from the Integrated Microbial Genomes (IMG)[32] database to extract homologs of *E. coli* TFs and their corresponding promoter sequences. First, for each *E. coli* TF, we found the corresponding orthologs in 48 selected bacterial species based on bidirectional best BLAST hits and tabulated each TF orthogroup with conserved Pfam domains found in them. We searched *E. coli* proteins against predicted genes in metagenomes using MMseqs2[52] with E-value 1e-5 and selected all hits which have a start codon (starting with Met) and at least 100bp upstream sequence from gene. Corresponding promoters were extracted as regions (-100 to +10) around the start codon. Selected orthologs were further filtered to keep only those which have the same Pfam domain(s) and the length within the range of protein lengths of the corresponding orthogroup. To remove the redundant sequences, for each TF, all of its metagenome homologs were clustered using UCLUST[53] at the percent identity cutoff of 80%, and only one TF and its corresponding promoter were kept for each cluster. Motifs in promoter sequences were predicted using FIMO[50] with default options.

Data availability

All raw FASTQ files and peak files are available at:
https://portal.nersc.gov/cfs/m342/jgi_usa/multidap_datashare/

Acknowledgements

The work conducted by the U.S. Department of Energy Joint Genome Institute (JGI), a U.S. Department of Energy Office of Science User Facility, is supported under contract no. DE-AC02-05CH11231. We are grateful to Dr. Adam Deutschbauer for generously providing genomic DNA from many of the bacterial species used in this work. We also thank Dr. Jacqueline Humphries, Dr. Nigel Mouncey, Dr. Len Pennacchio, Dr. Axel Visel, and Dr. Zhuzhu Zhang for helpful discussions and assistance with editing.

Author contributions

LB and RCO designed experiments. LB, JL, HN, YZ, DD, and YY performed experiments. LB, AS, MM, and MB analyzed the data. LB prepared the figures. LB and RCO prepared the manuscript.

Declaration of Interests

The authors declare no competing interests**.**

References

1. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
2. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
3. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
4. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
5. Berger, M. F. & Bulyk, M. L. Protein Binding Microarrays (PBMs) for the Rapid, High-Throughput Characterization of the Sequence Specificities of DNA Binding Proteins. *Methods Mol. Biol. Clifton NJ* **338**, 245–260 (2006).

6. Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.* **9**, 2944–2949 (1989).
7. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
8. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
9. O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280–1292 (2016).
10. Bartlett, A. *et al.* Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* **12**, 1659–1672 (2017).
11. Fischer, M. S., Wu, V. W., Lee, J. E., O'Malley, R. C. & Glass, N. L. Regulation of Cell-to-Cell Communication and Cell Wall Integrity by a Network of MAP Kinase Pathways and Transcription Factors in Neurospora crassa. *Genetics* **209**, 489–506 (2018).
12. de Mendoza, A., Pflueger, J. & Lister, R. Capture of a functionally active methyl-CpG binding domain by an arthropod retrotransposon family. *Genome Res.* **29**, 1277–1286 (2019).
13. Galli, M. *et al.* The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat. Commun.* **9**, 4526 (2018).
14. Uygun, S., Azodi, C. B. & Shiu, S.-H. Cis-Regulatory Code for Predicting Plant Cell-Type Transcriptional Response to High Salinity. *Plant Physiol.* **181**, 1739–1751 (2019).
15. Brooks, M. D. *et al.* Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. *Nat. Commun.* **10**, 1569 (2019).
16. Ricci, W. A. *et al.* Widespread long-range cis -regulatory elements in the maize genome. *Nat. Plants* **5**, 1237–1249 (2019).
17. Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4**, e04837 (2015).
18. Hemberg, M. & Kreiman, G. Conservation of transcription factor binding events predicts gene expression across species. *Nucleic Acids Res.* **39**, 7092–7102 (2011).
19. Kurzchalia, T. V. *et al.* tRNA-mediated labelling of proteins with biotin. *Eur. J. Biochem.* **172**, 663–668 (1988).
20. Santos-Zavaleta, A. *et al.* RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).
21. Ishihama, A., Shimada, T. & Yamazaki, Y. Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res.* **44**, 2058–2074 (2016).
22. Orenstein, Y. & Shamir, R. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.* **42**, e63–e63 (2014).
23. Shimada, T., Fujita, N., Maeda, M. & Ishihama, A. Systematic search for the Cra-binding promoters using genomic SELEX system. *Genes Cells Devoted Mol. Cell. Mech.* **10**, 907–918 (2005).

24. Ishida, Y., Kori, A. & Ishihama, A. Participation of regulator AscG of the beta-glucoside utilization operon in regulation of the propionate catabolism operon. *J. Bacteriol.* **191**, 6136–6144 (2009).

25. Ogasawara, H., Shinohara, S., Yamamoto, K. & Ishihama, A. Novel regulation targets of the metal-response BasS-BasR two-component system of Escherichia coli. *Microbiol. Read. Engl.* **158**, 1482–1492 (2012).

26. Shimada, T., Yokoyama, Y., Anzai, T., Yamamoto, K. & Ishihama, A. Regulatory Role of PlaR (YiaJ) for Plant Utilization in Escherichia coli K-12. *Sci. Rep.* **9**, 20415 (2019).

27. Lamers, J., Schippers, B. & Geels, F. Soil-borne diseases of wheat in the Netherlands and results of seed bacterization with Pseudomonas against Gaeumannomyces graminis var. tritici. in *Cereal Breeding Related to Integrated Cereal Production* 134–139 (Pudoc Wageningen, 1988).

28. Cole, B. J. *et al.* Genome-wide identification of bacterial plant colonization genes. *PLOS Biol.* **15**, e2002860 (2017).

29. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).

30. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **35**, D21–D25 (2007).

31. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

32. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).

33. Browning, D. F. & Busby, S. J. W. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* **2**, 57–65 (2004).

34. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

35. Eraso, J. M. *et al.* The highly conserved MraZ protein is a transcriptional regulator in Escherichia coli. *J. Bacteriol.* **196**, 2053–2066 (2014).

36. Tamames, J., González-Moreno, M., Mingorance, J., Valencia, A. & Vicente, M. Bringing gene order into bacterial shape. *Trends Genet.* **17**, 124–126 (2001).

37. Feng, D.-F., Cho, G. & Doolittle, R. F. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci.* **94**, 13028–13033 (1997).

38. Grenier, F., Matteau, D., Baby, V. & Rodrigue, S. Complete Genome Sequence of Escherichia coli BW25113. *Genome Announc.* **2**, (2014).

39. Xu, C., Shi, W. & Rosen, B. P. The chromosomal arsR gene of Escherichia coli encodes a trans-acting metalloregulatory protein. *J. Biol. Chem.* **271**, 2427–2432 (1996).

40. Chen, J., Yoshinaga, M., Garbinski, L. D. & Rosen, B. P. Synergistic interaction of glyceraldehydes-3-phosphate dehydrogenase and ArsJ, a novel organoarsenical efflux permease, confers arsenate resistance. *Mol. Microbiol.* **100**, 945–953 (2016).

41. Chen, Y. M., Zhu, Y. & Lin, E. C. The organization of the fuc regulon specifying L-fucose dissimilation in Escherichia coli K12 as determined by gene cloning. *Mol. Gen. Genet. MGG* **210**, 331–337 (1987).

42. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/.

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

44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

45. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

46. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

49. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).

50. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

51. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

52. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

53. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

Supplemental Information

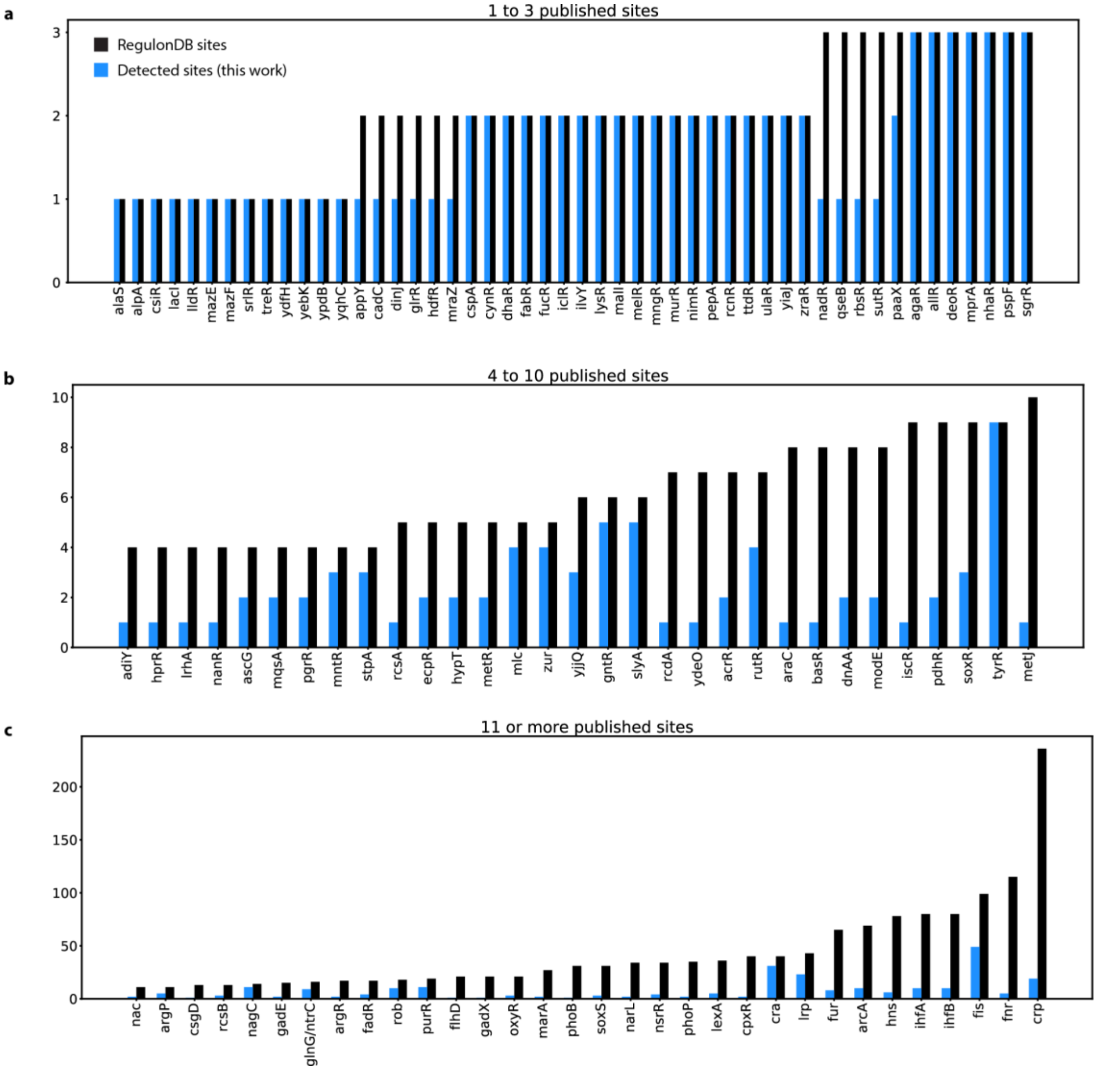


Figure S1. The biotin-DAP-seq method, using TF biotin tagging and streptavidin-coated bead capture, successfully detects *E. coli* TF binding sites that are annotated in RegulonDB. We observed that binding sites for TFs with smaller numbers of published binding sites (panels a and b) are particularly well represented in our dataset. This may indicate that TFs with fewer binding sites have properties that result in stronger signal and better performance in the DAP-seq assays, such as stronger binding affinity or specificity.

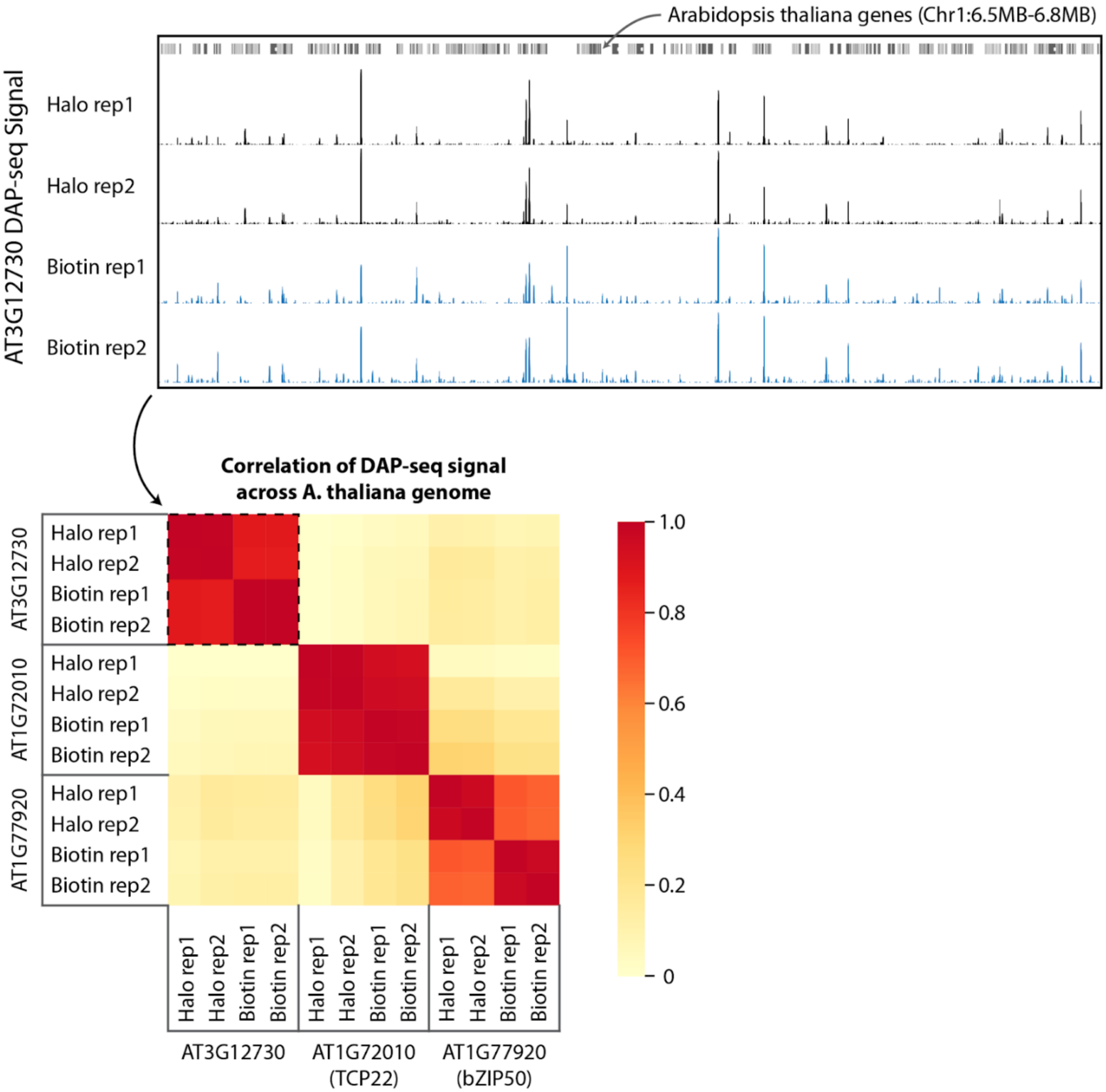


Figure S2. Demonstration of successful biotin-DAP-seq with eukaryotic transcription factors amplified directly from cDNA. The *Arabidopsis thaliana* TFs AT3G12730, AT1G72010, and AT1G77920 were selected to represent three distinct TF protein families: MYB, TCP, and bZIP, respectively. The identified binding signal obtained using biotin-tagged proteins matches closely to DAP-seq with Halo-tagged proteins as used in the original 2016 DAP-seq publication.

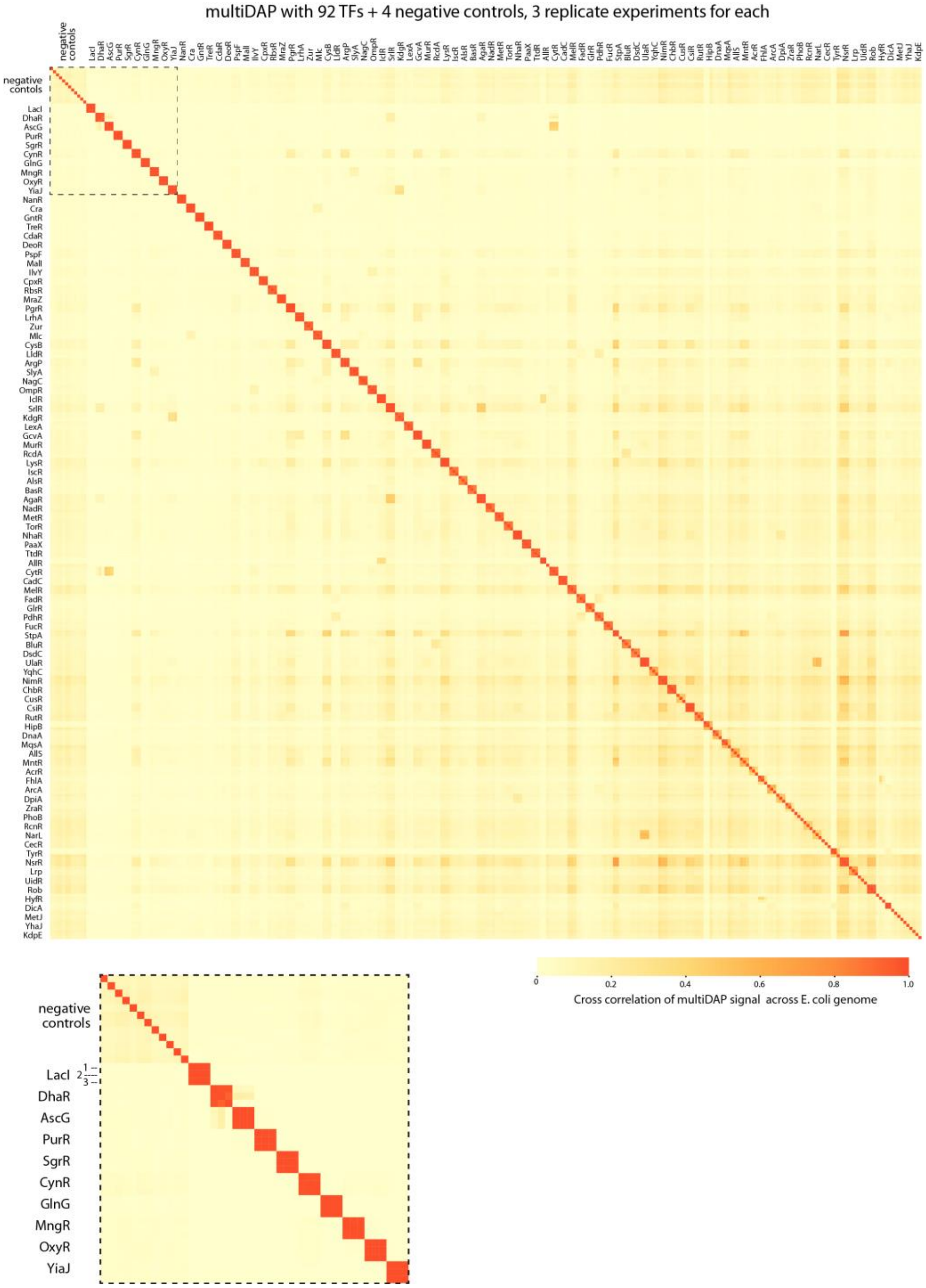
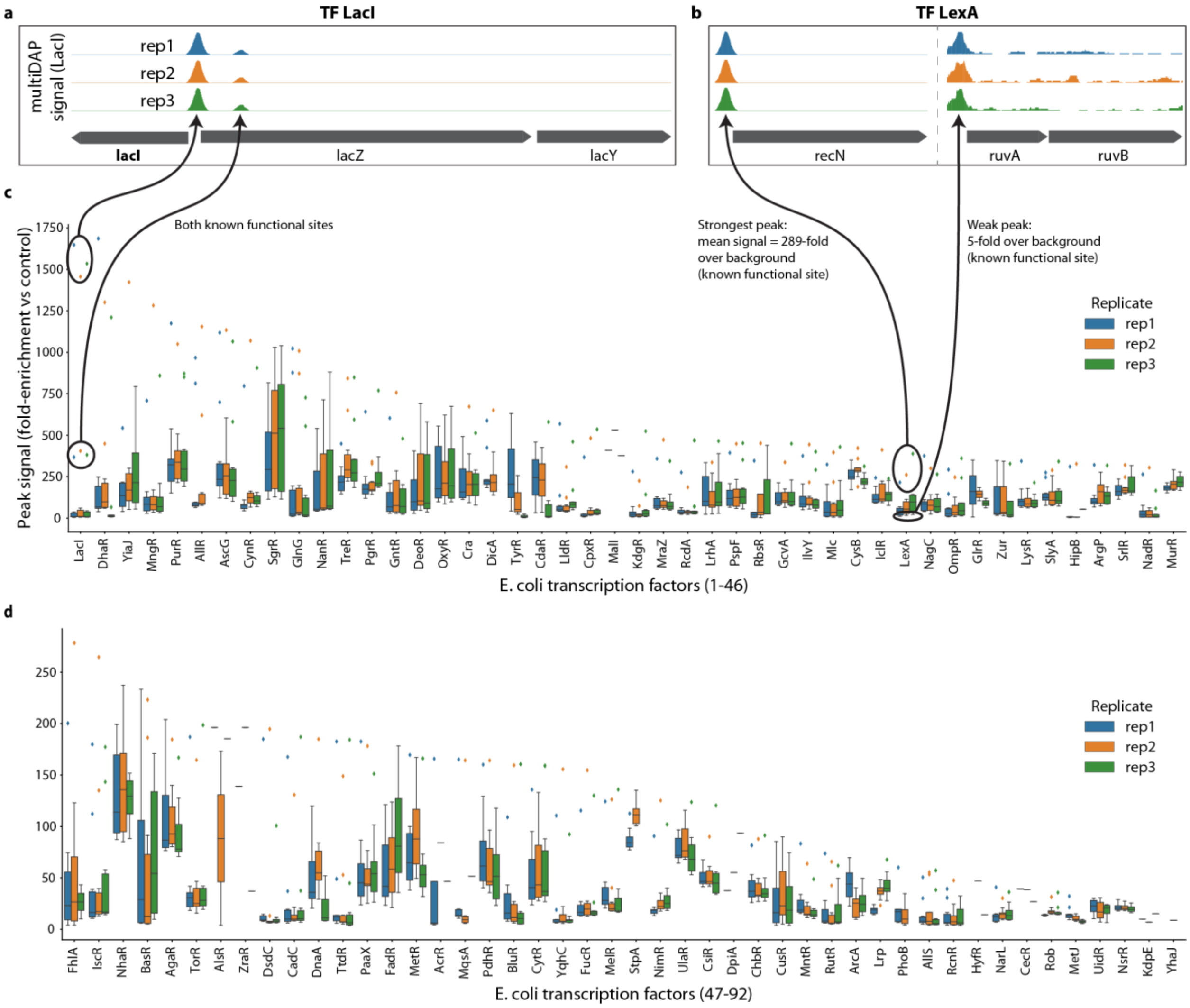


Figure S3. MultiDAP with 92 *E. coli* TFs along with 4 negative control samples, run in three independent triplicate 96-well plate experiments demonstrates unique binding signatures for each TF, which are highly reproducible. Signal correlation among these 288 samples was assessed by splitting the entire *E. coli* genome into 25 bp bins, each with an assigned signal value determined by the sequence read pileup within that bin. Inset shows detail with strong agreement between triplicates, while negative control wells with mock TF expression are largely un-correlated. This implies that background signal is primarily random noise, while individual TFs specifically enrich for binding site regions. Median correlation between replicates is 94%. See also Table S4 for peak numbers and correlations.



Both known functional sites

Strongest peak:
mean signal = 289-fold
over background
(known functional site)

Weak peak:
5-fold over background
(known functional site)

Figure S4. Triplicate multiDAP experiment with 92 *E. coli* TFs show that most TFs have strong affinity to a few genomic sites, and some also reproducibly bind weakly to many additional sites. Scatterplots depict the top 10 peaks ranked by fold-enrichment, as compared to merged negative control samples. (a) LacI is known to bind at and repress the promoter upstream of *lacZ*, along with a weaker accessory binding site just inside the *lacZ* coding sequence. Both of these binding sites were detected as strong peaks, while the additional weaker detected peaks do not have known biological functions. (b) LexA is known to have multiple binding sites at various promoters. The strongest detected peak corresponds to the known target *recN*, however in this case even the weakest detected peak in the *ruvA* promoter is a known functional regulatory site of LexA. (c) and (d) Top 10 peaks detected in triplicate experiments with 92 *E. coli* TFs.

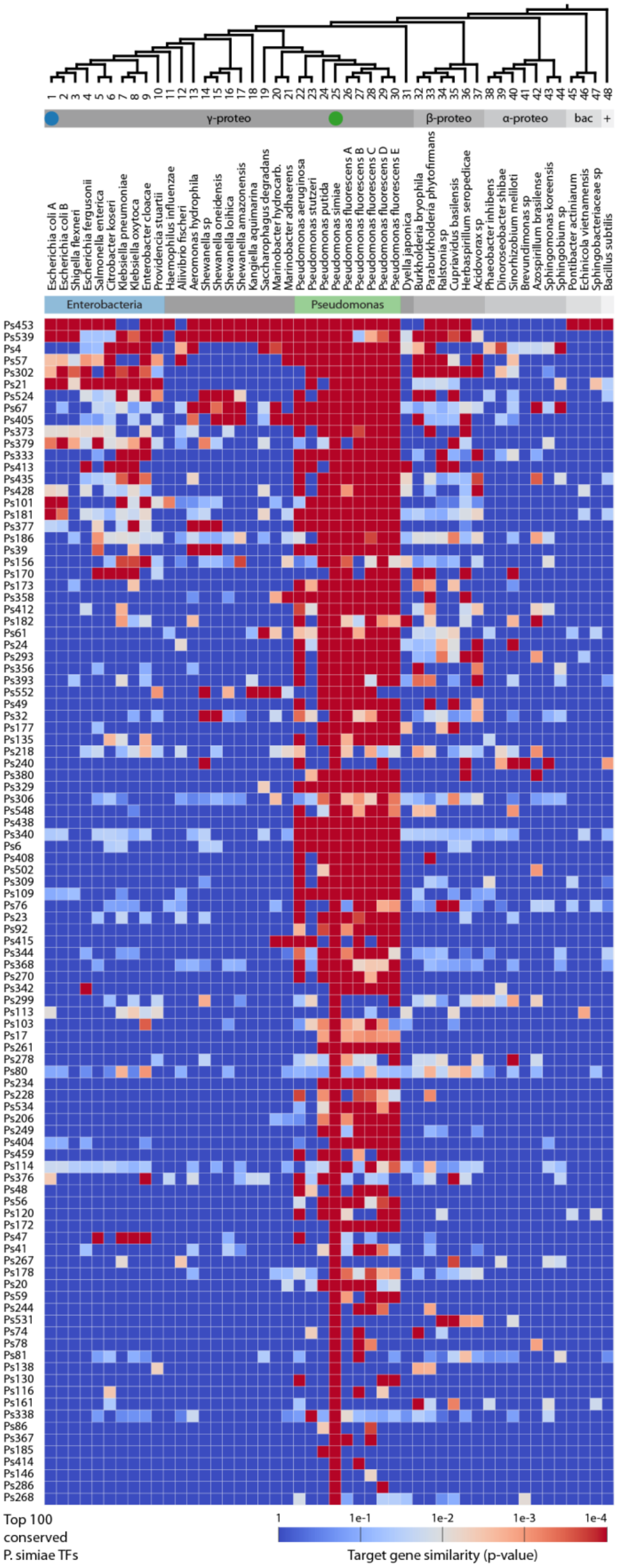


Figure S5. Quantification of TF target gene conservation across species reveals global patterns of conservation and evolution. Gray-shaded vertical bars mark phylogenetic clades (top to bottom): Gammaproteobacteria, Betaproteobacteria, Alphaproteobacteria, Bacteroidetes, Gram-positive. For each *P. simiae* TF (rows), the set of target genes in *P. simiae* was compared to the set of target genes in each of the other 47 species (columns, labeled 1-48). The two species used in this study as a source of TFs (*E. coli* and *P. simiae*) are indicated in the phylogenetic tree as colored dots (blue and green, respectively). Target gene similarity was quantified as the number of matching orthologs appearing in pairs of target gene sets. P-values were determined by comparing to a mock set of target genes randomly selected from each genome for 10,000 iterations. Blue-to-red shades correspond to significance, with darkest red representing the most significant degree of conservation (p-value <= 1e-4).

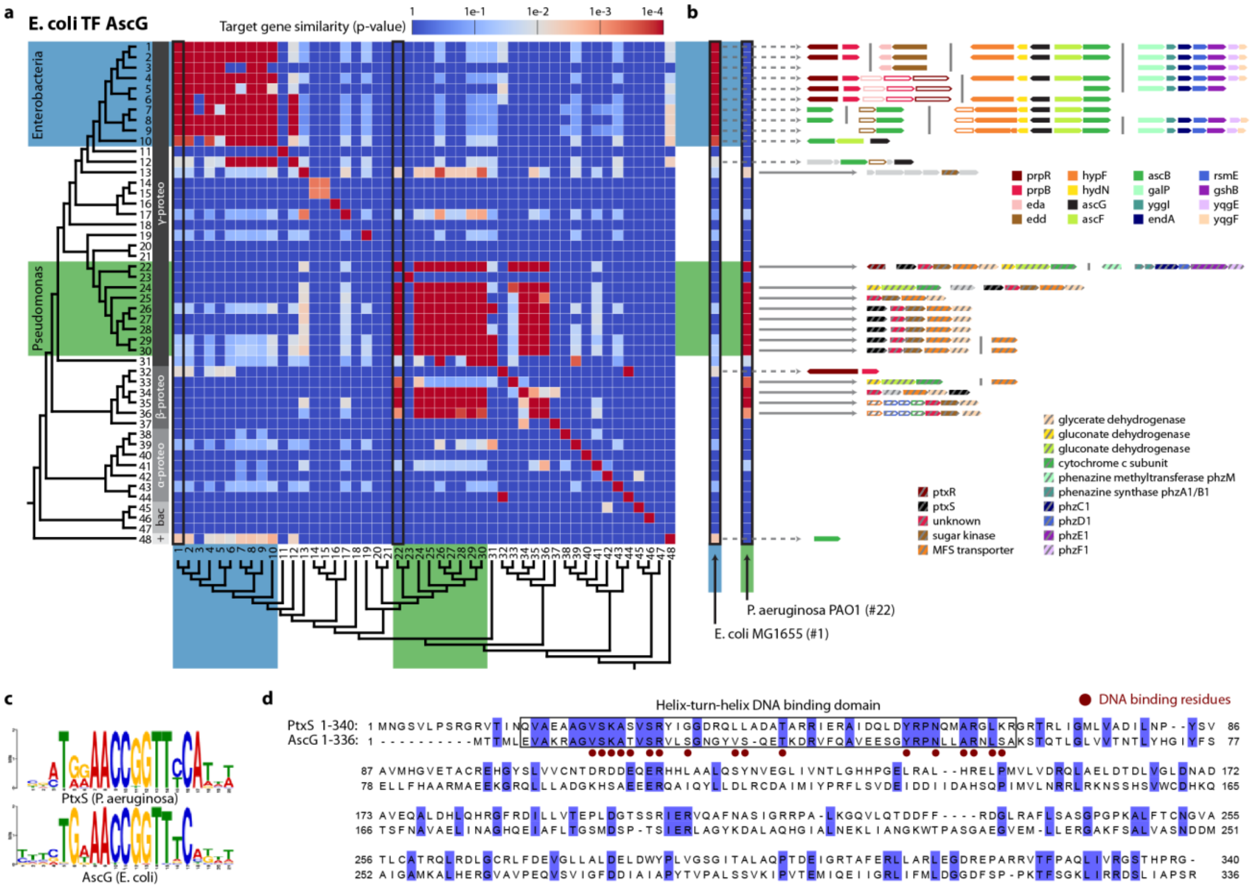


Figure S6. All-vs-all comparison of gene sets targeted by a given TF in each species reveals distinct clusters of conservation in different bacterial clades, suggesting TF functional rewiring. An ancestral form of the *E. coli* TF AscG appears to have diverged to regulate new metabolic functions in the Pseudomonas clade, while the DNA binding specificity has been maintained. (a) The *E. coli* TF AscG provides an example of two distinct clusters of target genes: one cluster mainly limited to the Enterobacteria, and another extending across the genus Pseudomonas and into the class β-Proteobacteria. (b) A closer inspection of the AscG target operons in the model organisms *E. coli MG1655* and *P. aeruginosa PAO1*, along with predicted orthologs of these genes in other species, suggests that the TF's function has diverged between the two clusters. Genes are colored according to their orthogroup: *E. coli* genes and orthologs in solid colors, and those of *P. aeruginosa* with stripes. Functional predictions or gene names from RefSeq annotations are shown in legend. (c) Comparison of the *Pseudomonas aeruginosa PAO1* PtxS DNA binding sequence motif (top) to that of the *Escherichia coli MG1655* TF AscG (bottom) shows high similarity (p-value = 2e-7 as calculated by Tomtom).(d) Despite the nearly identical binding motifs, alignment of the AscG and PtxS amino acid sequences reveals they only share an average 24% amino acid identity across the entire protein sequence (95% coverage). The helix-turn-helix DNA binding domain is conserved at a higher 43% identity, while the C-terminal ligand binding domain shows only 21% amino acid identity. While AscG is known to be induced by the ligand salicin, PtxS is induced by 2-ketogluconate.

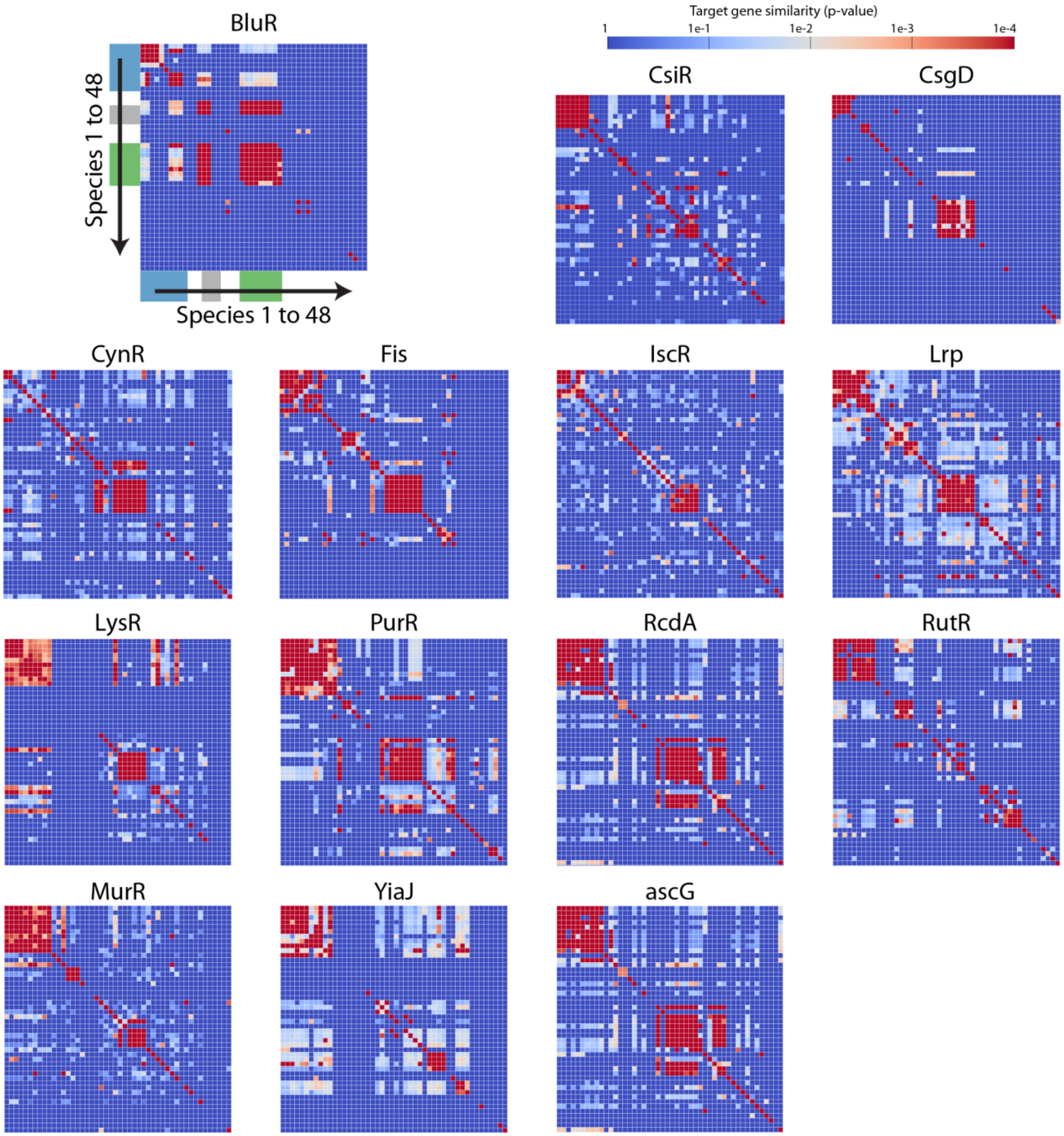


Figure S7. All-vs-all comparison of gene sets targeted by a given TF in each species reveals distinct clusters of conservation in different bacterial clades. In addition to AscG (as detailed in Figure 5), multiple clusters of conserved gene target sets are apparent for 13 other *E. coli* TFs. Clusters tend to appear in the clades representing Enterobacteria (blue), Shewanella (gray), and Pseudomonas (green). This is expected, because within the 48 species, we sampled more densely from within these three clades, while other lineages were sampled much more sparsely. In regions of sparse sampling, any existing clusters of conservation appear as a single red square, which precludes identification as a true cluster.

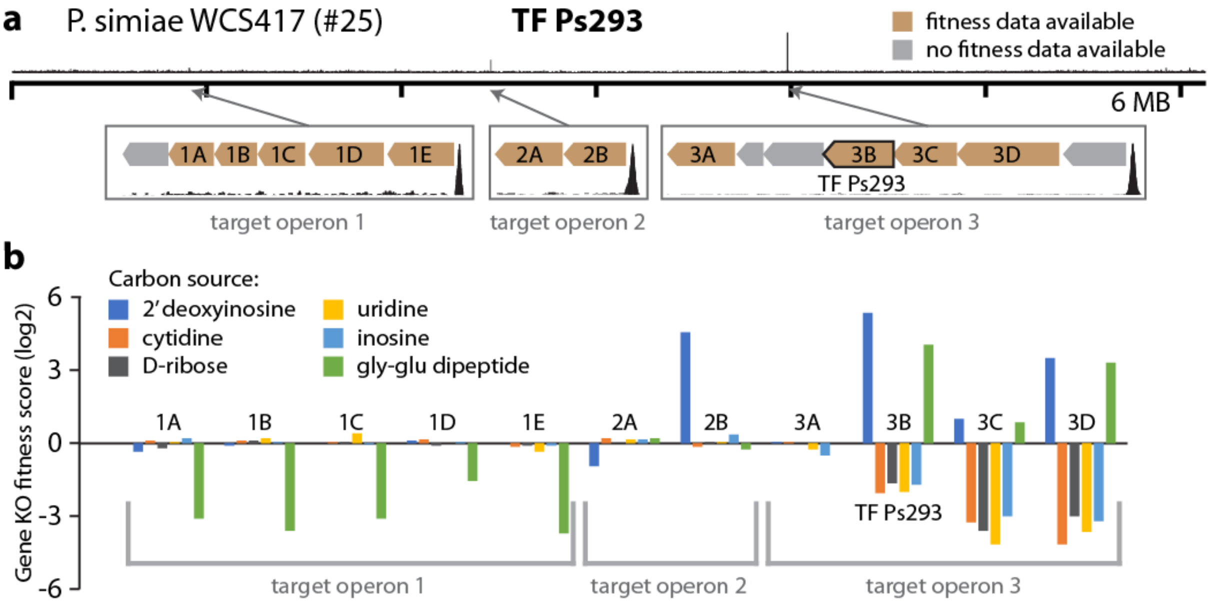


Figure S8. Functional predictions of *P. simiae* TF Ps293 regulon based on multiDAP (see Figure 4c) are supported by phenotypic measurements. (a) MultiDAP data indicates that TF Ps293 targets three distantly located operons, here designated target operons 1, 2, and 3. (b) As evidenced by phenotypic measurements of gene knockouts, the genes in each operon are responsible for distinct metabolic functions: operon 1 is involved in gly-glu dipeptide metabolism, operon 2 in 2'deoxyinosine metabolism, and operon 3 (which contains the TF Ps293 itself) in both of these functions as well as metabolism of several additional carbon sources. The TF knockout results in a phenotype in all of these conditions, while other genes in the regulon appear to only be important for growth on a subset of these carbon sources. This is consistent with the model that TF Ps293 acts as a master regulator for these diverse metabolic pathways. Gene knockouts in operon 1 result in strong growth defects when grown in the presence of gly-glu as the sole carbon source, while knockouts of the TF itself confers a growth advantage under these conditions. In contrast, the knockout phenotypes of genes in operons 2 and 3 do not show this opposing relationship. Taken together, this allows us to predict that TF Ps293 acts as a repressor of operon 1, and an activator of operons 2 and 3.

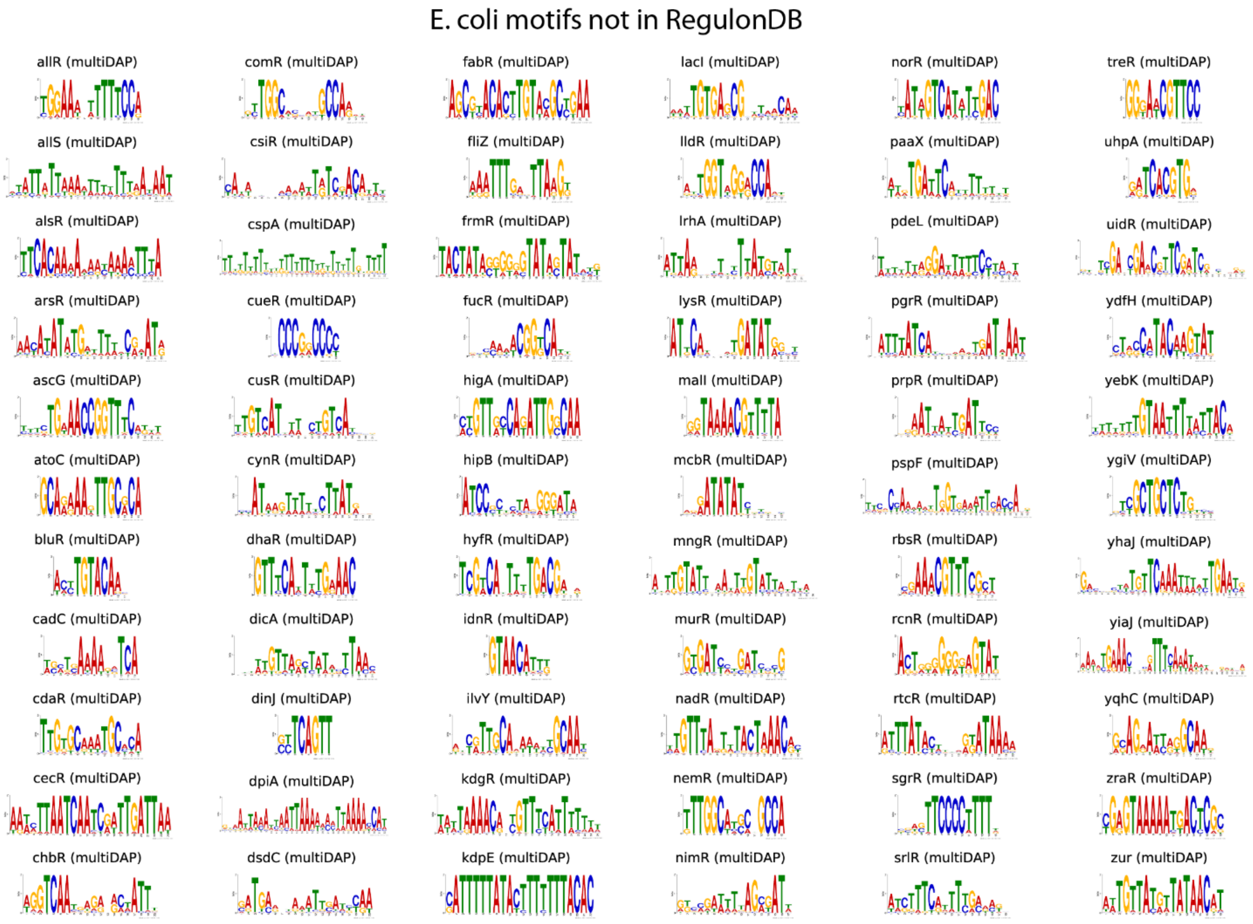


Figure S9. New *E. coli* motifs from this work that are not represented in RegulonDB (n = 66).

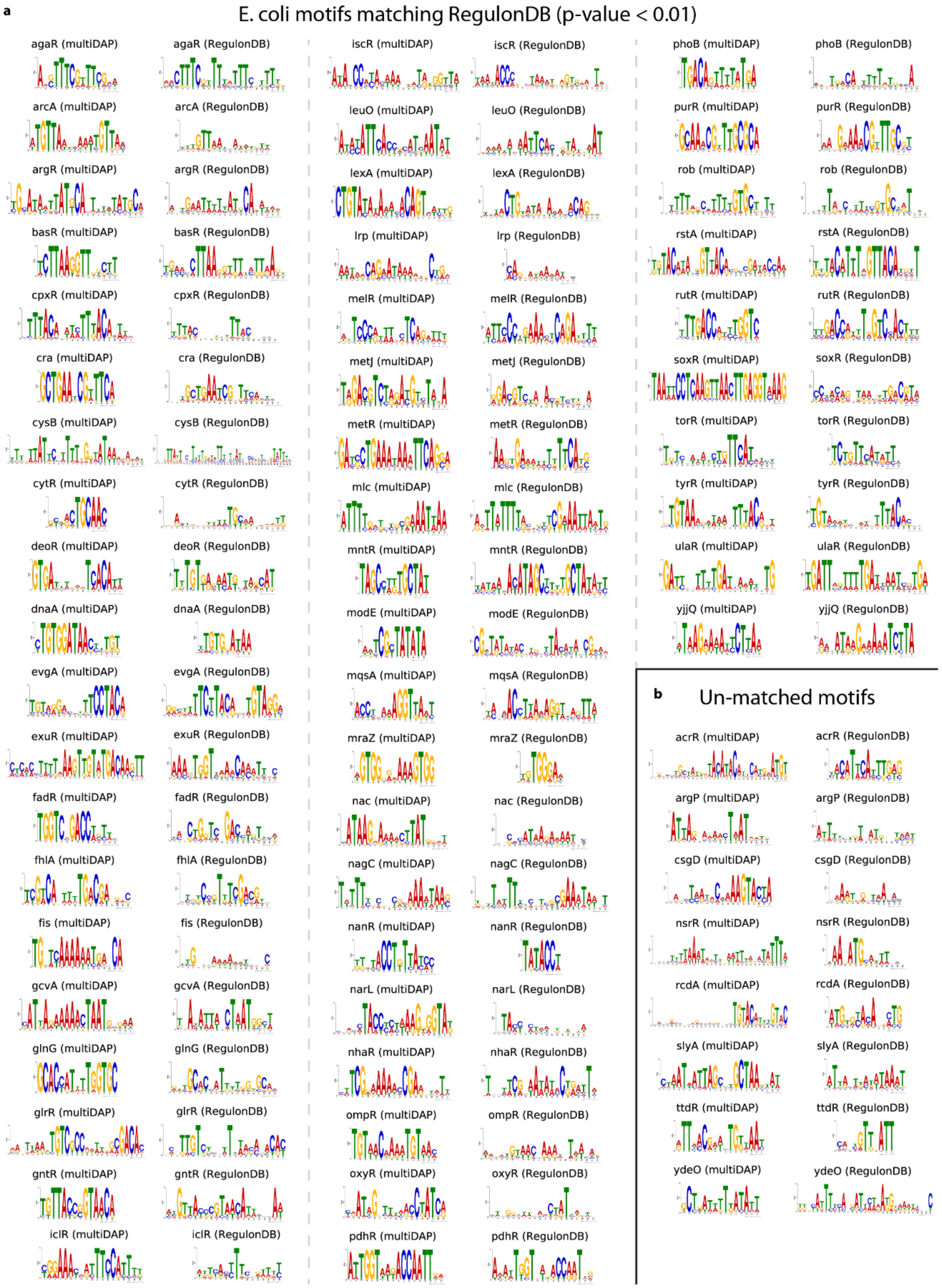


Figure S10. *E. coli* motifs compared to known motifs in RegulonDB. We found good agreement between motifs computed from the multiDAP datasets and RegulonDB: 50 matches (86%) of 58 motifs represented in both datasets. Motifs were considered to be matches if the p-value was less than 0.01, as scored by Tomtom.

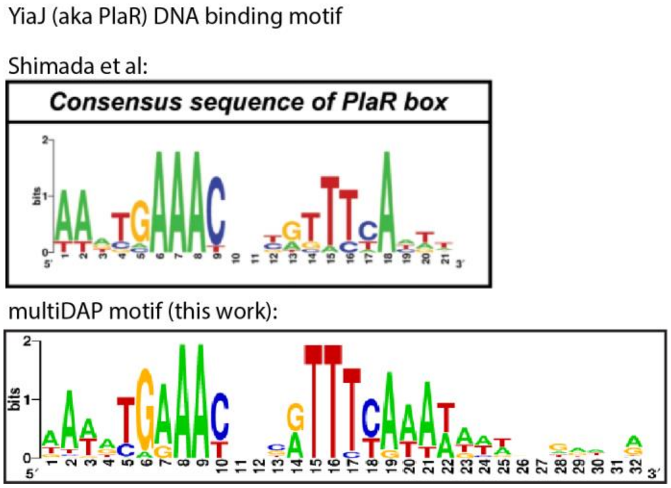

YiaJ (aka PlaR) DNA binding motif
Shimada et al:
Consensus sequence of PlaR box
bits
2
1
0
5′
3′
multiDAP motif (this work):
bits
2
1
0
5′
3′

Figure S11. The motif for *E. coli* TF YiaJ was recently described by Shimada et al (top), who established the TFs function in plant breakdown product utilization and renamed it to PlaR. The motif compares closely to the motif that we identified using multiDAP (bottom).

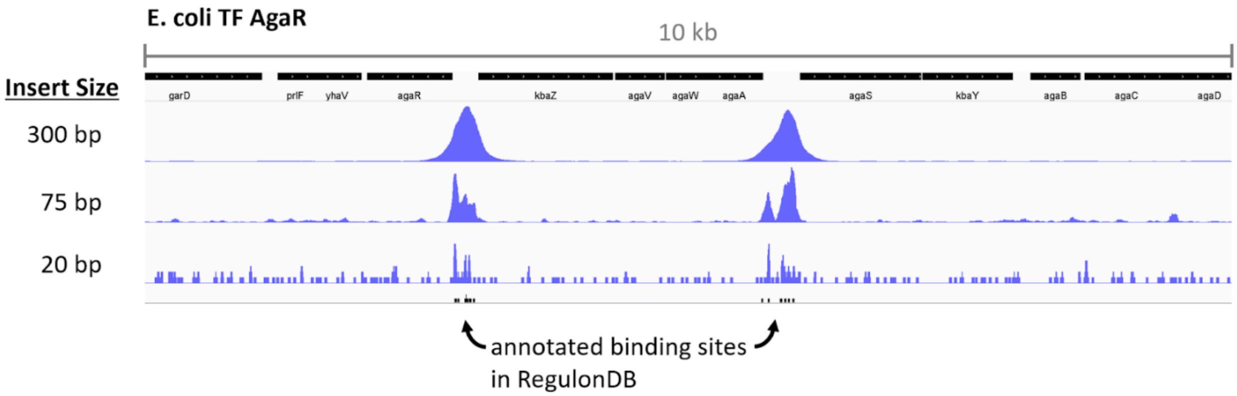


Figure S12. DAP-seq fragment library insert size is a tunable parameter, depending on the desired resolution. For this work, libraries were constructed from genomic DNA sheared to an average size of 75bp, because we found this to offer high resolution while still accurately capturing known binding sites. Note that at extremely short insert sizes (20bp, bottom track) the signal begins to decay, likely because the size is too short to capture the local DNA context of clustered AgaR motifs.

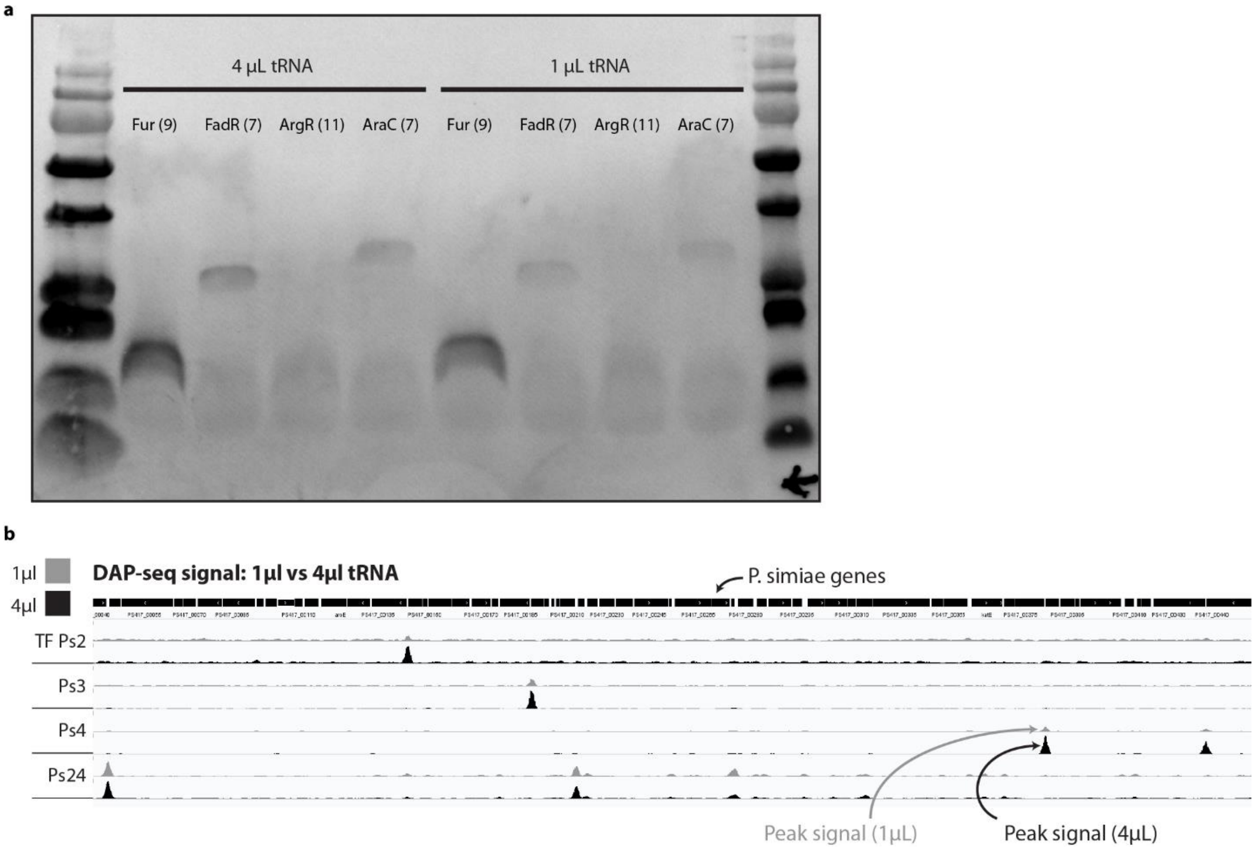


Figure S13. Optimization of biotin-lysine-tRNA reagent used in biotin-DAP-seq. (a) Western blot analysis of four example E. coli TFs demonstrating the slightly increased pull-down achieved by using 4x the amount of biotin-lysine-tRNA in each protein expression reaction. Numbers in parentheses after gene names indicate the number of lysine residues in the respective amino acid sequence. (b) Biotin-DAP-seq signal of four example P. simiae TFs demonstrating the slightly increased pull-down, and resulting increase in signal achieved by using 4x the amount of biotin-lysine-tRNA in each protein expression reaction. For each TF, a pair of tracks is shown: a track in gray (above) corresponding to protein expressed with 1 µL tRNA reagent, and a track in black (below) corresponding to 4 µL tRNA.

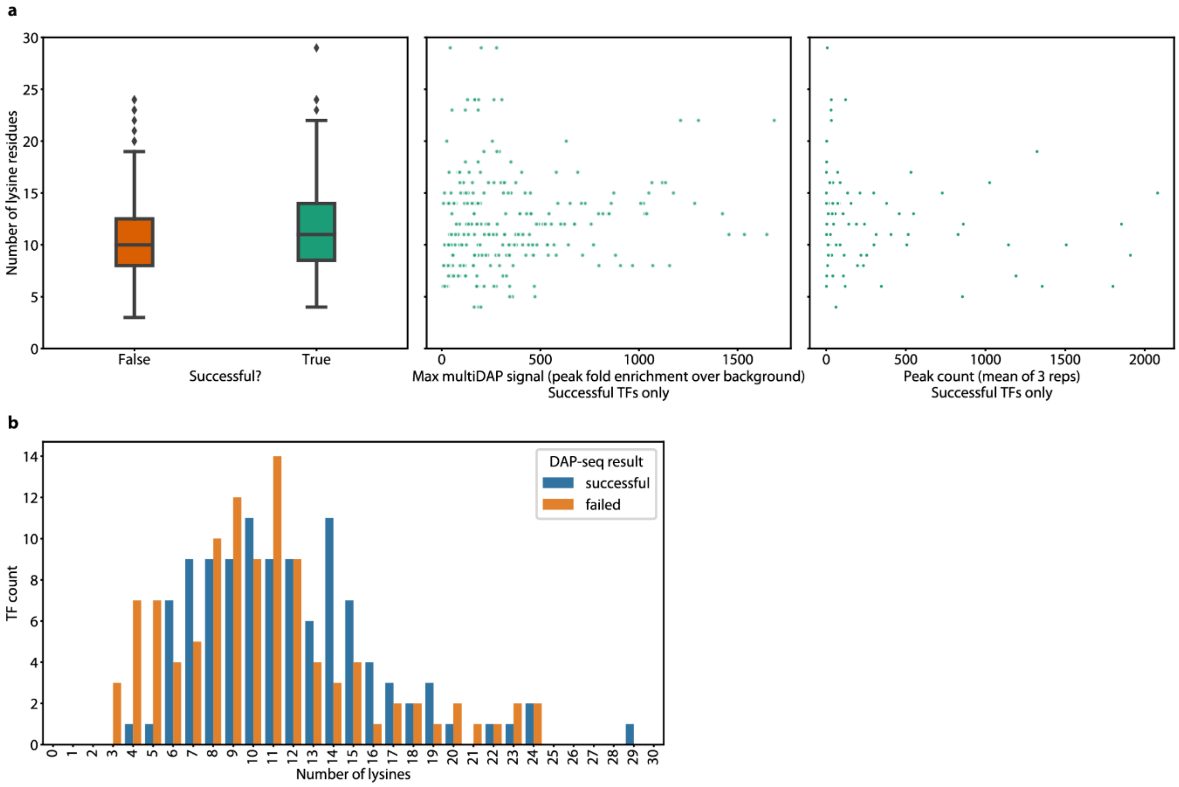


Figure S14. The number of lysine residues in a TF amino acid sequence is a poor predictor of performance in the biotin-DAP-seq assay. (a) The number of lysine residues in E. coli TF amino acid sequences does not predict success in the multiDAP assay (p = 0.35 as calculated by two-sided independent t-test), and also does not predict signal strength or peak number. (b) Counts of successful and failed DAP-seq experiments for TFs with different numbers of lysine residues. Although a small number of TFs having fewer than 6 lysines may suffer increased failure rates, most TFs are not affected. In cases where very few lysines are present, additional lysines can be added at either the N or C terminus by incorporating these directly in the 5' or 3' PCR primers.

Supplementary Notes

Success rate in the biotin-DAP-seq and multiDAP assays

The success rate of 125/216 (58%) for the biotin-DAP-seq E. coli screen is relatively high compared to other large-scale in vitro screens of full-length TF binding. In the original DAP-seq paper, the success rate reported was 30% for Arabidopsis (529/1812).[9] In large-scale SELEX screens published by other labs, reported success rates of full length TFs are 15% (151/984) for human,[S1] and 30% (100/301) for bacteria.[S2] The primary reasons for TF failure for all in vitro assays, including DAP-seq, are likely related to limitations inherent to any protein expression and purification system applied to a large set of TFs with distinct biochemical and binding properties. These include proper protein folding and stability and additional requirements such as co-factors.

In vitro expressed protein levels

We did not observe any significant drop-off in signal when moving from a single species up to 48 multiplexed species. As described in the methods section, we also spiked-in a large amount of salmon sperm DNA (1000x the amount of each genomic DNA) to suppress background signals caused by non-specific DNA interactions. Based on these results, we infer that the relative concentration of genomic DNA is not a major determinant of signal intensity. In the Nature Protocols paper describing the DAP-seq method it was shown that the TF yield is a limiting factor for the assay.[10] If concentrations are lower than optimal, we generally see the loss of the weakest binding sites, though the stronger binding sites are maintained even with a magnitude lower protein yield. Once an optimal protein amount is reached the assay will produce very similar results regardless of additional protein, possibly because binding to the ligand limits total bound protein. During optimization of biotin-DAP-seq we tested multiple conditions to find reaction conditions that balanced success rate and reagent expense. For small experiments and in cases where cost is not a major limiting factor, using a larger amount of protein expression mix in the assay may yield improved results for a subset of poorly expressed TFs.

The Nature Protocols DAP-seq paper also reported tests from multiple cell-free expression systems.[10] Overall, it was found that the rabbit reticulocyte performed better in terms of success rates and signal-to-noise. In initial experiments of the biotin-DAP-seq on bacterial TFs, we tested several different in vitro protein expression systems and again found that the Promega TnT T7 Quick derived from rabbit reticulocyte yielded the best results. In addition, the use of a non-native expression system may have some desirable properties, in particular avoiding confounding factors that could be introduced by the interference of native proteins (e.g. hetero-dimerization or indirect binding due to formation of TF protein complexes).

Influence of genome context on TF binding

In the original DAP-seq paper[9] it was demonstrated that genomic sequence encoded properties of cis-element architecture (motif arrangement and spacing), DNA shape, and chemical modifications (5-methylcytosine) all have a major impact on the TF binding landscape that can be measured by the DAP-seq assay. This previous work demonstrated how cis-element architecture is critical for TF binding affinity and TF-target gene specificity for the Auxin Response Factors (ARF) family of TFs. In this mulitDAP study, we demonstrate that cis-element architecture is important for specificity and binding for some prokaryotic TFs and that it can be a highly conserved feature (**Figure 6**). We are using the term DNA shape to specifically refer to local DNA sequence features (i.e. sequence flanking the motif) that can impact the major/minor groove accessibility and other properties of the DNA double helix that are not captured in DNA motif analysis.[S3] These can include features such as AT-stretches or other sequence properties that are known to impact DNA shape. There is an emerging recognition that DNA shape is an important DNA feature impacting TF binding.[S3] The capability of the DAP-seq assay to capture DNA shape effects on binding has been demonstrated in the original DAP-seq paper.[9] As DNA bases proximal to the motif are the ones that primarily impact DNA shape and thereby impact binding site accessibility, the 75 bp average fragment size is sufficient to capture this property. However, this fragment size is also a tunable parameter that can be tailored to fit the experimental requirements (**Figure S12**). Independently, other groups have also observed the capability of DAP-seq to directly identify complex biologically-relevant native cis-element architectures that impact gene target selectivity and binding affinity,[S4,S5] impact of DNA methylation,[S6] and DNA shape.[S7]

In its original form and as described in this work, the DAP-seq assay allows visualization of global TF binding events in a chromatin-free context, and therefore does not capture important tissue-specific dynamics that are driven by chromatin state. One way to overcome this limitation is to overlay the DAP-seq dataset on tissue- and cell-type specific chromatin accessibility (e.g. DNase-seq and ATAC-seq), and long-range chromatin contact information (HiC). The complementarity of combining DAP-seq with this in vivo chromatin structural information has been highlighted in multiple publications including the original 2016 DAP-seq paper.[9,S8-S10] In addition, new versions of the DAP-seq assay have been developed by other groups to specifically measure the impact of DNA-nucleosome[S11] and additional protein interactors[S4] on TF binding.

Supplementary References

S1. Jolma, A. et al. DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013).

S2. Fan, L. et al. A compendium of DNA-binding specificities of transcription factors in Pseudomonas syringae. Nat. Commun. 11, 4947 (2020).

S3. Zhou, T. et al. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. U. S. A. 112, 4654–4659 (2015).

S4. Lai, X. et al. Genome-wide binding of SEPALLATA3 and AGAMOUS complexes determined by sequential DNA-affinity purification sequencing. Nucleic Acids Res. 48, 9637–9648 (2020).

S5. Galli, M. et al. The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat. Commun. 9, 4526 (2018).

S6. Ichino, L. et al. MBD5 and MBD6 couple DNA methylation to gene silencing through the J-domain protein SILENZIO. Science (2021).

S7. Sielemann, J., Wulf, D., Schmidt, R. & Bräutigam, A. Local DNA shape is a general principle of transcription factor binding specificity in Arabidopsis thaliana. bioRxiv 2020.09.29.318923 (2020).

S8. Lu, Z. et al. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants 5, 1250–1259 (2019).

S9. Song, Q. et al. Prediction of condition-specific regulatory genes using machine learning. Nucleic Acids Res. 48, e62 (2020).

S10. Ricci, W. A. et al. Widespread long-range cis-regulatory elements in the maize genome. Nat Plants 5, 1237–1249 (2019).

S11. Lai, X. et al. The LEAFY floral regulator displays pioneer transcription factor properties. Mol. Plant (2021).